UNIVERSIDADE ABERTA

Missing Data Handling in Health Sciences
A Neuro-Fuzzy Methods Approach

José António Ferreira Lobo Pereira

Doutoramento em Matemática Aplicada e Modelação

2024

UNIVERSIDADE ABERTA

Missing Data Handling in Health Sciences
A Neuro-Fuzzy Methods Approach

José António Ferreira Lobo Pereira

Doutoramento em Matemática Aplicada e Modelação

Advisors: Professora Doutora Teresa Paula Costa Azinheira Oliveira
Professor Doutor Davide Maurício Costa Carvalho
Professor Doutor Anuj Mubayi

August 2024

Acknowledgments

I would like to express my heartfelt gratitude to Professor Dr. Teresa Paula Costa Azinheira Oliveira for the friendship with which she has always honored me, as well as for her motivation, guidance, and availability during both good and challenging times. I owe the completion of this dissertation to her unwavering support.

I also extend my sincere thanks to Professor Davide Carvalho for his friendship and unconditional support in the pursuit of this dissertation. His encouragement has been invaluable.

Furthermore, I am deeply grateful to Professor Anuj Mubayi, a friend for all seasons, for his constant availability and support. His willingness to assist at any time has been a crucial part of my journey.

I would also like to extend my gratitude to the faculty of the Doctoral Program in Applied Mathematics and Modeling for their dedication to their students and for the knowledge they imparted, which allowed me to complete this journey.

Dedication

À minha querida mulher, pela sua infinita paciência, apoio incondicional e amor constante ao longo desta jornada.
Sem o seu suporte e partilha, este trabalho não teria sido possível.

Ao meus queridos pais *in memoriam* pelo seu exemplo e incentivo.

Ao meus queridos filhos e netos a quem roubei tempo precioso.

UNIVERSIDADE ABERTA
www.uab.pt

DECLARAÇÃO DE INTEGRIDADE

STATEMENT OF INTEGRITY

Declaro ter atuado com integridade na elaboração da presente dissertação/tese. Confirmo que em todo o trabalho conducente à sua elaboração não recorri à prática de plágio ou a qualquer outra forma de falsificação de resultados.

Mais declaro que tomei conhecimento integral do Regulamento Disciplinar da Universidade Aberta, publicado no Diário da República, 2.ª série, n.º 215, de 6 de novembro de 2013.

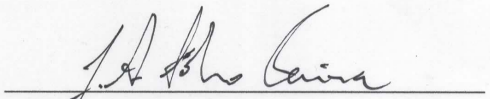
I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged Disciplinary Regulations of the Universidade Aberta (regulation published in the official journal Diário da República, 2.ª série, N.º 215, de 6 de novembro de 2013).

Universidade Aberta, 7 de agosto de 2024

Nome completo/Full name: José António Ferreira lobo Pereira

Assinatura/Signature:



manuscrita ou digital / handwritten or digital

Imp 12-26_A00

Resumo

Os levantamentos epidemiológicos de saúde periodontal exigem um tempo de exame extenso quando realizados através de avaliações completas da boca, o que sobrecarrega participantes e examinadores. Para aliviar isso, utiliza-se exames parciais da boca, omitindo intencionalmente alguns dados. No entanto, esses métodos podem levar a estimativas enviesadas de prevalência. Esta dissertação aborda essa questão aplicando métodos IA-fuzzy para imputação de dados ausentes em levantamentos epidemiológicos com avaliações unilaterais das arcadas dentárias. Usando técnicas avançadas de machine learning, esta investigação visa melhorar a predição de valores omitidos intencionalmente, garantindo estimativas de prevalência não enviesadas.

A revisão da literatura destaca a complexidade da periodontite e as limitações dos métodos tradicionais de imputação de dados na gestão de dados ausentes. O estudo utiliza dados de exames periodontais do *National Health and Nutrition Examination Survey* (NHANES) e emprega métodos estatísticos avançados para avaliar a simetria dentária. O método de imputação Mãe-Filha (MoDau), baseado em modelos XGBoost, foi desenvolvido para aprimorar a qualidade dos dados imputados.

Os resultados indicam que a função de medição de simetria (SM) avalia eficazmente a simetria da profundidade de sondagem periodontal (pocket probing depth, PPD), mostrando forte correlação com as avaliações profissionais. O método Modau exibe uma precisão preditiva robusta, especialmente para categorias mais altas de PPD, embora apresente discrepâncias nos valores mais baixos.

Este estudo ressalta a utilidade clínica e epidemiológica da função SM e o potencial dos métodos *Artificial Inteligence-fuzzy* (AI-fuzzy) para melhorar a imputação de dados. Pesquisas futuras devem explorar a aplicação desses métodos em outras áreas odontológicas e integrar técnicas adicionais de inteligência artificial (IA) para refinar ainda mais os modelos preditivos.

Palavras chave: Periodontal, Dados Ausentes Planeados, Imputação de Dados, Inteligência Artificial, Simetria Difusa, Simetria Oral.

Abstract

Periodontal health surveys require extensive examination time when conducting full-mouth assessments, which can burden participants and examiners. To alleviate this, partial-mouth examinations are employed, intentionally omitting some data. However, these methods can lead to biased prevalence estimates. This dissertation addresses this issue by applying neuro-fuzzy methods for imputing missing data in epidemiological surveys with unilateral dental arch assessments. Using advanced machine learning techniques, the research aims to improve the prediction of intentionally omitted values, ensuring unbiased prevalence estimates.

The literature review highlights the complexity of periodontitis and the limitations of traditional data imputation methods in handling missing data. Neuro-fuzzy techniques offer a more adaptive and precise approach, learning intrinsic patterns in the data for better imputation. The study utilizes data from National Health and Nutrition Examination Survey (NHANES) periodontal examinations and employs advanced statistical methods to assess dental symmetry. The Mother-Daughter (MoDau) imputation method, based on XGBoost models, was developed to enhance the quality of imputed data.

Results indicate that the symmetry measurement function (SM) effectively evaluates periodontal probing depth (PPD) symmetry, showing strong correlation with professional evaluations. The Daughter models exhibit robust predictive accuracy, especially for higher pocket probing depth (PPD) categories, though discrepancies in lower PPD values suggest room for improvement. Hyperparameter tuning significantly enhanced model performance, and introducing a noise level of 0.4 in data simulation ensured realistic model testing.

This study underscores the clinical and epidemiological utility of the SM function and the potential of artificial intelligence-fuzzy (AI-fuzzy) methods to improve data imputation. Future research should explore the application of these methods in other dental fields and integrate additional AI techniques to refine predictive models further.

Keywords: Periodontal, Planed Missing Data, Data Imputation, Artificial intelligence, Fuzzy Symmetry, Oral Symmetry.

Resumo Extenso

Introdução

Os levantamentos epidemiológicos da saúde periodontal, quando realizados através de exame dentário completo, necessitam de mais tempo de exame (ou de mais examinadores treinados), além de constituírem uma sobrecarga para os participantes. Para reduzir o tempo necessário e a carga sobre o participante, utiliza-se o exame oral parcial, que, pela sua natureza, omite uma quantidade planejada de dados. No entanto, está demonstrado que os resultados obtidos por estes métodos conduzem a vieses na prevalência de casos, ao contrário do exame dentário completo, cuja prevalência de casos depende apenas da forma como estes são definidos. Esta dissertação aborda este problema ao aplicar métodos *Artificial Inteligence-fuzzy* (AI-fuzzy) para a imputação de dados ausentes em levantamentos epidemiológicos com avaliação unilateral das arcadas dentárias. Utilizando técnicas avançadas de machine learning, a investigação visa melhorar a predição dos valores intencionalmente omitidos, permitindo uma prevalência de casos não enviesada.

Revisão da Literatura

A revisão da literatura começa com uma descrição sumária da periodontite, uma doença inflamatória que afeta os tecidos de suporte dos dentes. A complexidade dessa condição, juntamente com a variabilidade dos protocolos de exame periodontal, torna o processamento de dados ausentes um desafio considerável. Métodos tradicionais de imputação de dados, como imputação da média e regressão, são amplamente utilizados, mas possuem limitações significativas, especialmente em cenários cujos mecanismos de omissão são complexos. Técnicas de machine learning, incluindo métodos IA-fuzzy, oferecem uma abordagem mais adaptativa e precisa para lidar com dados ausentes. Estas técnicas podem aprender padrões intrínsecos nos dados, melhorando a imputação em comparação com métodos tradicionais. A simetria dentária também é explorada como informação complementar no processo de imputação.

Material e Métodos

Os dados utilizados no estudo foram obtidos a partir da base do *National Health and Nutrition Examination Survey* (NHANES) através de exames periodontais completos, permitindo uma análise abrangente e detalhada. Foram aplicados métodos estatísticos avançados para avaliar a simetria dentária, e desenvolvida e utilizada uma função de quantificação da simetria (SM) e estimativas de densidade de kernel (kernel density estimates, KDE). A preparação dos dados para a modelação por XGBoost envolveu imputação prévia pelo método hot-deck.

A imputação Mãe-Filha (MoDau) foi desenvolvida para melhorar a qualidade dos dados imputados, com base em dois modelos XGBoost consecutivos (modelos Mãe e Filha). Em todos os passos do processo em que os dados foram manipulados, foi avaliada a sua conformidade com as suas características pré-manipulação, para garantir as qualidades originais dos dados. Para garantir a robustez dos modelos, foi realizada uma afinação extensiva dos hiperparâmetros, que é um processo crucial para otimizar o desempenho dos modelos de machine learning. A simulação de dados com diferentes níveis de ruído foi utilizada para testar a eficácia dos modelos, assegurando que eles pudessem lidar com a variabilidade inerente dos dados clínicos.

Resultados Os resultados indicam que a função de medição de simetria (SM) é eficaz na avaliação da simetria da PPD. A forte correlação com as avaliações dos profissionais de medicina dentária valida a sua precisão e potencial para aplicação clínica. Os modelos Filha, especialmente em categorias mais altas de profundidade à sondagem (pocket probinh depth, PPD), exibem uma precisão preditiva robusta. No entanto, há algumas discrepâncias em valores mais baixos de PPD , sugerindo a necessidade de melhorias adicionais nos modelos, porém sem enviesamento significativo na estimação das prevalências de casos.

A afinação dos hiperparâmetros melhorou significativamente o desempenho dos modelos XGBoost. Esta abordagem metodológica, juntamente com a validação cruzada, assegura que os modelos sejam robustos e generalizáveis, capazes de fornecer previsões precisas. A introdução de um nível de ruído de 0.4 na simulação de dados de PPD permitiu gerar dados realistas, que capturam a variabilidade inerente da PPD e assim permitem testar o método Modau.

Discussão

Os achados deste estudo têm várias implicações importantes. Primeiro, a validação da função SM destaca a sua utilidade como uma ferramenta clínica e epidemiológica. A forte correlação com as avaliações profissionais sugere que esta ferramenta pode ser integrada na prática clínica para melhorar os diagnósticos periodontais.

Em segundo lugar, os modelos Filha desenvolvidos no estudo demonstram uma precisão preditiva robusta, particularmente em categorias mais altas de PPD. Isto é crucial, pois essas categorias são clinicamente significativas e exigem intervenções de saúde pública e clínicas mais assertivas. No entanto, as discrepâncias observadas em valores mais baixos de PPD indicam que há espaço para melhorias, possivelmente

através de um maior refinamento dos modelos e da incorporação de mais dados de treino.

A afinação dos hiperparâmetros é um processo meticuloso que permite melhorar significativamente o desempenho dos modelos de machine learning, assegurando que eles possam fornecer previsões precisas e confiáveis. A inclusão de medidas de simetria, especialmente a medida de simetria direcional (γSM), aumentou a precisão preditiva e a confiabilidade dos modelos. Este achado sublinha a importância de considerar a simetria dentária na modelação preditiva de dados periodontais.

A introdução de ruído na simulação de dados foi uma estratégia eficaz para testar a robustez dos modelos. O nível de ruído de 0.4 foi escolhido para capturar a variabilidade inerente dos dados clínicos, garantindo que os modelos fossem testados em condições que refletem a realidade clínica. Esta abordagem assegura que os modelos sejam robustos e possam lidar com a variabilidade dos dados reais. A análise da importância das características nos modelos preditivos revelou que as medidas de simetria direcional são críticas para o desempenho dos modelos. Isso sugere que futuros estudos devem continuar a explorar e refinar essas características para melhorar ainda mais a precisão preditiva. A geração de dados sintéticos com ruído controlado também se mostrou uma estratégia eficaz, especialmente quando há insuficiência de dados reais. Esta abordagem garante que os dados gerados mantenham a plausibilidade da distribuição, facilitando comparações precisas entre os dados simulados e os dados reais.

Conclusões

As conclusões retiradas do nosso trabalho foram as seguintes:

1. Validação da Função SM: A função de medição de simetria (SM) foi validada como uma ferramenta eficaz para avaliar a simetria da PPD, com forte correlação com pontuações profissionais.

2. Desempenho dos Modelos Filha: Os modelos Filha mostraram uma precisão preditiva robusta, especialmente em categorias mais altas de PPD, destacando a eficácia dos métodos neuro-fuzzy.

3. Importância das Medidas de Simetria: A incorporação de medidas de simetria, particularmente a medida de simetria direcional (γSM), aumentou a precisão preditiva e a confiabilidade dos modelos.

4. Acurácia dos Métodos de Imputação: O método de imputação de *hot-deck* (H-D) preservou efetivamente as propriedades estatísticas originais dos dados de PPD.

5. Geração de Dados Sintéticos: A geração de dados sintéticos com ruído controlado é um método eficaz quando há falta de dados reais para testar o método MoDau, garantindo a plausibilidade da distribuição.

6. Aplicabilidade Epidemiológica: Os modelos desenvolvidos têm forte potencial para incorporar métodos de imputação que usem avaliação oral parcial unilateral em levantamentos epidemiológicos periodontais.

7. Robustez Estatística: O uso de métodos estatísticos como a modelação GAMLSS, a validação cruzada e métodos de bootstrap para comparações de distribuições assegura a robustez das avaliações dos modelos.

8. Importância das Características nos Modelos Preditivos: A análise da importância das características revela que medidas de simetria direcional são críticas para o desempenho preditivo dos modelos periodontais.

Implicações e Estudos Futuros

As descobertas desta investigação têm implicações significativas para a prática clínica e epidemiológica em periodontologia. A validação da função SM e a eficácia dos modelos Filha sugerem que os métodos machine learning-fuzzy podem melhorar a predição de dados omissos, permitindo a obtenção de estimativas não enviesadas de prevalências. Futuras investigações podem explorar a aplicação destes métodos a outras áreas da medicina dentária e investigar a integração de outras técnicas de IA para aprimorar ainda mais os modelos preditivos.

A aplicação de técnicas de IA, como a modelação XGBoost, e a imputação hot deck mostrou-se eficaz na gestão de dados ausentes do tipo MAR. A robustez e a generalizabilidade dos modelos desenvolvidos neste estudo fornecem uma base sólida para futuras investigações e desenvolvimentos. Além disso, a incorporação de medidas de simetria na avaliação periodontal destaca a importância de uma abordagem integrada e multidimensional para a análise de dados clínicos.

Futuros estudos podem focar em aprimorar ainda mais os modelos desenvolvidos, explorando diferentes técnicas de inteligência artificial em combinação com métodos "crisp" e/ou "fuzzy". A investigação de métodos de imputação mais avançados e a integração de dados periodontais não utilizados neste estudo, bem como dados demográficos, também podem contribuir para melhorar a precisão e a validade das previsões. A expansão da aplicação dos métodos neuro-fuzzy para outras áreas da odontologia pode revelar novas oportunidades para melhorar a gestão de dados e os cuidados clínicos.

Conclusão

Esta dissertação demonstra a viabilidade e a eficácia dos métodos IA-fuzzy na gestão de dados ausentes do tipo MAR em Periodontologia. A aplicação de técnicas de imputação avançadas de IA, como a modelação XGBoost e a imputação de hot deck, melhora a predição em dados omissos planejados. As conclusões destacam a importância da simetria dentária na avaliação periodontal e o potencial dos métodos desenvolvidos para aplicação clínica e epidemiológica em investigações futuras. As descobertas e metodologias apresentadas nesta investigação fornecem uma base sólida para avanços contínuos na área, contribuindo para a melhoria da qualidade nos métodos de imputação. A integração de técnicas avançadas de machine learning e a validação rigorosa das metodologias desenvolvidas sublinham o compromisso com a excelência científica.

Palavras chave: Periodontal, Dados Ausentes Planeados, Imputação de Dados, Inteligência Artificial, Simetria Difusa, Simetria Oral.

Contents

List of Abbreviations, Acronyms, and Symbols **xxviii**

1 Introduction **1**

1.1 Context and Rationale 1

2 Literature Review **7**

2.1 Case Definition of Periodontitis 7

2.2 Missing Data - An Overview 8

2.2.1 Missing Data Challenges 9

2.2.2 Traditional Imputation Methods 11

2.2.3 Advanced Machine Learning-Based 11

2.2.4 Characterization of Missing Data 12

2.3 Dealing with Missing Data 15

2.3.1 Missing Data Handling Methods 15

2.4 Imputation Techniques 19

2.4.1 Artificial Intelligence - Machine Learning 25

2.5 Periodontal Examination Protocols:
Full-Mouth and Partial-Mouth Approaches 30

2.6 Symmetry Assumption 32

2.6.1 Symmetry of the Mouth 32

2.6.2 Research Question 37

3 Material and Methods **38**

3.1 Data 38

3.2 Assessing Symmetry Through Statistical Methods 41

3.2.1 Basic Methods 42

3.2.2 Advanced Methods 49

3.2.3 Symmetry Measure Function 54
3.3 Visualization of Similarity - Kernel Density Estimates 58
3.3.1 Kernel Density Estimation (KDE) 59
3.3.2 Optimization of the Kernel Parameters 63
3.3.2.1 Boostrap Method 64
3.4 Clinical Perception of Symmetry 65
3.5 Mother-Daughter Imputation Methodology 66
3.5.1 Imputation with Hot Deck 68
3.6 Mother-Daughter Models 70
3.6.1 Mother Models 72
3.6.2 Validation of Mother-Daughter Imputation 74
3.6.3 Assessing Daughter Models Imputation Results 76

4 Results **79**
4.1 Symmetry Assessment 79
4.1.1 Upper Central Incisors (11,21) 79
4.1.2 Upper Lateral Incisors (12,22) 90
4.1.3 Upper Canines (13,23) 94
4.1.4 Upper First Pre Molars (14, 24) 97
4.1.5 Upper Second Premolars (15, 25) 99
4.1.6 Upper First Molars (16, 26) 102
4.1.7 Upper Second Molars (17, 27) 105
4.2 Symmetry Measure 108
4.2.1 Numerical Assessment 108
4.2.2 Visual Assessment: Bar Plots and Kernel Density Estimates 110
4.2.3 Clinical perception of symmetry 114
4.3 Data preparation for MoDau 114
4.3.1 Hot Deck Imputation 114
4.4 Modau 139
4.4.1 Simulated Data 139
4.4.1.1 Overall analysis of kernels depicting simulated (noise = 0.4) and original 145
4.5 Mother-Daughter Models Imputation 145
4.5.1 Tooth 21 - MoDau Imputation 145
4.5.2 Tooth 22 - MoDau Imputation 159
4.5.3 Tooth 23 - MoDau Imputation 161

4.5.4 Tooth 24 - MoDau Imputation 163
4.5.5 Tooth 25 - MoDau Imputation 165
4.5.6 Tooth 26 - MoDau Imputation 167
4.5.7 Tooth 27 - MoDau Imputation 169

5 Discussion **176**
5.1 Introduction 176
5.2 Assessing symmetry 176
5.3 MoDau Imputation 179

6 Conclusion **189**

Bibliography **191**

I Appendix: Assessing Symmetry **1**
I.1 Upper Central Incisors - 11, 21 2
I.1.1 Distal Vestibular Sites 3
I.1.2 Vestibular Sites 4
I.1.3 Mesial Vestibular Sites 5
I.1.4 Distal Lingual Sites 6
I.1.5 Lingual Sites 7
I.1.6 Mesial Lingual Sites 8
I.2 Upper Lateral Incisors - 12, 22 9
I.2.1 Distal Vestibular Sites 10
I.2.2 Vestibular Sites 11
I.2.3 Mesial Vestibular Sites 12
I.2.4 Distal Lingual Sites 13
I.2.5 Lingual Sites 14
I.2.6 Mesial Lingual Sites 15
I.3 Upper Canines - 13, 23 16
I.3.1 Distal Vestibular Sites 17
I.3.2 Vestibular Sites 18
I.3.3 Mesial Vestibular Sites 19
I.3.4 Distal Lingual Sites 20
I.3.5 Lingual Sites 21
I.3.6 Mesial Lingual Sites 22
I.4 Upper First Premolar - 14, 24 23

I.4.1 Distal Vestibular Sites 24
I.4.2 Vestibular Sites 25
I.4.3 Mesial Vestibular Sites 26
I.4.4 Distal Lingual Sites 27
I.4.5 Lingual Sites 28
I.4.6 Mesial Lingual Sites 29
I.5 Upper Second Premolars - 15, 25 30
I.5.1 Distal Vestibular Sites 31
I.5.2 Vestibular Sites 32
I.5.3 Mesial Vestibular Sites 33
I.5.4 Distal Lingual Sites 34
I.5.5 Lingual Sites 35
I.5.6 Mesial Lingual Sites 36
I.6 Upper First Molars - 16, 26 37
I.6.1 Distal Vestibular Sites 38
I.6.2 Vestibular Sites 39
I.6.3 Mesial Vestibular Sites 40
I.6.4 Distal Lingual Sites 41
I.6.5 Lingual Sites 42
I.6.6 Mesial Lingual Sites 43
I.7 Upper Second Molars - 17, 27 44
I.7.1 Distal Vestibular Sites 45
I.7.2 Vestibular Sites 46
I.7.3 Mesial Vestibular Sites 47
I.7.4 Distal Lingual Sites 48
I.7.5 Lingual Sites 49
I.7.6 Mesial Lingual Sites 50
I.8 Descriptive SM for Pairs of Contralateral Sites of Upper Teeth 51
I.9 Upper Central Incisors (11, 21) 52
I.10 Upper Lateral Incisors (12, 22) 53
I.11 Upper Canines (13, 23) 54
I.12 Upper First Premolars (14, 24) 55
I.13 Upper Second Premolars (15, 25) 56
I.14 Upper First Molars (16, 26) 57
I.15 Upper Second Molars (17, 27) 58

II Appendix: Hot Deck Imputation **1**
II.1 Upper Right Central Incisor (11) 1
II.2 Upper Right Lateral Incisor (12) 5
II.3 Upper Right Canine (13) 8
II.4 Upper Right First Pre Molar (14) 11
II.5 Upper Right Second Premolar (15) 14
II.6 Upper Right First Molar (16) 17
II.7 Upper Right Second Molar (17) 20
II.8 Upper Left Central Incisor (21) 23
II.9 Upper Left Lateral Incisor (22) 26
II.10 Upper Left Canine (23) 29
II.11 Upper Left First Premolar(24) 32
II.12 Upper Left Second PremMolar (25) 35
II.13 Upper Left First Molar (26) 38
II.14 Upper Left Second Molar (27) 41

III Appendix: Comparison of Original versus Simulated **1**
III.1 Upper Central Incisors - Original *vs* Simulated 2
III.2 Upper Lateral Incisors - Original *vs* Simulated 5
III.3 Upper Canines - Original *vs* Simulated 8
III.4 Upper First Premolar - Original *vs* Simulated 11
III.5 Upper Second Premolars - Original *vs* Simulated 14
III.6 Upper First Molars - Original *vs* Simulated 17
III.7 Upper Second Molars - Original *vs* Simulated 20

IV Appendix: Mother-Daughter Method Imputation Results **1**
IV.1 Upper Left Central Incisor 2
IV.2 Upper Lateral Incisor 13
IV.3 Upper Left Canine 23
IV.4 Upper First Premolar 33
IV.5 Upper Second Premolar 43
IV.6 Upper First Molar 53
IV.7 Upper Second Molar 63

V Appendix V: R Scripts **1**
V.0.1MoDau IMputation 6

List of Figures

2.1 Dental formula 33
2.2 Representation of dental arch symmetry 34

3.1 Representation Symmetry Assessment 41
3.1 Buterfly Bar Plot for Symmetry Score 66
3.2 Diagram of MoDau Imputation Method 78

I.1 Upper Central Incisors Percentages of SM Values by Site 110
I.2 KDE of Upper Central Incisors by Site 111
II.1 Kernel Density Estimates of Imputed and Original 11 by Site 120
II.2 Kernel Density Estimates of Imputed and Original 12 by Site 122
4.2 Kernel Density Estimates of Imputed and Original 21 by Site 139
III.1 Kernel Density Plot of Simulated and Original 11DV, 21DV (Noise: 0.4) 142
III.2 Kernel Density Plot of Simulated and Original 11V, 21V (Noise: 0.4) . 142
III.2 Kernel Density Plot of Simulated and Original 11MV, 21MV (Noise: 0.4) 143
III.4 Kernel Density Plot of Simulated and Original 11DL, 21DL (Noise: 0.4) 143
III.5 Kernel Density Plot of Simulated and Original 11L, 21L (Noise: 0.4) . . 144
III.6 Kernel Density Plot of Simulated and Original 11ML, 21ML (Noise: 0.4) 144
III.7 Histograms of Simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 149
III.9 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 149
IV.5 Histograms of Simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 151
III.13 Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 152

III.15 Histograms of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm 153
III.17 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 154
III.8 Optimal Kernel Density Plots of 21DV for Simulated and Predicted Data 156
III.10 Optimal Kernel Density Plots of 21V for Simulated and Daughter Predicted 156
III.12 Optimal Kernel Density Plots of 21MV for Simulated and Daughter Predicted 157
III.14 Optimal Kernel Density Plots of 21DL for Simulated and Daughter Predicted 157
III.16 Optimal Kernel Density Plots of 21L for Simulated and Daughter Predicted 158
III.18 Optimal Kernel Density Plots of 21ML for Simulated and Daughter Predicted 158

I.1 Upper Central Incisors Percentages of SM Values by Site 52
I.2 Kernel Density Plots Estimates of Upper Central Incisors by Site . . . 52
I.3 Upper Lateral Incisors Percentages of SM Values by Site 53
I.4 Kernel Density Plots Estimates of Upper Lateral Incisors by Site . . . 53
I.5 Upper Canines Percentages of SM Values by Site 54
I.6 Kernel Density Plots Estimates of Upper Canines by Site 54
I.7 Upper First Pre Molars Percentages of SM Values by Site 55
I.8 Kernel Density Plots Estimates of Upper First Pre Molars by Site . . 55
I.9 Upper Second Premolars Percentages of SM Values by Site 56
I.10 Kernel Density Plots Estimates of Upper Second Pre Molars by Site . 56
I.11 Upper First Molars Percentages of SM Values by Site 57
I.12 Kernel Density Plots Estimates of Upper First Molars by Site 57
I.13 Upper Second Molars Percentages of SM Values by Site 58
I.14 Kernel Density Plots Estimates of Upper Second Molars by Site . . . 58

II.1 Kernel Density Plot of Imputed and Original 11 by Site 4
II.2 Kernel Density Plot of Imputed and Original 12 by Site 7
II.3 Kernel Density Plot of Imputed and Original 13 by Site 10
II.4 Kernel Density Plot of Imputed and Original 14 by Site 13
II.5 Kernel Density Plot of Imputed and Original 15 by Site 16

II.6 Kernel Density Plot of Imputed and Original 16 by Site 19
II.7 Kernel Density Plot of Imputed and Original 17 by Site 22
II.8 Kernel Density Plot of Imputed and Original 21 by Site 25
II.9 Kernel Density Plot of Imputed and Original 22 by Site 28
II.10 Kernel Density Plot of Imputed and Original 23 by Site 31
II.11 Kernel Density Plot of Imputed and Original 24 by Site 34
II.12 Kernel Density Plot of Imputed and Original 25 by Site 37
II.13 Kernel Density Plot of Imputed and Original 26 by Site 40
II.14 Kernel Density Plot of Imputed and Original 27 by Site 43

III.1 Kernel Density Plot of Simulated and Original 11DV, 21DV (Noise: 0.4) 2
III.2 Kernel Density Plot of Simulated and Original 11V, 21V (Noise: 0.4) . 2
III.3 Kernel Density Plot of Simulated and Original 11MV, 21MV (Noise: 0.4) 3
III.4 Kernel Density Plot of Simulated and Original 11DL, 21DL (Noise: 0.4) 3
III.5 Kernel Density Plot of Simulated and Original 11L, 21L (Noise: 0.4) . . 4
III.6 Kernel Density Plot of Simulated and Original 11ML, 21ML (Noise: 0.4) 4
III.7 Kernel Density Plot of Simulated and Original 12DV, 22DV (Noise: 0.4) 5
III.8 Kernel Density Plot of Simulated and Original 12V, 22V (Noise: 0.4) . 5
III.9 Kernel Density Plot of Simulated and Original 12MV, 22MV (Noise: 0.4) 6
III.10 Kernel Density Plot of Simulated and Original 12DL, 22DL (Noise: 0.4) 6
III.11 Kernel Density Plot of Simulated and Original 12L, 22L (Noise: 0.4) . 7
III.12 Kernel Density Plot of Simulated and Original 12ML, 22ML (Noise: 0.4) 7
III.13 Kernel Density Plot of Simulated and Original 13DV, 23DV (Noise: 0.4) 8
III.14 Kernel Density Plot of Simulated and Original 13V, 23V (Noise: 0.4) . 8
III.15 Kernel Density Plot of Simulated and Original 13MV, 23MV (Noise: 0.4) 9
III.16 Kernel Density Plot of Simulated and Original 13DL, 23DL (Noise: 0.4) 9
III.17 Kernel Density Plot of Simulated and Original 13L, 23L (Noise: 0.4) . 10
III.18 Kernel Density Plot of Simulated and Original 13ML, 23ML (Noise: 0.4) 10
III.19 Kernel Density Plot of Simulated and Original 14DV, 24DV (Noise: 0.4) 11
III.20 Kernel Density Plot of Simulated and Original 14V, 24V (Noise: 0.4) . 11
III.21 Kernel Density Plot of Simulated and Original 14MV, 24MV (Noise: 0.4) 12
III.22 Kernel Density Plot of Simulated and Original 14DL, 24DL (Noise: 0.4) 12
III.23 Kernel Density Plot of Simulated and Original 14L, 24L (Noise: 0.4) . 13
III.24 Kernel Density Plot of Simulated and Original 14ML, 24ML (Noise: 0.4) 13
III.25 Kernel Density Plot of Simulated and Original 15DV, 25DV (Noise: 0.4) 14
III.26 Kernel Density Plot of Simulated and Original 15V, 25V (Noise: 0.4) . 14

III.27 Kernel Density Plot of Simulated and Original 15MV, 25MV (Noise: 0.4) 15
III.28 Kernel Density Plot of Simulated and Original 15DL, 25DL (Noise: 0.4) 15
III.29 Kernel Density Plot of Simulated and Original 15L, 25L (Noise: 0.4) . 16
III.30 Kernel Density Plot of Simulated and Original 15ML, 25ML (Noise: 0.4) 16
III.31 Kernel Density Plot of Simulated and Original 16DV, 26DV (Noise: 0.4) 17
III.32 Kernel Density Plot of Simulated and Original 16V, 26V (Noise: 0.4) . 17
III.33 Kernel Density Plot of Simulated and Original 16MV, 26MV (Noise: 0.4) 18
III.34 Kernel Density Plot of Simulated and Original 16DL, 26DL (Noise: 0.4) 18
III.35 Kernel Density Plot of Simulated and Original 16L, 26L (Noise: 0.4) . 19
III.36 Kernel Density Plot of Simulated and Original 16ML, 26ML (Noise: 0.4) 19
III.37 Kernel Density Plot of Simulated and Original 17DV, 27DV (Noise: 0.4) 20
III.38 Kernel Density Plot of Simulated and Original 17V, 27V (Noise: 0.4) 20
III.39 Kernel Density Plot of Simulated and Original 17MV, 27MV (Noise: 0.4) 21
III.40 Kernel Density Plot of Simulated and Original 17DL, 27DL (Noise: 0.4) 21
III.41 Kernel Density Plot of Simulated and Original 17L, 27L (Noise: 0.4) . 22
III.42 Kernel Density Plot of Simulated and Original 17ML, 27ML (Noise: 0.4) 22

IV.1 Histograms of Simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 4
IV.2 Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 5
IV.3 Histograms of Simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 6
IV.4 Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 7
IV.5 Histograms of Simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 8
IV.6 Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm 9
IV.7 Optimal Kernel Density Plots of 21DV for Simulated and Predicted Data 10
IV.8 Optimal Kernel Density Plots of 21V for Simulated and Daughter Predicted 11
IV.9 Optimal Kernel Density Plots of 21MV for Simulated and Daughter Predicted 11

IV.10 Optimal Kernel Density Plots of 21DL for Simulated and Daughter Predicted 12
IV.11 Optimal Kernel Density Plots of 21L for Simulated and Daughter Predicted 12
IV.12 Optimal Kernel Density Plots of 21ML for Simulated and Daughter Predicted 12
IV.13 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 17
IV.14 Optimal Kernel Density Plots of 22DV for Simulated and Daughter Predicted 18
IV.15 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 18
IV.16 Optimal Kernel Density Plots of 22V for Simulated and Daughter Predicted 19
IV.17 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 19
IV.18 Optimal Kernel Density Plots of 22MV for Simulated and Daughter Predicted 20
IV.19 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 20
IV.20 Optimal Kernel Density Plots of 22DL for Simulated and Daughter Predicted 21
IV.21 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 21
IV.22 Optimal Kernel Density Plots of 22L for Simulated and Daughter Predicted 22
IV.23 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 22
IV.24 Optimal Kernel Density Plots of 22ML for Simulated and Daughter Predicted 23
IV.25 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 27
IV.26 Optimal Kernel Density Plots of 23DV for Simulated and Daughter Predicted 28

IV.27 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 28
IV.28 Optimal Kernel Density Plots of 23V for Simulated and Daughter Predicted 29
IV.29 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 29
IV.30 Optimal Kernel Density Plots of 23MV for Simulated and Daughter Predicted 30
IV.31 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 30
IV.32 Optimal Kernel Density Plots of 23DL for Simulated and Daughter Predicted 31
IV.33 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 31
IV.34 Optimal Kernel Density Plots of 23L for Simulated and Daughter Predicted 32
IV.35 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 32
IV.36 Optimal Kernel Density Plots of 23ML for Simulated and Daughter Predicted 33
IV.37 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 37
IV.38 Optimal Kernel Density Plots of 24DV for Simulated and Daughter Predicted 38
IV.39 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 38
IV.40 Optimal Kernel Density Plots of 24V for Simulated and Daughter Predicted 39
IV.41 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 39
IV.42 Optimal Kernel Density Plots of 24MV for Simulated and Daughter Predicted 40
IV.43 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 40

IV.44 Optimal Kernel Density Plots of 24DL for Simulated and Daughter Predicted 41
IV.45 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 41
IV.46 Optimal Kernel Density Plots of 24L for Simulated and Daughter Predicted 42
IV.47 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 42
IV.48 Optimal Kernel Density Plots of 24ML for Simulated and Daughter Predicted 43
IV.49 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 47
IV.50 Optimal Kernel Density Plots of 25DV for Simulated and Daughter Predicted 48
IV.51 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 48
IV.52 Optimal Kernel Density Plots of 25V for Simulated and Daughter Predicted 49
IV.53 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 49
IV.54 Optimal Kernel Density Plots of 25MV for Simulated and Daughter Predicted 50
IV.55 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 50
IV.56 Optimal Kernel Density Plots of 25DL for Simulated and Daughter Predicted 51
IV.57 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 51
IV.58 Optimal Kernel Density Plots of 25L for Simulated and Daughter Predicted 52
IV.59 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 52
IV.60 Optimal Kernel Density Plots of 25ML for Simulated and Daughter Predicted 53

IV.61 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 57
IV.62 Optimal Kernel Density Plots of 26DV for Simulated and Daughter Predicted 58
IV.63 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 58
IV.64 Optimal Kernel Density Plots of 26V for Simulated and Daughter Predicted 59
IV.65 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 59
IV.66 Optimal Kernel Density Plots of 26MV for Simulated and Daughter Predicted 60
IV.67 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 60
IV.68 Optimal Kernel Density Plots of 26DL for Simulated and Daughter Predicted 61
IV.69 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 61
IV.70 Optimal Kernel Density Plots of 26L for Simulated and Daughter Predicted 62
IV.71 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 62
IV.72 Optimal Kernel Density Plots of 26ML for Simulated and Daughter Predicted 63
IV.73 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 67
IV.74 Optimal Kernel Density Plots of 27DV for Simulated and Daughter Predicted 68
IV.75 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 68
IV.76 Optimal Kernel Density Plots of 27V for Simulated and Daughter Predicted 69
IV.77 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 69

IV.78 Optimal Kernel Density Plots of 27MV for Simulated and Daughter Predicted 70
IV.79 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 70
IV.80 Optimal Kernel Density Plots of 27DL for Simulated and Daughter Predicted 71
IV.81 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 71
IV.82 Optimal Kernel Density Plots of 27L for Simulated and Daughter Predicted 72
IV.83 Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm 72
IV.84 Optimal Kernel Density Plots of 27ML for Simulated and Daughter Predicted 73

List of Tables

2.1 Universal dental enumerating system. 33
2.2 FDI dental enumerating system. 34

4.1 Summary of statistical tests comparing central incisors 11 and 21 PPD medians and variances across six dental sites; distances between distributions. 81
4.2 Assessing Side effect on PPD at Upper Central Incisors Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test 82
4.3 Statistical Comparison of Cross Validation Metrics for Models: PPD ∼ 1 and PPD ∼ Side 82
4.4 Assessing Side effect on PPD at Upper Central Incisors Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 83
4.5 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 84
4.6 Assessing Side effect on Pocket Probing Depth (PPD) at Upper Central Incisors Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 84
4.7 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 85
4.8 Assessing Side effect on PPD at Upper Central Incisors Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 86
4.9 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 87

4.10 Assessing Side effect on PPD at Upper Central Incisors Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 88
4.11 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 88
4.12 Assessing Side Effect on PPD at Upper Central Incisors mesial lingual (ML) Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test (LRT) with ex-Gaussian Distribution 89
4.13 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 90
4.14 Summary of SM values for upper contralateral teeth pairs across six dental sites. 109
4.15 Missing Data by Tooth Site in Upper Arch 115
4.16 Comparison of Original and Imputed Periodontal Pocket Depth Values Across Different Sites and Unique Value Categories (k) 117
4.17 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site . . 118
4.18 PPD Statistics Comparison: Before and After H-D Imputation by Site 119
4.19 Bootstrap Test for Kernel Density Estimates (KDE) Difference Across Noise Levels 141
4.20 Distinctive Characteristics of M21 Mother Models by Site 145
4.21 Performance Metrics for the M21 Mother Models by Site 146
4.22 Mother Models M21 Features Importance Metrics by Site 146
4.24 Performance Metrics by Site for D21 147
4.23 Distinctive Characteristics of the D21 by Site 147
4.25 Feature Importance Metrics of D21 by Site 148
4.26 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 149
4.27 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 150
4.28 Chi-squared Test Results for Comparison of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm 151
4.29 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 152
4.30 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 153

4.31 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 154
4.32 KDE Differences and Kolmogorov-Smirnov Test 156

I.1 Summary of statistical tests comparing central incisors 11 and 21 PPD medians and variances across six dental sites; distances between distributions. 2
I.2 Assessing Side effect on PPD at Upper Central Incisors Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test 3
I.3 Statistical Comparison of Cross Validation Metrics for Models: PPD ∼ 1 and PPD ∼ Side 3
I.4 Assessing Side effect on PPD at Upper Central Incisors Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 4
I.5 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 4
I.6 Assessing Side effect on PPD at Upper Central Incisors Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 5
I.7 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 5
I.8 Assessing Side effect on PPD at Upper Central Incisors Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 6
I.9 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 6
I.10 Assessing Side effect on PPD at Upper Central Incisors Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 7
I.11 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 7
I.12 Assessing Side Effect on PPD at Upper Central Incisors Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 8
I.13 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 8

I.14 Summary of statistical tests comparing lateral incisors 12 and 22 PPD medians and variances across six dental sites; distances between distributions 9
I.15 Assessing Side Effect on PPD at Upper Lateral Incisors Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution. 10
I.16 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 10
I.17 Assessing Side Effect on PPD at Upper Lateral Incisors Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 11
I.18 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 11
I.19 Assessing Side Effect on PPD at Upper Lateral Incisors Mesio Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 12
I.20 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 12
I.21 Assessing Side Effect on PPD at Upper Lateral Incisors Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 13
I.22 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 13
I.23 Assessing Side Effect on PPD at Upper Lateral Incisors Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 14
I.24 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 14
I.25 Assessing Side Effect on PPD at Upper Lateral Incisors Mesio Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 15
I.26 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 15
I.27 Summary of statistical tests comparing canines 13 and 23 PPD medians and variances across six dental sites; distances between distributions. 16

I.28 Assessing Side Effect on PPD at Upper Canines Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 17
I.29 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 17
I.30 Assessing Side Effect on PPD at Upper Canines Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 18
I.31 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 18
I.32 Assessing Side Effect on PPD at Upper Canines Mesio-Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 19
I.33 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 19
I.34 Assessing Side Effect on PPD at Upper Canines Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 20
I.35 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 20
I.36 Assessing Side Effect on PPD at Upper Canines Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 21
I.37 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 21
I.38 Assessing Side Effect on PPD at Upper Canines Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 22
I.39 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 22
I.40 Summary of Statistical Tests Comparing Upper First Premolar 14, 24 PPD medians and variances across six dental sites; distances between distributions 23
I.41 Assessing Side Effect on PPD at Upper First Premolar Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 24

I.42 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 24
I.43 Assessing Side Effect on PPD at Upper First Premolar Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 25
I.44 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 25
I.45 Assessing Side Effect on PPD at Upper First Premolar Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 26
I.46 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 26
I.47 Assessing Side Effect on PPD at Upper First Premolar Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 27
I.48 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 27
I.49 Assessing Side Effect on PPD at Upper First Premolar Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 28
I.50 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 28
I.51 Assessing Side Effect on PPD at Upper First Premolar Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 29
I.52 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 29
I.53 Summary of statistical tests comparing Upper Second Premolars 15 and 25 PPD medians and variances across six dental sites; distances between distributions. . . . 30
I.54 Assessing Side Effect on PPD at Upper Second Premolars Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 31
I.55 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 31

I.56 Assessing Side Effect on PPD at Upper Second Premolars Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 32
I.57 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 32
I.58 Assessing Side Effect on PPD at Upper Second Premolars Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 33
I.59 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 33
I.60 Assessing Side Effect on PPD at Upper Second Premolars Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 34
I.61 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 34
I.62 Assessing Side Effect on PPD at Upper Second Premolars Lingual Sites: Null Model vs Side Model Using Ex-Gaussian Distribution 35
I.63 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 35
I.64 Assessing Side Effect on PPD at Upper Second Premolars Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 36
I.65 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 36
I.66 Summary of Statistical Tests Comparing Upper First Molars 16 and 26 PPD Medians and Variances Across Six Dental Sites; Distances Between Distributions 37
I.67 Assessing Side Effect on PPD at Upper First Molars Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 38
I.68 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 38
I.69 Assessing Side Effect on PPD at Upper First Molars Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 39

I.70 Statistical Comparison of Cross Validation Metrics for Models PPD ∼ 1 and PPD ∼ Side . . . 39
I.71 Assessing Side Effect on PPD at Upper First Molars Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution . . . 40
I.72 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side . . . 40
I.73 Assessing Side Effect on PPD at Upper First Molars Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution . . . 41
I.74 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side . . . 41
I.75 Assessing Side Effect on PPD at Upper First Molars Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution . . . 42
I.76 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side . . . 42
I.77 Assessing Side Effect on PPD at Upper First Molars Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution . . . 43
I.78 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side . . . 43
I.79 of Statistical Tests Comparing Upper Second Molars 17 and 27 PPD Medians and Variances Across Six Dental Sites; Distances Between Distributions . . . 44
I.80 Assessing Side Effect on PPD at Upper Second Molars Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution . . . 45
I.81 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side . . . 45
I.82 Assessing Side Effect on PPD at Upper Second Molars Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution . . . 46
I.83 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side . . . 46

I.84 Assessing Side Effect on PPD at Upper Second Molars Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 47
I.85 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 47
I.86 Assessing Side Effect on PPD at Upper Second Molars Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 48
I.87 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 48
I.88 Assessing Side Effect on PPD at Upper Second Molars Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 49
I.89 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 49
I.90 Assessing Side Effect on PPD at Upper Second Molars Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution 50
I.91 Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side 50
I.92 Summary of SM values for upper contralateral teeth pairs across six dental sites. 51

II.1 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 2
II.2 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 3
II.3 PPD Statistics Comparison: Before and After H-D Imputation by Site 3
II.4 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 5
II.5 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 6
II.6 PPD Statistics Comparison: Before and After H-D Imputation by Site 6
II.7 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 8
II.8 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 9
II.9 PPD Statistics Comparison: Before and After H-D Imputation by Site 9

II.10 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 11
II.11 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 12
II.12 PPD Statistics Comparison: Before and After H-D Imputation by Site 12
II.13 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 14
II.14 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 15
II.15 PPD Statistics Comparison: Before and After H-D Imputation by Site 15
II.16 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 17
II.17 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 18
II.18 PPD Statistics Comparison: Before and After H-D Imputation by Site 18
II.19 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 20
II.20 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 21
II.21 PPD Statistics Comparison: Before and After H-D Imputation by Site 21
II.22 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 23
II.23 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site 24
II.24 PPD Statistics Comparison: Before and After H-D Imputation by Site 24
II.25 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 26
II.26 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site . . 27
II.27 PPD Statistics Comparison: Before and After H-D Imputation by Site 27
II.28 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 29
II.29 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site . . 30
II.30 PPD Statistics Comparison: Before and After H-D Imputation by Site 30
II.31 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 32
II.32 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site . . 33
II.33 PPD Statistics Comparison: Before and After H-D Imputation by Site 33
II.34 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 35
II.35 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site . . 36

II.36 PPD Statistics Comparison: Before and After H-D Imputation by Site 36
II.37 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 38
II.38 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site . . 39
II.39 PPD Statistics Comparison: Before and After H-D Imputation by Site 39
II.40 Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k) 41
II.41 Comparison of PPD > 3 Proportions: Original vs. Imputed by Site . . 42
II.42 PPD Statistics Comparison: Before and After H-D Imputation by Site 42

IV.1 Distinctive Characteristics of M21 Mother Models by Site 2
IV.2 Performance Metrics for the M21 Mother Models by Site 2
IV.3 Mother Models M21 Features Importance Metrics by Site 2
IV.4 Distinctive Characteristics of the D21 by Site 3
IV.5 Performance Metrics by Site for D21 3
IV.6 Feature Importance Metrics of D21 by Site 3
IV.7 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 4
IV.8 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 5
IV.9 Chi-squared Test Results for Comparison of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm 6
IV.10 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 7
IV.11 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 8
IV.12 Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site 9
IV.13 Kernel Density Estimates Differences and Kolmogorov-Smirnov Test . . 10
IV.14 Distinctive Characteristics of the M22 Mother Models by Site 13
IV.15 Performance Metrics for the M22 Mother Models by Site 13
IV.16 Mother Models features Importance Metrics - M22 by Site 13
IV.17 Distinctive Characteristics of the D22 Daughter Models by Site 14
IV.18 Performance Metrics for the Daughter Models - D22 Data 14
IV.19 Feature Importance Metrics Of Daughter Models D22 by Site 14

IV.20Chi-squared Test Results for proportions of PPD values ≥ 4mm comparisons between Original and predicted for Site 22 15
IV.21Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site 16
IV.22Kernel Density Estimates Differences and Kolmogorov-Smirnov Test . . 17
IV.23Distinctive Characteristics of the M23 Mother Models by Site 23
IV.24Performance Metrics for the M23 Mother Models by Site 24
IV.25Mother Models features Importance Metrics - M23 by Site 24
IV.26Distinctive Characteristics of the D32 Daughter Models by Site 24
IV.27Performance Metrics for the Daughter Models - D23 Data 25
IV.28Feature Importance Metrics Of Daughter Models D23 by Site 25
IV.29Chi-squared Test Results for proportions of PPD values ≥ 4mm comparisons between Original and predicted for Site 23 25
IV.30Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site 26
IV.31Kernel Density Estimates Differences and Kolmogorov-Smirnov Test . . 27
IV.32Distinctive Characteristics of the M24 Mother Models by Site 33
IV.33Performance Metrics for the M24 Mother Models by Site 34
IV.34Mother Models features Importance Metrics - M24 by Site 34
IV.35Distinctive Characteristics of the D42 Daughter Models by Site 34
IV.36Performance Metrics for the Daughter Models - D24 Data 35
IV.37Feature Importance Metrics of Daughter Models D24 by Site 35
IV.38Chi-squared Test Results for proportions of PPD values ≥ 4mm comparisons between Original and predicted for Site 24 35
IV.39Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site 36
IV.40Kernel Density Estimates Differences and Kolmogorov-Smirnov Test . . 37
IV.41Distinctive Characteristics of the M25 Mother Models by Site 43
IV.42Performance Metrics for the M25 Mother Models by Site 44
IV.43Mother Models features Importance Metrics - M25 by Site 44
IV.44Distinctive Characteristics of the D25 Daughter Models by Site 44
IV.45Performance Metrics for the Daughter Models - D25 Data 45
IV.46Feature Importance Metrics Of Daughter Models D25 by Site 45
IV.47Chi-squared Test Results for proportions of PPD values ≥ 4mm comparisons between Original and predicted for Site 25 45

IV.48Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site 46
IV.49Kernel Density Estimates Differences and Kolmogorov-Smirnov Test . . 47
IV.50Distinctive Characteristics of the M26 Mother Models by Site 53
IV.51Performance Metrics for the M26 Mother Models by Site 54
IV.52Mother Models features Importance Metrics - M26 by Site 54
IV.53Distinctive Characteristics of the D26 Daughter Models by Site 54
IV.54Performance Metrics for the Daughter Models - D26 Data 55
IV.55Feature Importance Metrics Of Daughter Models D26 by Site 55
IV.56Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted for Site 55
IV.57Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site 56
IV.58Kernel Density Estimates Differences and Kolmogorov-Smirnov Test . . 57
IV.59Distinctive Characteristics of the M27 Mother Models by Site 63
IV.60Performance Metrics for the M27 Mother Models by Site 64
IV.61Mother Models features Importance Metrics - M27 by Site 64
IV.62Distinctive Characteristics of the D27 Daughter Models by Site 64
IV.63Performance Metrics for the Daughter Models - D27 Data 65
IV.64Feature Importance Metrics Of Daughter Models D27 by Site 65
IV.65Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted by Site 27 65
IV.66Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site 66
IV.67Kernel Density Estimates Differences and Kolmogorov-Smirnov Test . . 67

List of Abbreviations, Acronyms, and Symbols

AE Autoencoder

AI Artificial Intelligence

ANN Artificial Neural Networks

AIC Akaike Information Criteria

BC Bhattacharyya Coefficient

BEST Branch-Exclusive Splits Trees

BP Back Propagation

CAL Clinical Attachment Loss

CCA Complete Case Analysis

CDC Centers for Disease Control and Prevention

CDF Cumulative Distribution Function

CNN Convolutional Neural Network

CPITN Community Periodontal Index of Treatment Needs

DA Data Augmentation

DBN Deep Belief Networks

DeLe Deep Learning

DL distal lingual

D-L Deep Learning

DV distal vestibular

DNN Deep Neural Networks

DRL Deep Reinforcement Learning

DT Decision Trees

EM Expectation-Maximisation algorithm

XGBoost eXtreme Gradient Boosting

FCS Fully Conditional Specification

FD Full Dataset

FIML Full Information Maximum Likelihood

FMPE Full-Mouth Periodontal Examination

f-NN fuzzy-Neural Network

GA Genetic Algorithm

GAM Generalised Additive Models

GAMLSS Generalized Additive Models for Location Scale and Shape

GLM Generalized Linear Models

GAN Generative Adversarial Network

GR Gingival Recession

GBM Gradient Boosting Machine

GRU Gated Recurrent Unit

H-D Hot Deck

IPW Inverse Probability Weighting

k-NN k-Nearest Neighbors

KDE Kernel Density Estimates

K-S Kolmogorov-Smirnov

L lingual

LOCF Last Observation Carried Forward

LRT Likelihood Ratio Test

LSTM Long Short Term Memory

LVCF Last Value Carried Forward

MAE Mean Absolute Error

MaLe Machine Learning

MAR Missing at Random type

MCAR Missing Completely at Random type

MI Multiple Imputation

MIA Missing Incorporated in Attribute

MICE Multivariate Imputation by Chained Equations

ML mesial lingual

MLE Maximum Likelihood Estimation

MLP Multilayer Perceptron

MNAR Missing Not at Random type

MoDau Mother-Daughter Imputation Method

MSE Mean Square Error

MV mesial vestibular

NHANES National Health and Nutrition Examination Surveys

NIDCR The National Institute of Dental and Craniofacial Research

NLP Natural Language Processing

NN Neural Network

PCA Principal Component Analysis

PMF Probability Mass Function

PMPE Partial-Mouth Periodontal Examination

PPD Pocket Probing Depth

PDF Probability Density Function

RHMP Random Half-Mouth

RBF Radial Basis Function

RICC Robustness Interpretability Completeness Correctness

RL Reinforcement Learning

RF Random Forest

RNN Recurrent Neural Network

RAR Robust Association Rules

RBM Restricted Boltzmann Machines

RMSE Root Mean Square Error

SL Supervised Learning

SD Symmetry Distance

SM Symmetry Measure

ST Symmetry Transformation

SOM Self-Organized Map

SRPI Similar Response Pattern Imputation

SSL Semi-Supervised Learning

SVM Support Vector Machines

SVR Support Vector for Regression

UL Unsupervised Learning

V vestibular

VAE Variational Autoencoders

VDeLe Very Deep Learning

WWCPPC 2017 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions

Chapter 1

Introduction

1.1 Context and Rationale

In health sciences research, handling missing data, particularly when the missingness is planned, is a critical issue that can significantly influence the validity and reliability of study outcomes. This dissertation, titled "Missing Data Handling in Health Sciences - A Neuro-Fuzzy Methods Approach," was initially proposed with the aim of exploring neuro-fuzzy methods for the imputation of planned missing data. Neuro-fuzzy methods were chosen for their theoretical potential to manage uncertainties and adaptively learn from data patterns, which seemed well-suited for this complex task.

As the research progressed, it became clear that while neuro-fuzzy methods hold promise, their application in the context of imputing planned missing data requires further development and refinement. Extensive evaluation and empirical testing revealed that these methods did not achieve the desired level of effectiveness. Consequently, the research shifted towards alternative techniques that demonstrated superior performance in this context.

Despite this methodological shift, the title of the dissertation has remained unchanged. This decision was made to preserve the original framework and intention of the research proposal. The title reflects the foundational aim of addressing missing data challenges in health sciences, specifically through an initially hypothesized neuro-fuzzy approach.

Adapting research methods in response to empirical evidence is a hallmark of rigorous scientific inquiry, ensuring that the research remains relevant and accurate. This dissertation documents the methodological evolution and justifies the shift from neuro-fuzzy methods to more effective alternatives for imputing planned missing data. The

findings underscore the need for further research and development in the application of neuro-fuzzy methods within this domain.

By maintaining the original title, the dissertation acknowledges the starting point of the research journey, while the content reflects the adaptive nature of scientific investigation. The discussions and findings provide valuable insights into the imputation of planned missing data, contributing to the broader discourse in health sciences and highlighting areas where neuro-fuzzy methods require further exploration and refinement.

In conclusion, this work represents a comprehensive exploration of missing data handling, demonstrating flexibility and responsiveness to new evidence. The title serves as a tribute to the initial research proposal, while the body of work showcases the evolved methodologies that ultimately offered more effective solutions. Moreover, it underscores the ongoing need for research into neuro-fuzzy methods, emphasizing their potential and the necessity for continued advancement in this area.

Periodontal disease is a prevalent and significant public health issue, affecting a large portion of the adult population worldwide. It is characterized by the destruction of the supporting structures of the teeth, including the gums, periodontal ligament, and alveolar bone. This condition can lead to tooth loss and is associated with systemic conditions such as cardiovascular disease and diabetes. Accurate diagnosis and monitoring of periodontal disease are critical for effective treatment and management. Therefore, surveillance of periodontal disease is of major importance.

One of the principal challenges in periodontal research is the issue of missing data. Missing data can arise from various factors, including patient non-compliance, measurement errors, limitations in the data collection process, and planned missingness. In periodontal studies, missing data is particularly problematic because it can lead to biased estimates and reduced statistical power, ultimately compromising the validity of the research findings. This concern is similarly significant when missing data is planned in periodontal surveys.

The missing data observed in periodontal research can be broadly classified into three categories: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Understanding these mechanisms is crucial for selecting appropriate methods to handle missing data. MCAR occurs when the probability of missing data is unrelated to both observed and unobserved data. MAR occurs when the probability of missing data is related to observed data but not to the missing data itself. MNAR occurs when the probability of missing data is related to the missing

data itself, posing significant challenges for data analysis and imputation.

An important characteristic of periodontal lesions is their tendency to exhibit a degree of symmetry. This means that the condition of one side of the dental arch can often provide information about the condition of the other side. Leveraging this symmetry can significantly enhance imputation methods for missing data in periodontal research. By assessing the symmetry of periodontal lesions, researchers can develop more accurate and reliable imputation methods that utilise the inherent structure of the data.

The primary objective of this dissertation is to develop and validate a new imputation method for handling missing data in periodontal surveys, aiming for the correct prediction of the prevalence of pathological contralateral values of PPD (PPD $>$ 3 mm), using the concept of fuzzy symmetry in periodontal lesions. The scope of this work includes assessing the symmetry of periodontal lesions and its quantification, and using this information to enhance a non-parametric modelling imputation method. The proposed method will be tested with a different dataset, and its results will be compared with the real data.

This dissertation is structured as follows:

1. **Literature Review**

- **Periodontal Disease**: Provides an overview of periodontal disease, its prevalence, risk factors, and clinical implications.

- **Missing Data - An Overview**: Discusses the challenges, issues, and characterisation of missing data, along with an extension on missing data mechanisms (Missing Completely at Random type (MCAR), Missing at Random type (MAR), Missing Not at Random type (MNAR)) and their assessment, furthermore the methods to transform MNAR to MAR will be addressed using of auxiliary variables, modelling the missing data process, and using instrumental variables.

- **Dealing with Missing Data**: Reviews various methods for handling missing data.

- **Periodontal Examination Protocols**: Describes full-mouth and partial-mouth examination protocols.

2. **Methodology**

- **Data**: Details the data collection process, including considerations on probing

pocket depth, data preparation, and the selection of individuals with full periodontal examination.

- **Assessing Symmetry**: Statistical methods for assessing symmetry in periodontal lesions will explored, including Wilcoxon Signed-Rank Test, Levene's Test, Kolmogorov-Smirnov Test, Bhattacharyya Coefficient, correlation coefficients, and gamlss models with respective comparison tests, this approach will completed with kernel plots and SM bar plots.

- **Imputation with "Hot Deck" Method**: Describes the algorithm for Hot Deck imputation and methods used to evaluate the results, including proportion tests and bootstrap tests.

- **Validation of the Imputation Method**: Explains the creation of a simulated data set for validation and the use of bootstrap tests for KDE.

This research contributes to the field of periodontal studies by providing a novel method for impute data in planned missing data surveys design, a mew methodology to reduce participant and researchers burden, time of examination, and consequently the operational cost. By leveraging the symmetry of periodontal lesions, the proposed imputation method aims to produce more reliable and accurate estimates, ultimately improving the quality of periodontal research and patient outcomes.

In conclusion, this dissertation addresses the critical issue in periodontal research by developing a new imputation method that utilises the inherent symmetry of periodontal lesions. Through a comprehensive literature review, detailed methodology, and rigorous validation, this research aims to provide a tool to impute data in half mouth periodontal exam protocol.

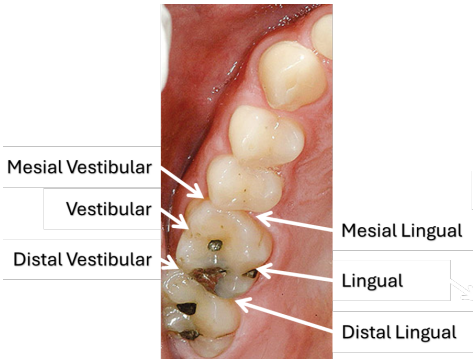
Periodontal diseases is defined as a group of multifactorial inflammatory diseases of supporting tissues of teeth, associated with bacterial dysbiosis. Its clinical most destructive form, periodontitis, leads to a progressive destruction of the periodontal ligament and alveolar bone, with periodontal pocket formation, gingival recession or both (M. G. Newman et al., 2006), which present a rough reflexive symmetry of its lesions and progression in both sides of the mouth (Mombelli & Meier, 2001). Together with dental caries are the two most important oral diseases (Bratthall et al., 2006), contributing to the global burden of chronic disease (Jamison, 2006). Periodontal disease with a prevalence of up 90% of the worldwide population periodontal disease became major public health problem (Pihlstrom et al., 2005), with the most severe

forms affecting 10.8% or 743 million people aged 15 – 99 worldwide (Butt et al., 2019) and a prevalence of up 90% when mild forms are included periodontal disease became major public health problem (Pihlstrom et al., 2005).

The basic features of periodontitis include the loss of periodontal tissue manifested through clinical attachment and radiographic bone loss, presence of periodontal pocketing, gingival bleeding, and (Fritz et al., 2018) is both site specific and episodic in nature (Mascarenhas, Okunseri, Dye, et al., 2020). The bone support loss is episodic and its rate per year of age allows for lifelong estimation.

The diagnosis and assessment of periodontitis severity are primarily conducted through clinical means, utilizing periodontal probing to measure clinical attachment loss Clinical Attachment Loss (CAL) and pocket depth, also known as periodontal pocket depth (PPD). This probing process targets six specific sites on each tooth to obtain comprehensive measurements. The sites on the tooth facing towards the face or lip are termed vestibular sites, while those facing towards the tongue are referred to as lingual sites. For each tooth, measurements are conducted at six distinct locations: one near the mouth's front or centerline (mesial), another towards the back of the mouth (distal), and a third site positioned approximately halfway between these two (vestibular or lingual) (Figure 1.1). These six probing sites are identified as mesial vestibular (MV), vestibular (V), distal vestibular (DV), ML, lingual (L), distal lingual (DL) sites, with the mesial and distal sites also known as interproximal (interdental) sites due to their proximity to the spaces between adjacent teeth (Mascarenhas, Okunseri, Dye, et al., 2020).

Figure 1.1



The six sites where PPD measures are taken

Periodontal pocket depth is measured from the gingival margin to the pocket's base, with a PPD exceeding 3 mm considered indicative of pathological conditions such as gingivitis or periodontitis. The presence of clinical attachment loss CAL and PPD, measured from the cement enamel junction to the base of the gingival sulcus or pocket, signifies either current or past instances of periodontitis. The severity of periodontitis is gauged by the highest CAL measurement, while its extent is determined by the percentage of probed sites (or teeth) affected. If more than 30% of teeth are involved, the condition is classified as generalised periodontitis; otherwise, it is considered localised. Bone level assessments can be directly obtained from x-ray images or indirectly inferred through CAL measurements. The progression rate of bone loss is calculated by dividing the percentage of bone loss by the patient's age in years, providing an indication of the disease's advancement speed.

Chapter 2

Literature Review

2.1 Case Definition of Periodontitis

The case definition of periodontitis has been evolving to encompass a variety of clinical conditions and new knowledge on aetiology and pathogenesis. The case definition adopted by the Centers for Disease Control and Prevention (CDC) in the National Health and Nutrition Examination Surveys (NHANES) has been widely used for epidemiologic surveys of periodontitis. Recently, a new definition of periodontitis case in both epidemiological and clinical care contexts was issued by the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions (2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions (WWCPPC 2017)) (Tonetti et al., 2018).

The relevance of case definition is intrinsically linked to the methodology of periodontal health assessment. The type of assessment, whether it is a Full-Mouth Periodontal Examination (FMPE) or a Partial-Mouth Periodontal Examination (PMPE), affects the value of prevalence estimators. Biases in prevalence estimations are dependent on the case definition used.

The CDC periodontal diseases classification system defines an individual as having mild periodontitis if there are ≥ 2 interproximal sites with CAL ≥ 3 mm, and ≥ 2 interproximal sites with PPD ≥ 4 mm (not on the same tooth) or one site with PPD ≥ 5 mm. Moderate periodontitis is defined by the presence of ≥ 2 interproximal sites with CAL ≥ 4 mm (not on the same tooth) or ≥ 2 interproximal sites with PPD ≥ 5 mm, also not on the same tooth. Severe periodontitis is diagnosed when there are ≥ 2 interproximal sites with CAL ≥ 6 mm (not on the same tooth) and ≥ 1 interproximal site(s) with PPD ≥ 5 mm. Participants with no evidence of mild, moderate, or severe

periodontitis are classified as having no periodontitis (P. I. Eke, Page, et al., 2012).

According to the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions classification framework, periodontitis diagnosis is positive if interdental CAL is detectable at $\geq$ 2 non-adjacent teeth, or buccal or oral CAL $\geq$ 3 mm with PPD $>$ 3 mm is detectable at $\geq$ 2 teeth. The severity and complexity are classified into four stages based on the severity (CAL, PPD, type of bone loss, and complexity factors).

In Stage I (mild disease), the PPD is $\leq$ 4 mm, with CAL $\leq$ 1–2 mm.

In Stage II (moderate disease), PPD is $\leq$ 5 mm, with CAL $\leq$ 3–4 mm.

In Stage III (severe disease), PPD is $\geq$ 6 mm, with CAL $\geq$ 5 mm.

In Stage IV (very severe disease), PPD is $\geq$ 6 mm, with CAL $\geq$ 2 mm (Tonetti et al., 2018).

The extent of periodontitis is related to the number of teeth affected, with $\leq$ 30% classified as localized and more than 30% as generalized. Staging and extent provide a comprehensive overview of the disease status throughout the mouth, relative to its severity and complexity. Correct staging of periodontitis is crucial for determining the type and complexity of treatment (maintenance, non-surgical or surgical), post-treatment tooth loss, and prognosis. Thus, accurate classification is vital not only for patient-level treatment planning and prognosis but also for public health planning, as it allows for the estimation of resources needed to treat a population effectively.

2.2 Missing Data - An Overview

Missing data, or values not recorded for certain variables in an observation, is a pervasive issue across all research fields, significantly impacting the accuracy of results and the reliability of conclusions (Srijan & Rajagopalan, 2024). This broad definition encompasses all situations where one or more cells belonging to one or more observations are void, regardless of their causes, which can be attributed to study participants, study design, or the interaction between participants and the study design (McKnight et al., 2007). Such omissions can lead to significant problems, including biased parameter estimates and issues in hypothesis testing, such as inaccurate standard errors and reduced statistical power (D. A. Newman, 2014).

The nature of missing data can be further dissected into its mechanisms, patterns, and extent:

Mechanisms of Missing Data: These describe whether the missingness is

random or systematic. There are three primary mechanisms:

Missing Completely at Random (MCAR): The probability of missing data on a variable is unrelated to any other measured or unmeasured variables.

Missing at Random (MAR): The probability of missing data on a variable is related to some of the measured variables in the study but not the variable itself.

Missing Not at Random (MNAR): The probability of missing data on a variable is related to the variable itself.

Patterns of Missing Data: These identify whether there is a systematic reason behind the missing data, such as a specific group of participants or a particular time point.

Extent of Missing Data: This refers to the amount or proportion of missing data, which is crucial for assessing its potential impact on the analysis and results.

Understanding these distinctions is crucial for determining the impact of missing data on research outcomes and for selecting appropriate methods to address this issue. Effective handling of missing data can enhance the validity of research findings, reduce bias, and increase the robustness of statistical inferences.

2.2.1 Missing Data Challenges

We begin with a brief review of the challenges associated with missing data in statistics. This review will cover the following: the characterisation and classification of missing data, traditional imputation methods such as mean imputation, regression techniques, and multiple imputation, alongside advanced machine learning-based approaches. Each method will be critically evaluated for its assumptions (e.g., Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR)) and suitability in handling the unique characteristics of periodontal data.

The most relevant issues arising from missing data are bias, reduced precision, and increased likelihood of Type II errors (Agiwal & Chaudhuri, 2024).

Bias due to missing data occurs when the sample containing only the observed data is not representative of the original population. This bias can be quantified under

different missing data mechanisms. Mathematically, if θ is the parameter of interest estimated by $\hat{\theta}$, the bias introduced by missing data can be represented as:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}|\text{Observed Data}) - \theta \tag{2.1}$$

Under MCAR, the missingness is independent of both observed and unobserved data, theoretically allowing $E(\hat{\theta}|\text{Observed Data})$ to approach θ if the model is correctly specified. However, under MAR or MNAR, where missingness depends on the observed data or the missing data itself, $\hat{\theta}$ may not converge to θ without adjustments for the missing data.

Reduced precision due to missing data is typically reflected in increased variances and wider confidence intervals. Suppose $\hat{\theta}$ is an estimator of θ, and the variance of $\hat{\theta}$ increases as data points are missing. The variance of $\hat{\theta}$ considering the missing data can be approximated as:

$$\text{Var}(\hat{\theta}) = \frac{\sigma^2}{n - m} \tag{2.2}$$

where:

σ^2 is the variance of the estimator if no data were missing,

n is the total intended sample size,

m is the number of missing data points.

This increase in variance results in less precise estimates and consequently less confidence in the statistical inferences made.

Increased Likelihood of Type II Errors with missing data, the effective sample size decreases, which affects the study's power, defined as $1 - \beta$, where β is the probability of a Type II error. The relationship between power, sample size, effect size (δ), and variance (σ^2) can be illustrated as:

$$\text{Power} = 1 - \beta = \Phi\left(\frac{n^{1/2}\delta}{\sigma} - z_{\alpha/2}\right) \tag{2.3}$$

where:

Φ is the cumulative distribution function of the standard normal distribution,

$z_{\alpha/2}$ is the critical value for the significance level α.

As n decreases due to missing data, the power decreases, making it more difficult to detect a true effect, thus increasing the likelihood of committing a Type II error.

2.2.2 Traditional Imputation Methods

Traditional imputation methods include mean imputation, regression techniques, and multiple imputation. Each of these methods comes with its own set of assumptions and limitations:

Mean Imputation: Assumes the missing data is MCAR and imputes missing values with the mean of the observed data, potentially reducing variability.

Regression Imputation: Uses observed data to predict missing values based on a regression model, assuming the relationship between variables remains consistent.

Multiple Imputation: Generates several possible imputed datasets, analyzes each one, and combines the results to account for the uncertainty associated with missing data.

2.2.3 Advanced Machine Learning-Based

Advanced machine learning-based approaches, such as neural networks and ensemble methods, offer more sophisticated mechanisms for handling missing data. These methods can capture complex relationships and patterns within the data, providing more accurate and reliable imputations:

Neural Networks: Capable of learning non-linear relationships and handling large amounts of data, making them suitable for complex imputation tasks.

Ensemble Methods: Combine multiple models to improve predictive performance and robustness in imputation, such as using random forests or gradient boosting.

In summary, addressing the challenges of missing data is critical for ensuring the validity and reliability of statistical analyses in periodontal research. By employing advanced imputation methods and critically evaluating their assumptions and performance, we can mitigate the impacts of missing data and improve the accuracy of our findings.

2.2.4 Characterization of Missing Data

The characterization of missing data involves understanding its extent, the underlying mechanisms, and the patterns that can be represented by a set of properties and variables. This understanding is crucial for effective handling and imputation of missing values.

Extension

The extension of missing data is defined as the percentage of cases with missing values. According to different authors, the classification varies: it is considered not extensive when up to 10% (Cohen, 1983), 15% (Hertel, 1976), or 40% (Raymond & Roberts, 1987) of cases have missing values. Beyond these thresholds, the variable may be eliminated, while below these values, the missing data can be managed effectively.

Missing Data Mechanisms

Understanding the mechanism behind data missingness is critical as it significantly affects the generalizability of the results. According to Donald B. Rubin's 1976 classification (Rubin, 1976), there are three categories of missing data mechanisms: Missing at Random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (MNAR).

MAR occurs when the probability of a missing value is explained entirely by the observed values of other variables and is independent of the unobserved missing value. This systematic missingness can be modeled as:

$$P(\text{Missing}|Y_{\text{obs}}, Y_{\text{mis}}) = P(\text{Missing}|Y_{\text{obs}}) \tag{2.4}$$

where Y_{obs} are the observed values, and Y_{mis} are the missing values.

MCAR describes a situation where the missingness of data is entirely random and does not depend on any observed or unobserved data. This is represented as:

$$P(\text{Missing}|Y_{\text{obs}}, Y_{\text{mis}}) = P(\text{Missing}) \tag{2.5}$$

A dataset with MCAR can be considered a random sample of the Full Dataset (FD).

MNAR suggests that missingness depends directly on the unobserved data itself, which can be modeled as:

$$P(\text{Missing}|Y_{\text{obs}}, Y_{\text{mis}}) \neq P(\text{Missing}|Y_{\text{obs}}) \tag{2.6}$$

In this scenario, the probability of a data value being missing is related to the unobserved values, making it challenging to use existing data to approximate the missing values.

Testing for MCAR

Using a missing data dummy code (e.g., $y_{\text{ob}} = 1$ for observed, $y_{\text{mi}} = 0$ for missing) can help assess the MCAR assumption. Correlations with other variables $(z_1, z_2, \dots)$ in the dataset (Acock, 1997; Cohen & Cohen, 2003), and comparisons of means and frequencies (using t-tests or chi-square tests) can identify associations (Acock, 1997; Huisman, 1999).

If significant associations are found, the MCAR assumption is ruled out. Logistic regression models can further test if other variables predict missingness (Choi et al., 2019; Hair et al., 2019; Little & Rubin, 2020; Orme & Reis, 1991). If no significant predictors are found, it supports the MCAR assumption.

Visualizing Missing Data

Visual tools complement statistical tests and intuitively expose data structures. Graphical checks for statistical test requirements, such as outliers or skewed distributions, are useful for understanding data characteristics, as noted by Templ and Filzmoser (2008).

In practical scenarios, MCAR is rare, and MAR is a safer assumption. Some MNAR cases can be transformed into MAR, making methods effective under MAR particularly useful.

Assessing Missing Data Mechanisms

The MCAR assumption can be verified through hypothesis tests for association (correlations and comparisons) and regression methods (Little, 1988). χ^2 test is used for multivariate, partially observed quantitative data. Rejection of the null hypothesis in t-tests indicates data are not MCAR.

Transforming MNAR to MAR

Transforming MNAR to MAR can make imputations more reliable. Common approaches include the inclusion of auxiliary variables, modeling the missing data process, and using instrumental variables.

Inclusion of Auxiliary Variables Assume a variable Y with missing values, where the missingness depends on Y itself (MNAR). If an auxiliary variable X is correlated with both Y and its missingness:

$$P(R = 1|Y, X) = f(Y) \tag{2.7}$$

If X captures the influence of Y on missingness:

$$P(R = 1|Y, X) \approx P(R = 1|X) \tag{2.8}$$

This approximates a MAR scenario (Hardt et al., 2012).

Modeling the Missing Data Process Generative models can address MNAR by modeling latent variables influencing missingness. For instance, if Y depends on an unobserved variable Z, and Z is modeled from observed data X:

$$Z = g(X) + \epsilon \tag{2.9}$$

where g is the functional relationship, and ϵ is the error term. The probability of observing Y given Z:

$$P(R = 1|Y, Z) = h(Z) \tag{2.10}$$

Substituting Z with $g(X) + \epsilon$:

$$P(R = 1|Y, X) \approx P(R = 1|g(X) + \epsilon) \tag{2.11}$$

This approximates MAR conditions (Ma & Zhang, 2021).

Using Instrumental Variables In cases where missingness depends on an unobserved variable, instrumental variables (Z) help isolate the mechanism. If Z affects

missingness but is independent of Y:

$$P(R = 1|Y, Z) = k(Z) \tag{2.12}$$

A valid Z allows analysis that approximates MAR, improving statistical conclusions (Sun et al., 2018).

Patterns

Little and Rubin (2019) considered two main missing data patterns: arbitrary and monotone. The arbitrary missing data pattern is characterized by the occurrence of missing observations at random locations throughout the dataset, where the order of columns is not relevant. In contrast, monotone missing patterns are defined by the sequential nature of missing data, where the order of columns is crucial and the occurrence of missing data is not random.

2.3 Dealing with Missing Data

2.3.1 Missing Data Handling Methods

The problem of missing data can be addressed using a variety of techniques described by Vahdati et al. (2024). These include:

1. **Deletion Techniques:**

Complete Case Analysis: Involves analyzing only the cases with no missing values.

Available Case Analysis: Utilizes all available data without imputing missing values.

2. **Weighting Techniques**: These techniques adjust the weights of observed data to account for missingness.

3. **Imputation Methods:**

Single Imputation: Replaces missing values with a single value, such as the mean, median, or mode.

Multiple Imputation: Generates multiple datasets by replacing missing values with multiple sets of plausible values and then combining the results.

4. **Model-Based Methods:** These involve using statistical models to estimate the missing data, incorporating the uncertainty around the missing values.

5. **Data Augmentation (DA) Techniques:** These involve augmenting the dataset with additional information to better estimate the missing values.

6. **Artificial Intelligence (AI) and Deep Learning (D-L) Methods:** These advanced techniques utilize AI and deep learning algorithms to predict missing values based on the patterns in the data (Liu et al., 2023; Myrtveit et al., 2001).

These techniques vary in their impact on the amount of dispersion around true scores, the degree of bias in the final results (Roth & Switzer III, 1995), the availability of software packages, and the computational expense. Therefore, the choice of a method must be carefully considered. Since the MAR mechanism is more realistic than the MCAR mechanism, handling methods that only perform well under the MCAR assumption are of limited use.

Complete Case Analysis (CCA)

Techniques that ignore incomplete observations involve omitting cases with missing data from statistical calculations. These methods are only applicable to MCAR type data sets since they can be considered a random sample of a complete data set. Deletion results only in the reduction of sample size and consequent loss of statistical power.

Complete case analysis, also known as "list-wise deletion," involves deleting all observations that have one or more missing values. An alternative method is "specific deletion," which deletes only those observations that have more than a given percentage of missing values (Jönsson & Wohlin, 2004; Kumutha & Palaniammal, 2013; Scheffer, 2002). Another approach, partial deletion—also called "variable deletion" or "pair-wise deletion" omits only the observations with missing values in the variables of interest.

However, when a substantial amount of data is missing, these methods result in severely biased parameter estimates and standard errors (Graham, 2009). In extreme cases, this can lead to a waste of the data set (Royston, 2004). As stated by Wilkinson, 1999, deletion methods "are among the worst methods" for handling missing data. These methods do not make full use of available information and can significantly compromise the validity and reliability of statistical analyses.

To mitigate these issues, researchers often prefer more sophisticated techniques such as multiple imputation or model-based methods that better handle missing data while preserving the integrity of the data set and improving the robustness of the analysis.

Adjusting for Biases in Complete Case Analysis Using Weighting Techniques

Little and Rubin (2020) propose the use of weighting techniques to adjust for biases inherent in CCA, particularly within the framework of randomization inference for fi-

nite population surveys. They emphasize that utilizing only complete cases can lead to biased results if the missing data are not MCAR. In such scenarios, the remaining data may not be representative of the entire dataset. To mitigate this bias, weights are assigned to the observed cases based on the inverse probability of being observed, aiming to make the analyzed sample more representative of the original intended sample.

The Inverse Probability Weighting (IPW) method is a foundational technique for this adjustment. Consider a dataset with n units, where each unit i has a probability π_i of being included in the sample. The weight for each unit is defined as:

$$w_i = \frac{1}{\pi_i} \tag{2.13}$$

This weighting compensates for non-random missingness or selection by up-weighting the influence of units that were less likely to be included. The weighted estimator for the mean of a variable Y is then expressed as:

$$\hat{\mu}_w = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} \frac{y_i}{\pi_i}}{\sum_{i=1}^{n} \frac{1}{\pi_i}} \tag{2.14}$$

where y_i are the observed values of Y.

This method ensures that the resulting estimates are less biased by giving more weight to the observations that are underrepresented due to the missing data mechanism. The approach assumes that the probability of being observed π_i can be estimated accurately. These weights can be derived from logistic regression models or other probability models that predict the likelihood of an observation being complete based on available data.

In practice, IPW involves several steps:

1. Modeling the Missingness Mechanism: Estimate the probability π_i for each unit using logistic regression or other appropriate models.

2. Calculating Weights: Compute the weights $w_i = \frac{1}{\pi_i}$ for each observed unit.

3. Applying Weights: Use these weights in subsequent analyses to adjust for the missing data, ensuring that the weighted sample better represents the original population.

By employing IPW, researchers can mitigate the biases associated with CCA, leading to more accurate and reliable estimates. This technique is particularly valuable in survey research and other fields where missing data is a common issue.

Randomization Inference in Finite Population Surveys

In finite population surveys, where units are randomly sampled, the concept of randomization inference plays a crucial role. It involves calculating the probability of observing the collected data under various random assignments of units to treatment or control groups. This method ensures that the inference accurately reflects the random nature of both sample selection and treatment assignment (Hansen et al., 1953; Rao & Fuller, 2017; Ritzwoller et al., 2024).

Randomization inference calculates the probabilities associated with different possible outcomes of the study based on various configurations of assigning units to either the treatment or control group. This approach leverages the randomization distribution, which represents the distribution of potential outcomes over all possible random assignments. This ensures that the conclusions drawn from the analysis are valid under the randomization model, providing an unbiased estimation of the treatment effect (Ritzwoller et al., 2024).

The formula to represent the probability of observing a specific sample under random assignment is given by:

$$P(S = s) = \prod_{i \in s} \pi_i \prod_{i \notin s} (1 - \pi_i)$$

where S is the set of sampled units, s represents a particular sample, and π_i is the probability of selecting unit i. This formula highlights how the likelihood of any given sample configuration can be decomposed into the product of inclusion probabilities for units that are selected and exclusion probabilities for units that are not selected (Ritzwoller et al., 2024).

Randomization inference provides several advantages in the context of finite population surveys:

Unbiased Estimation: By leveraging the randomization distribution, estimators derived under this framework are unbiased, ensuring that the expected value of the estimator equals the true parameter value (Hansen et al., 1953).

Validity Under Randomization: The method remains valid as long as the randomization process is properly conducted, making it robust to various assumptions that are often required in model-based approaches (Hansen et al., 1953; Ritzwoller et al., 2024).

Flexibility: It can be applied to various types of experimental designs, including completely randomized designs, stratified designs, and more complex survey designs

(Rao & Fuller, 2017; Ritzwoller et al., 2024).

In summary, randomization inference is a powerful tool in the analysis of finite population surveys, providing a rigorous framework for making valid inferences about treatment effects while accounting for the inherent randomness in the sampling and assignment processes.

Weighting in Randomization Inference

The weighting methods used in randomization inference mirror those applied in broader survey analysis contexts. They ensure that the presence of each unit in the population is appropriately represented in the sample. This is especially important in finite population surveys to adjust the influence of each unit based on its probability of selection.

Weights are typically assigned as the inverse of the probability of selection, which helps in adjusting the analysis to reflect the entire population structure more accurately:

$$w_i = \frac{1}{\pi_i}$$

where w_i is the weight assigned to unit i, and π_i is its probability of selection (Lohr, 2019). These weights help ensure that the inference reflects the structure of the entire population, compensating for over- or under-representation of any particular unit in the sample.

By applying these weights, the analysis can better account for how data were collected or the patterns of missing data, making the results from complete case analysis more robust and representative of the entire population or the intended study sample. This approach is essential in survey methodology and causal inference, underscoring the importance of considering the collection and relevance of each data point within the larger population or dataset (Imbens & Rubin, 2015).

2.4 Imputation Techniques

Imputation techniques involve calculating an estimate of each missing value and replacing, or imputing, each value with its respective estimate, considering any other known information for that observation (Kalton, 1986). This approach aims to achieve a complete data set, allowing for standard statistical analysis.

These methods are particularly useful for handling missing values of the MAR or MCAR types, especially when each record or variable in the data set is relevant, and

no single record has missing values across many variables (Royston, 2004). Imputation can be performed using single or multiple imputation techniques.

Single Imputation

Single imputation aims to obtain one complete data set without accounting for the variability of the imputed values. This approach uses a set of external covariates to generate a range of plausible values for each missing value, based on correlations between the covariates and the item to be imputed.

The missing values can be replaced by group means, medians, or modes (constant replacement methods) or any other predefined value, which is a common approach in single imputation. Additionally, the values that "approximate" the missing response can be obtained through methods such as regression, stochastic regression, the Expectation-Maximisation algorithm (EM) algorithm, similar response patterns, hot deck techniques, last value carried forward for longitudinal data, and a variety of other methods (Scheffer, 2002).

- **Mean/Median/Mode Imputation:** Replaces missing values with the mean, median, or mode of the observed data.
- **Regression Imputation:** Uses regression models to predict the missing values based on other available information.
- **Stochastic Regression:** Adds a random error term to the predicted values from regression imputation to reflect the uncertainty of the prediction.
- **Expectation-Maximization (EM):** Iteratively estimates the missing values by maximizing the likelihood function.
- **Hot Deck Imputation:** Replaces missing values with observed values from similar records.
- **Last Observation Carried Forward (LOCF):** For longitudinal data, replaces missing values with the last observed value.

Multiple Imputation

The complexity introduced by missing data in statistical analyses necessitates sophisticated estimation techniques such as Multiple Imputation (MI). MI involves creating

several complete datasets by imputing the missing values based on distributions estimated from observed data. The variability between these imputations reflects the uncertainty due to missing data and is combined to give:

$$\hat{\theta}_{MI} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{2.15}$$

$$\text{Var}(\hat{\theta}_{MI}) = \bar{V} + \left(1 + \frac{1}{M}\right) B \tag{2.16}$$

where:

- M is the number of multiple imputations,
- $\hat{\theta}_m$ is the estimate from the m-th imputed dataset,
- $\bar{V}$ is the average within-imputation variance,
- B is the between-imputation variance.

MI is a sophisticated statistical technique used to address the issue of missing data in datasets, particularly where the missing data mechanism is understood to be MAR or possibly MNAR. The fundamental principle behind MI is to create several plausible imputations for missing values, rather than filling these gaps with a single estimate. This technique reflects the uncertainty inherent in determining what value to impute (Kleinke et al., 2020). The mathematical representation of this process is:

$$Y^{(m)} = \text{impute}(Y^{(obs)}, \theta^{(m)}), \quad m = 1, 2, \ldots, M$$

where $Y^{(obs)}$ represents the observed parts of the dataset, $Y^{(m)}$ denotes the m-th imputed dataset, $\theta^{(m)}$ are the parameters estimated from the observed data, and M is the total number of imputed datasets.

The implementation of multiple imputation typically involves three key steps: imputation, analysis, and pooling.

The Imputation Process Initially, the imputation step involves replacing missing data multiple times to generate several complete datasets. This is often done using regression models conditioned on observed data and incorporating random variation:

$$X_j = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon$$

where ϵ represents the normally distributed error term, introducing variability across the different imputations. This variability is crucial as it captures the uncertainty around each imputed value.

Analysis and Pooling of Results Following the generation of multiple complete datasets, each dataset is analyzed using standard statistical methods as though it were a fully observed dataset. The results from these individual analyses are then combined, or pooled, to produce a single estimation result. The pooling process generally employs Rubin's rules, which calculate the overall estimates and their variances by considering both the variability within each imputation and the differences across the imputations:

$$\hat{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}^{(m)}$$

$$\text{Var}(\hat{\theta}) = \overline{V} + \left(1 + \frac{1}{M}\right) B$$

Here, $\overline{V}$ is the average of the within-imputation variances, B is the between-imputation variance, and M is the number of imputations. This method ensures that the final statistical inferences adequately reflect the uncertainty due to the missing data, providing a more robust and reliable foundation for decision-making based on the analyzed data.

Fully Conditional Specification (FCS)

FCS is an approach for defining multivariate models using conditional distributions. A well-known application of this approach, where each variable is imputed based on all other variables, is referred to as Multivariate Imputation by Multivariate Imputation by Chained Equations (MICE).

The MI process, as described by Rubin (1987a), involves imputing missing values using an appropriate model that incorporates random variation. This process is repeated M times (usually three to five times), producing M "complete" data sets. The desired statistical analysis is performed on each dataset using complete-data methods, followed by averaging the parameter estimates across the M samples to produce a single-point estimate. The standard errors are calculated by averaging the squared standard errors of the M estimates, and the variance of the M parameter estimates across samples is combined using Rubin's formula.

Regression Imputation

Regression imputation, or conditional mean imputation, replaces the missing values with predicted values from a regression of the missing value on other observed variables, exploiting the correlation among them (Buck, 1960; Myrtveit et al., 2001).

Stochastic Regression

Stochastic regression is a refinement of regression imputation that addresses correlation bias by adding noise to the predictions.

Expectation-Maximization (EM)

The EM algorithm can be used in missing data settings to compute maximum likelihood estimates from an incomplete dataset. This algorithm starts from an original idea by Fisher and McKendrick Fisher (1925) and M'Kendrick (1925) of estimating the missing value and iteratively re-estimating the parameters using the estimated missing values. It was further developed by Dempster et al. (1977a), Rubin (1987a), and Schafer (1997b) and popularized by Van Dyk and Meng (2001). The EM algorithm explores the interdependence between missing data Y_{mis} and parameters θ. The fact that Y_{mis} contains information relevant to estimating θ and θ in turn helps us find likely values of Y_{mis} suggests the following scheme for estimating θ in the presence of Y_{obs}. First, fill in the missing data Y_{mis} based on an initial estimate of θ, re-estimate θ based on Y_{obs}, and then fill in Y_{mis} and iterate until the estimates converge (Schafer, 1997b).

Similar Response Pattern Imputation (SRPI)

In SRPI, the missing values are replaced by observed values from a case that scored similarly, where the similarity is determined by a set of user-specified matching variables (Joreskog & Sorbom, 1993; Kingman & Albandar, 2002).

Hot Deck (H-D) Technique Hot deck imputation is a method for handling missing data in which each missing value is replaced with an observed value from a "similar" unit. The hot deck technique uses a completely observed "donor case" for the imputation of an incomplete case. The missing value is replaced by the corresponding value of the best donor case, which is found by minimizing the "distance" between the donee and all potential donor cases (typically, Euclidean distance computed in the "space" of covariates) (Myers, 2011).

Last Observation Carried Forward (LOCF) imputes the previous observed value as a replacement for the missing data. When multiple values are missing in succession, the method searches for the last observed value. These imputation methods are *ad hoc* for longitudinal data.

Behind the model-based methods already addressed in this chapter, the Full Information Maximum Likelihood (FIML) method also requires the researcher to make assumptions about the joint distribution of all variables.

FIML (Enders, 2001; Little & Rubin, 2019; Schafer, 1997a) does not involve the imputation of missing values. Instead, it directly estimates the parameters from available items using a maximum likelihood algorithm to maximize the likelihood function (Arbuckle, 1996). Along with the multiple-group approach and the EM algorithm, FIML assumes that the covariation among variables can be used to infer or estimate probable values for the missing data, though these methods differ in their mathematical approaches (Baraldi & Enders, 2010).

Data Augmentation (DA)

The methods referred as DA are based on Bayes' Rule for estimating the joint density refers to methods for iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables. If the data is MCAR or MAR then all the relevant information about the parameters are contained in the observed data likelihood $L(\theta \mid Y_{obs})$ or observed-data posterior $p(\theta \mid Y_{obs})$. The main idea of EM and DA algorithms is to impute the values in a way so that there is no change in the statistical properties of the data set (Imtiaz & Shah, 2008).

The methods referred to as DA are based on Bayes' Rule for estimating the joint density and involve methods for iterative optimization or sampling algorithms through the introduction of unobserved data or latent variables. When the data is MCAR or MAR, all the relevant information about the parameters is contained in the observed data likelihood $L(\theta \mid Y_{obs})$ or the observed-data posterior $p(\theta \mid Y_{obs})$. The main idea behind EM and DA algorithms is to impute the values in a way that maintains the statistical properties of the dataset (Dempster et al., 1977b; Tanner & Wong, 1987).

2.4.1 Artificial Intelligence - Machine Learning

More recently, new methods of handling missing data using neural networks to generate the missing values have been developed.

Traditional approaches in dealing with missing data, mainly model-based techniques considered the best methods, rely on normal or multivariate normal distribution assumptions.

AI can be defined as the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings (Copeland, 2020). It includes the entire universe of computing technology that exhibits anything remotely resembling human intelligence (IBM, 2020). Emerging in the 1980s as a set of rules in conjunction with conditional logic methods, AI evolved to data-related procedures. The modeling methodology paradigm underwent a deep change, extending iterative procedures beyond linear models, enabling the solution of problems where traditional models fail, especially for modeling non-linear relationships. Due to the structure of AI to mimic the nervous system, these types of models are designated as Artificial Neural Networkss (ANNs) or Neural Networks (NNs).

The building blocks of NNs is the neuron/perceptron, defined by Rosenblatt (1958) as a probabilistic model for information storage and organization in the brain (Rosenblatt, 1958). It consists of four main parts: the input values, weights and bias, net sum, and an activation function, behaving as a simple logic gate with binary outputs and considered the earliest supervised training algorithms capable of binary classification. The multilayer association of perceptrons (Multilayer Perceptron (MLP)) allowed for more complex systems to perform more elaborate tasks, giving birth to a new field of AI, ML, which can be defined as the study of computer algorithms that improve automatically through experience (Mitchell, 1997), relying primarily on data to optimize and "learn" how to perform tasks (Campesato, 2020).

MLPs are feed-forward neural networks with an architecture comprising four layers: the input, the hidden, and the output, each formed by neurons. The input signals are distributed forward from the input layer to the next layers, where each neuron receives a signal, which is a weighted sum of the outputs of the nodes in the previous layer. The activation function inside each neuron controls the input. This architecture determines a non-linear mapping from an input vector to the output vector, parameterized by a vector of weights.

With the development of the Back Propagation (BP) learning algorithm (Rumel-

hart, Hinton, & Williams, 1986), ANN became able to learn from data, allowing these models to adapt and establish almost any relationship between data without being explicitly programmed.

The BP algorithm, or backward error propagation, aims to train a neural network through the chain rule method. After each forward pass of the input through a network, backpropagation performs a backward pass while adjusting the model's parameters (weights and biases). This process is iterated to minimize a measure of the difference (loss function) between the actual output vector of the net and the desired output vector. The BP algorithm with the Levenberg-Marquardt for nonlinear least squares training function is probably the best-known learning algorithm to train supervised learning feedforward neural networks.

Stacking simpler networks (one hidden layer) Autoencoder (AE) creates deep networks, and Deep Belief Networks (DBN) arise from connecting Restricted Boltzmann Machines (RBM). These higher levels of AI are designated as Deep Learning (DeLe) or Very Deep Learning (VDeLe) when more than ten layers are involved. DeLe focuses NN on algorithms for training deep architectures composed of multiple levels of non-linear operations, i.e., neural nets with many hidden layers or in complicated propositional formulae re-using many sub-formulae (Bengio, 2009).

AI, which evolved in different branches to solve specific tasks, together with the complexity of its taxonomy, became more complex to encompass the new features necessary to accomplish each task. Thus, NN classification is related to the type of task to perform, NN architecture, learning procedures, learning rules applied, categorization of data, etc. The most important branches of AI include: Machine Learning (MaLe), DeLe, Natural Language Processing (NLP), Reinforcement Learning (RL), and Deep Reinforcement Learning (DRL).

The learning procedures, distinctive features of ML including DL and VDeLe, are classified according to the need for labeled data during the training phase as Supervised Learning (SL), Semi-Supervised Learning (SSL), Unsupervised Learning (UL), and RL. SL involves classifying the data point by labeling each one. Dealing with unlabeled data, UL methods involve clustering procedures (such as k-Means, hierarchical cluster analysis, and EM) and, in high dimension data sets, dimensionality reduction procedures (Principal Component Analysis (PCA), kernel PCA, locally linear embedding, and t-distributed stochastic neighbor embedding).

DL comprises different NN architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short Term Memorys (LSTMs),

a family of NN that have multiple variations such as LSTM, Gated Recurrent Unit (GRU), Variational Autoencoders (VAE), and Generative Adversarial Network (GAN). Since the beginning, ML has made use of several classifiers, regression, and clustering algorithms. Some of the most used are classifier algorithms, such as k-Nearest Neighbors (k-NN), Decision Trees (DT), Random Forest (RF), and Support Vector Machines (SVM) (Vapnik, 1995), and regression algorithms such as linear regression.

The performance of ANN models can be evaluated by different and NN-non-specific metrics such as classification accuracy, logarithmic loss, confusion matrix, area under curve, F1 score, mean absolute error, mean squared error, Akaike information criteria, and Bayesian information criteria. Some more NN-specific metrics have been proposed, like the Robustness Interpretability Completeness Correctness (RICC) dependability attributes of NN (Cheng et al., 2018). The option depends on the type of model and the aspect of the performance.

The versatility, variety, and suitable metrics have made NN an increasingly popular method for working with any kind of data, in any scientific area, and performing a wide variety of tasks, including dealing with missing data.

Missing Data Handling with Artificial Intelligence

Considering the concept of "learn from data," the quality of the data is more important than the MaLe algorithms themselves. The quality of data, particularly the handling of missing data, is of paramount importance because it can significantly affect the results, regardless of the algorithm's quality or the NN architecture. Therefore, considerable effort has been made to develop efficient methods to address the issue of missing data. In the following paragraphs, we will explore some of these methods as potential examples to follow and develop.

MaLe methods address missing data in two main ways: by imputing the missing values or by learning directly from the data without imputation through robust algorithms.

No Imputation Approach: No imputation methods are more limited in their enforcement and are much more suitable for classification and prediction tasks.

The most popular NN methods for dealing with missing data are classifier-type algorithms such as k-NN, DT, RF, and SVM (Garcia et al., 2010). The goal of these methods is to yield the best possible predictions on test data with missing values without completing the dataset. The advantage of DT or RF for classification tasks is that these models do not require imputing missing data nor encoding categorical

variables, unlike ANNs or other classifiers (Poulos J., 2018; Troyanskaya et al., 2001).

Imputation Approach: Imputation methods fill in the missing values before applying machine learning algorithms. Various techniques can be used, ranging from simple mean imputation to more complex methods like multiple imputation or model-based approaches.

One common method is the k-NN imputation, where the missing value is estimated based on the values of the nearest neighbors (Tutz, 2010). Another advanced method is the use of autoencoders in deep learning, which can learn a compressed representation of the data and reconstruct it, filling in the missing values in the process (Vincent et al., 2008).

Moreover, sophisticated algorithms like RF can also be adapted for imputation purposes. In this method, an ensemble of decision trees is used to estimate the missing values by averaging the predictions from multiple trees (Stekhoven & Bühlmann, 2012).

In conclusion, handling missing data effectively is crucial for the performance of MaLe algorithms. Whether through imputation or robust algorithms that can handle incomplete data, the choice of method depends on the specific context and the nature of the data.

The tree-based models can incorporate missing values as an attribute Missing Incorporated in Attribute (MIA) for handling incomplete input data without a previous estimation of missing values, especially when binary classification is the goal (Josse et al., 2019; Twala et al., 2008).

García-Laencina et al. (2010) and Ghorbani and Zou (2018) proposed a general embedding approach to learn representations for missingness, where the embedding acts as a modular layer in any neural network (NN) architecture. This embedding is learned concurrently with the network's predictions (Ghorbani & Zou, 2018) . The authors also introduced the Best Estimation for Sequential Trees (Branch-Exclusive Splits Trees (BEST)) algorithm, which shares several features with MIA and performs well under MCAR, (MAR, and MNAR conditions. Moreover, the BEST algorithm produces interpretable trees and achieves accuracy comparable to most missing value handling techniques.

Imputation Approach: Imputation methods have the advantage of filling in all the data values, enabling any type of statistical analysis.

However, the disadvantage is that the vast majority of imputation methods require unrealistic assumptions about the missing mechanism (García-Laencina et al., 2010; Ghorbani & Zou, 2018). The BP algorithm overcomes this limitation (Rumelhart, Hin-

ton, & Williams, 1986; Rumelhart, McClelland, & the PDP Research Group, 1986) by repeatedly adjusting the weights of the connections in the network to minimize the difference between the actual output vector and the desired output vector. This method has proved advantageous over traditional missing data reconstruction techniques due to its robustness to invalid distributional assumptions and lower computational complexity (Gupta & Lam, 1996; Sarle, 1994).

Some algorithms commonly used by MaLe for missing values imputation include AE, MLP, Self-Organized Map (SOM), fuzzy-Neural Network (f-NN), Robust Association Rules (RAR), and RNN.

The evaluation of imputation approaches is based on the performance of classification or regression tasks, rather than on how close the imputed values are to the real ones. The following examples illustrate the variety of methods and algorithms that can be applied to different problems across numerous research fields, demonstrating the versatility of NN models.

Bengio and Gingras (1996) applied a RNN with feedback into the input units for handling static and sequential missing data analysis problems. The temporal dimension of RNNs gives them an advantage over other NN models in dealing with the temporal and spatial dimensions of data (Zhu et al., 2019).

Lakshminarayan et al. (1996) used the autoclass (Cheeseman et al., 1988) unsupervised clustering strategy, employing a Bayesian approach to cluster the data into classes, which were then used to predict multiple options for the missing values. The same authors also applied the C4.5 (Quinlan, 1993) group of programs to model the missing variables by supervised induction of a decision tree-based classifier, predicting the most likely value for the missing data point.

Abdella and Marwala (2005) proposed an AE (or auto-associative Neural Networks) constructed using MLP networks and trained using BP in combination with Genetic Algorithm (GA), as well as a Radial Basis Function (RBF) (Haykin, 1999) network also combined with GA. These feedforward networks are typically configured with a single hidden layer of units activated by one basis function. The objective of the GA is to minimize an error function derived from an auto-associative neural network (autoencoder, implemented as a MLP) and a Radial Basis Function (RBF) network.

Silva-Ramírez et al. (2011) used a three-layered MLP with the hyperbolic tangent activation function in the hidden layer and the identity function as the activation function for the output layer.

Gad et al. (2020) evaluated the ability of CNN and dense layers for the prediction

and filling of missing data associated with k-NN.

Jerez et al. (2010), using a data set from the Spanish Breast Cancer Research Group, compared imputation methods based on statistical techniques (mean, hot-deck, and multiple imputation) with MaLe techniques (MLP, SOM, and k-NN) and concluded that the prognosis accuracy was significantly better with MaLe.

Choudhury and Pal (2019) tested an AE with a sigmoidal-type activation function with an innovative two-stage training scheme. This approach relies on the capability of NN type AE to reconstruct their input. When there are missing values in the data, which can be considered as noise, the AE can learn the most salient features of the data and then reconstruct the input data along with the missing values (Platias & Petasis, 2020).

2.5 Periodontal Examination Protocols: Full-Mouth and Partial-Mouth Approaches

The FMPE, the "gold standard" periodontal recording protocol, is widely used in clinical periodontics for accurate clinical diagnosis and unbiased estimators of periodontal disease prevalence under any classification framework. This protocol involves the examination of six sites (MV, V, DV, ML, L, and DL) on all teeth, excluding third molars, resulting in a maximum of 168 sites per mouth. This comprehensive method ensures that the true prevalence only depends on the case definition, providing full-mouth data (FD). According to Owens and Palmer (1999), conducting individual FMPE for measuring PPD and gingival recession takes an average of 28.8 minutes, and 40 minutes to measure PPD, bleeding on probing, calculus, and CAL (Benigeri et al., 2000).

However, due to the labor, time, resource, and patient burden, the FMPE method is less frequently used in large surveys (P. I. Eke, Dye, et al., 2012; Owens et al., 2003). To address these challenges, full-mouth restricted protocols using only selected few sites per tooth have been developed. The National Institute of Dental and Craniofacial Research The National Institute of Dental and Craniofacial Research (NIDCR) fully developed full-mouth protocols consisting of all MV, V, or all MV, V, and DV sites. These protocols aim to minimize the disadvantages of FMPE while maintaining accuracy.

To obtain acceptable estimates of prevalence and severity with minimal bias, while reducing costs, PMPE design protocols have been used. These protocols reduce the number of variables to be collected and consist of examining a selected number of teeth

(index teeth) or sites per tooth or some sites in selected teeth adjusted to the case definition of interest. The expectation is that the selected subset of data points represents the entire mouth, yielding information that can be applied universally (Kingman & Albandar, 1999; Kingman & Albandar, 2002). The specific sites within subjects that should be evaluated depend mainly on the study's objective, such as estimating prevalence or severity (Mascarenhas, Garetto, & Johnson, 2020; Mascarenhas, Okunseri, Dye, et al., 2020).

Several PMPE protocols use six sites per tooth, as described for the NHANES protocol, but on selected teeth. These include:

- **Ramfjord teeth subset** or "Ramfjord teeth" (Ramfjord, 1959), which examines the right maxillary first molar (16), left maxillary central incisor (21), left maxillary first premolar (24), left mandibular first molar (36), right mandibular central incisor (41), and right mandibular first premolar (44).

- **Community Periodontal Index of Treatment Needs (CPITN)** (Ainamo et al., 1982), which includes the right maxillary first and second molars (16, 17), right maxillary central incisor (11), left maxillary first and second molars (26, 27), left mandibular first and second molars (36, 37), left mandibular central incisor (31), and right mandibular first and second molars (46, 47).

- **Random Half-Mouth (RHMP)**, which uses homolateral quadrants, diagonal quadrants, and one maxillary and one mandibular quadrant randomly selected and examined. Some RHMP protocols use six sites per tooth in every tooth of the quadrant, while others select only a fixed number (< 6) of sites per tooth. The 1985-1986 NIDCR Adult and NHANES III used the MV and V sites. The NHANES IV (P. Eke et al., 2010) used the MV, V, and DV sites (Dye & Thornton-Evans, 2007).

Additionally, some protocols examine all teeth present in the dental arch but evaluate only selected sites such as MV–V; MV–V–DV; and ML–L–DL (for an extensive review: Tran et al., 2013). A number of studies (Borrell & Papapanou, 2012; Couto et al., 2018; P. I. Eke et al., 2015; Kingman et al., 2008; Kingman & Albandar, 2002; Kingman & Albandar, 2008; Papapanou, 2012; Tran et al., 2013), among others, have assessed the quality of these estimators. The biases associated with PMPE protocols vary with the number and sites of the protocol, disease level in the population, and periodontitis case definition.

In the context of missingness by design, data are intentionally missing (controlled) to save resources and reduce participant burden. This results in a loss of information critical for correct classification of disease status according to the case definition issued by the 2017 World Workshop on the Classification of Periodontal and Peri-Implant Diseases and Conditions (Caton et al., 2018; Tonetti et al., 2018). Consequently, methods to estimate and impute the omitted data are required.

To summarize, while the FMPE remains the gold standard for recording protocols in periodontal surveys due to its accuracy and ability to provide unbiased estimators of periodontal disease prevalence, its demanding nature limits its use in large-scale surveys. Alternative PMPE protocols have been developed to balance the need for accurate data with practical constraints, ensuring representative sampling with minimized biases.

2.6 Symmetry Assumption

"Some objects are more symmetrical than others" (Atkins, 1986).

2.6.1 Symmetry of the Mouth

Symmetry is a characteristic of animal body plans. Animals can be classified into three groups based on their body symmetry: radially symmetrical, bilaterally symmetrical, and asymmetrical. The bilateral symmetry that involves the division of the animal through a saggital plane, resulting in two mirror images, is a feature of the human body, with exceptions for some internal organs.

Symmetry plays a crucial role in providing fundamental biomechanical attributes to structures, creating a significant link between the body (including minor anatomical structures) and the physical environment (Holló, 2015). The correct formation of the body's overall structure, including the left-right symmetry, relies on axial patterning processes by which the body's basic structure is organized along its central axis during embryonic development. The molecular and genetic regulation of axial patterning is a highly complex process involving a number of clusters of homeotic genes and signaling pathways (Carroll, 1995).

Formal symmetry of the mouth has been the basis of all notations used to represent symbolically the number of teeth by type and order (dental formula).

The teeth are located within the alveolar processes of maxilla (upper dental arch or

upper jaw) and mandible (lower dental arch or lower jaw) accordingly with the formula:

Figure 2.1

$$I\frac{2}{2} + C\frac{1}{1} + P\frac{2}{2} + M\frac{3}{3}$$

Dental formula

Where I, C, P and M stands for incisors, canines, premolars and molars, respectively altogether on one side, right or left hemiarches and the number of each type of tooth is placed above the line for the maxilla (upper jaw) and below the line for the mandible (lower jaw). Two main tooth numbering systems are used the universal and the two-digit Fédération Dentaire Internationale (FDI) system.
In the universal notation system for the permanent dentition, the maxillary teeth are numbered from 1 through 16, beginning with the right third molar. Beginning with the mandibular left third molar, the teeth are numbered 17 through 32 as described in table 2.1.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17

Table 2.1: Universal dental enumerating system.

In the FDI two digit system incorporate the concept of quadrant that is used to divide the mouth into four sections from the 1 st , 2 nd , 3 rd , and 4 th quadrant defined from the upper right side of the mouth to the lower right side in a clockwise manner (Muslim et al., 2012). The first digit indicates the quadrant: 1 to 4 (the permanent dentition) and the second digit indicates the tooth within a quadrant: 1 to 8 for the permanent teeth as described in table 2.2. The representation in table dental formula and Table 2.2 teeth in the mouth by type and location suggests suggest a visible reflective symmetry of hemi-archs.

18	17	16	15	14	13	12	11	21	22	23	24	25	26	27	28
48	47	46	45	44	43	42	41	31	32	33	34	35	36	37	38

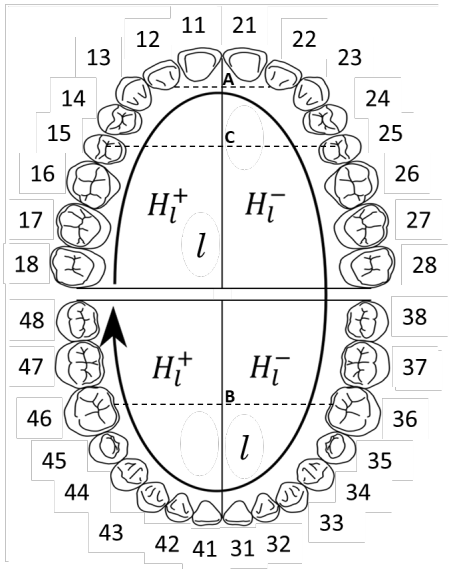
Table 2.2: FDI dental enumerating system.

Defining symmetry as "invariance of an object or process against some set of transformations called symmetry operations" (e.g., translation, rotation, reflection, inversion, etc.), which forms a mathematical group (Köhler, 1991). When considering dental arches in their perfect form, a hemiarch is the reflection of its contralateral. If we fold the representation of a dental arch along the maxillary or mandibular midlines, we would see both hemiarches match exactly. This meets the criteria for reflection according to the following definition:

Let l be any line in the plane. The reflection across axis l is a transformation that sends each point A to another point A', such that line l is the perpendicular bisector of the segment AA'. We call S_l the reflection on the line l, so $S_l(A) = A'$. If H_l^+ and H_l^- are the half-planes determined by l, then $S_l(H_l^+) = H_l^-$ and $S_l(H_l^-) = H_l^+$. All points on line l remain fixed when the reflection on line l is performed (Félix, 2015).

In mouth the symmetry axis l consist in the maxillary and mandibular midlines.The maxillary midline is obtained by drawing a line on the mid-palatal suture from the interincisor papillae (between teeth 8 and 9) to the tangential line to the distal third molar (teeth 1 and 16) surfaces, and perpendicular to l (Maurice & Kula, 1998).In figure 2.2 the mandibular midline (l) correspond to a line drawn using the mandibular references equivalent to those used in the maxilla, i.e mesially the papillae between 24 and 25 and distally the tangential line to the distal surfaces of 17 and 32 (Alavi et al., 1988). The midline (l) divides each dental arch (maxillary and mandible) in two hemiarches which are half-plane determined by l, and the midline l is a perpendicular bisector of the segment connecting the correspondent contralateral tooth (dashed lines A, B and C see Figure 2.2). Thus, each hemi-arch is reflexive or amphichiral.

Figure 2.2



Representation of dental arch symmetry with FDI enumerating system

The human mouth does not exhibit perfect bilateral symmetry according to any mathematical definition of symmetry. Perfect bilateral symmetry is uncommon due to a combination of genetic, acquired, and environmental factors. However, dental arches and other biological forms that show imperfect symmetry retain a recognizable "degree" of symmetry but do not belong to the set $\mathbf{S(G)}$ containing all objects fulfilling exactly the symmetry requirements of the symmetric group $\mathbf{G}$.

To address this difficulty in classifying objects with respect to their symmetry, Kohler and Wehling (Köhler, 1991), based on fuzzy set theory introduced by Zadeh (Zadeh, 1965), developed the concept of fuzzy symmetry described as follows:

A given form $\mathcal{F}$ belongs to the set $\mathbf{S(G)}$ containing all objects fulfilling the symmetry requirements of the symmetry group $\mathbf{G}$. To the set $\mathbf{S(G)}$ we may ascribe a membership function $\mu(\mathcal{F}, \mathbf{G})$ describing the "degree of membership" of $\mathcal{F}$ in $\mathbf{S(G)}$. For a crisp set, $\mu(\mathcal{F}, \mathbf{G})$ is a binary function. In the fuzzy symmetry concept, the set $\mathbf{S(G)}$ is replaced by a fuzzy set $\tilde{\mathbf{S}}(\mathbf{G})$ characterized by the fact that the degree of membership $\tilde{\mu}(\mathcal{F}, \mathbf{G})$ of a certain form $\mathcal{F}$ in this set (i.e., the degree of fulfillment by $\mathcal{F}$ of the symmetry requirements of $\mathbf{G}$) is no longer binary but continuous ($\tilde{\mu} \in [0, 1]$).

In the domain of physical chemistry, where the classification of symmetry plays an important role, Zabrodsky et al. (Zabrodsky et al., 1992) developed the concept of continuous symmetry to characterize shape properties of molecules in static and dy-

namic states. This characterization was achieved by quantifying the symmetry grade or the intermediate "amount of symmetry" through a continuous measure of symmetry, thereby capturing important information that would be missing if symmetry were regarded as dichotomous. The Symmetry Distance (SD) was defined as the minimum effort required to turn a given shape into a symmetric shape. This is measured by the mean of the square distances each point is moved from its location in the original shape to its location in the symmetric shape. No *a priori* symmetric reference shape is assumed.

Denote Ω the space of all shapes of a given dimension, where each shape P is represented by a sequence of n points $\{P_i\}_{i=0}^{n-1}$. A distance function between every two shapes in Ω is the metric d defined on Ω as:

$$\begin{aligned} &d : \Omega \times \Omega \longrightarrow R \\ &d\left(P, Q\right) = d\left(\left\{P_i\right\}, \left\{Q_i\right\}\right) = \frac{1}{n} \sum_{i=0}^{n-1} \left\|P_i - Q_i\right\|^2 \end{aligned} \tag{2.17}$$

The Symmetry Transformation (ST) of a shape P, as the symmetric shape closest to P in terms of the metric d. The SD of a shape P is now defined as the distance between P and it's Symmetry Transform ST

$$SD = d\left(P, ST\left(P\right)\right) \tag{2.18}$$

The SD of shape $P = \{P_i\}_{i=0}^{n-1}$ is evaluated by finding the symmetry transform $\hat{P}$ of P and computing.

$$SD = \frac{1}{n} \sum_{i=0}^{n-1} \left\|P_i - \hat{P}_i\right\|^2 \tag{2.19}$$

2.6.2 Research Question

Given the limitations of traditional methods and the potential advantages of AI methods and the availability of fuzzy-symmetry related information, this dissertation seeks to answer the following research question:

How can the application of AI-fuzzy methods enhance the accuracy and validity of imputing missing data in periodontal planned missing data surveys compared to traditional imputation methods?

Chapter 3

Material and Methods

3.1 Data

The data for this study were obtained from the NHANES 2011-2012 cycle. The NHANES program is a series of studies designed to evaluate the health and nutritional status of adults and children in the United States. This program uniquely combines structured interviews and physical examinations to provide a comprehensive dataset for a wide range of health-related research.

The study population comprises individuals aged 30 and above who participated in the NHANES 2011-2012 cycle and had complete periodontal examinations. To ensure data integrity, individuals with missing data for key periodontal variables were excluded from the analysis.

The demographic data from the NHANES 2011-2012 study was read into R using the `read.table` function. This dataset includes various demographic variables and is stored in the `NHANES_11.12_Demog` dataframe. Similarly, the periodontal data from the same study was read into R and stored in the `NHANES_11.12_Perio` dataframe.

Initially, a broad set of variables was selected based on their potential relevance to the study objectives. The collected variables included demographic information such as `RIAGENDR` (gender), `RIDAGEYR` (age), `RIDRETH1` (race/ethnicity), `DMDEDUC2` (education level), `INDFMPIR` (poverty index ratio), and `DMDHREDU` (education level of the household reference person). This dataset was compiled into a dataframe named `DEMO`.

The primary periodontal health-related variables considered were PPD, CAL, and Gingival Recession (GR). Periodontal examinations were meticulously conducted by trained dental professionals and included measurements of PPD at six specific sites per

tooth (MV, V, DV, ML, L, and DL).

Despite the initial inclusion of multiple periodontal health variables, the analysis was ultimately focused on PPD data and its derived variables from the NHANES dataset. This focus was determined to be sufficient for addressing the research objectives.

The PPD data, recorded as rounded integers but inherently continuous, will be analyzed in two distinct ways based on specific needs.

First, the PPD will be treated as a continuous variable for Generalized Additive Models for Location Scale and Shape (GAMLSS) modeling. Despite being recorded as integers, the PPD represents rounded measurements of an underlying continuous scale. Using continuous family distributions allows for accurate modeling of the data's distribution, capturing trends, and estimating parameters effectively. The rounding to integers does not significantly impact the analysis's resolution or granularity, as it still permits the observation of important trends and patterns. For continuous data, visualizations will use smooth (bandwidth of 2) KDE, providing a smooth estimate of the data's distribution and helping to understand the overall structure.

Second, the PPD will be treated as a discrete variable in certain contexts, adopting suitable statistical methods for discrete data. This approach is useful for specific aspects of the research, such as assessing symmetry between contralateral sites using the Symmetry Measure (SM) function. In this context, a KDE with a bandwidth of 1 will be used to analyze discrete variables, where the peaks for each unique PPD value are clear.

Selection of individuals with full periodontal examination

The periodontal dataset was filtered to include only participants who had completed the periodontal examination (`OHDPDSTS == 1`). This filtered dataset was stored in the `Perio.comp` dataframe.

Elimination of totally edentulous patients

Further filtering was done to identify participants who were completely edentulous (i.e., had no teeth). This was determined by checking if all tooth-related columns (e.g., `D17BoDV`, `D16BoDV`, etc.) had a value of 99, indicating missing teeth. The subset of completely edentulous participants was stored in the `QQQ` dataframe. The completely edentulous participants were then excluded from the `Perio.comp` dataframe using the `anti_join` function. This ensures that only participants with at least some teeth

were included in subsequent analyses.

Subseting the data frame `Perio.comp` by periodontal healt indicator

A new dataframe, `PPD.ds`, was created from the filtered periodontal data frame (`Perio.comp`). This dataframe includes only the `SEQN` and columns containing "Bo" in their names, which represent Pocket Probing Depth (PPD) measurements. Missing data in the `PPD.ds` dataframe were handled by replacing values of 99 with `NA` using the `mutate_at` function combined with `na_if`. This step ensures that missing values are properly coded as `NA` for further analysis and imputation.

Computing and Appending the Symmetry Measure

The next step involves computing and appending a new variable, the SM, which will be introduced later in this dissertation. This measure is calculated for each pair of contralateral dental sites. Since each tooth consists of six sites, it is necessary to create paired columns to evaluate symmetry. The computed SM is a fundamental component of the analytical method.

A comprehensive list of column pairs is defined. Each pair contains the sequence number (`SEQN`) and two columns representing measurements from contralateral sites. These pairs facilitate the comparison of corresponding dental sites.

An empty list, `df_list`, is initialized to store data frames derived from the specified column pairs. For each pair, the script extracts the relevant columns from the primary data frame and stores them in `df_list`.

Each data frame within `df_list` undergoes the following steps to calculate the SM: 1. The Spearman correlation coefficient (β) is computed between the second and third columns. 2. The value of α is fixed at 1. 3. The SM for each row is calculated using an exponential decay formula that accounts for the absolute difference between paired measurements, normalized by their average and correlation.

Post-calculation, the SM column in each data frame is renamed to include the specific source names, enhancing clarity in subsequent analyses. The column names within each data frame are then customized to remove specific substrings, streamlining the dataset for further processing.

All processed data frames are sequentially joined by the `SEQN` column using functions from the `purrr` package. This consolidation results in a comprehensive data frame, `ALL.all.Bo`.

Binary transformed matrix - New data frame

A new data frame, `ALL.all.Bo_IO`, was obtained by transforming the `ALL.all.Bo` into a new data frame by replacing `NA` values with "0" and all other values with "1", except for the `SEQN` column. This transformation is aimed at describing tooth missingness before the imputation process, within the data preparation phase.

Spliting into upper and lower arches data frame

Finally, the consolidated data frame is split into two separate data frames, `UABo` (Upper Arch) and `LABo` (Lower Arch). This separation allows for focused analysis on the upper dental arch's probing pocket depth (PPD) values.

3.2 Assessing Symmetry Through Statistical Methods

In the context of our work, assuming that the mouth is conceptually symmetric, we expect that the periodontal disease parameters, such as PPD of contralateral teeth sites, are equal or approximately equal. This approach relies on hypothesis testing, where perfect symmetry occurs when the difference between the two parameters being compared lies within a confidence interval at a given confidence level.

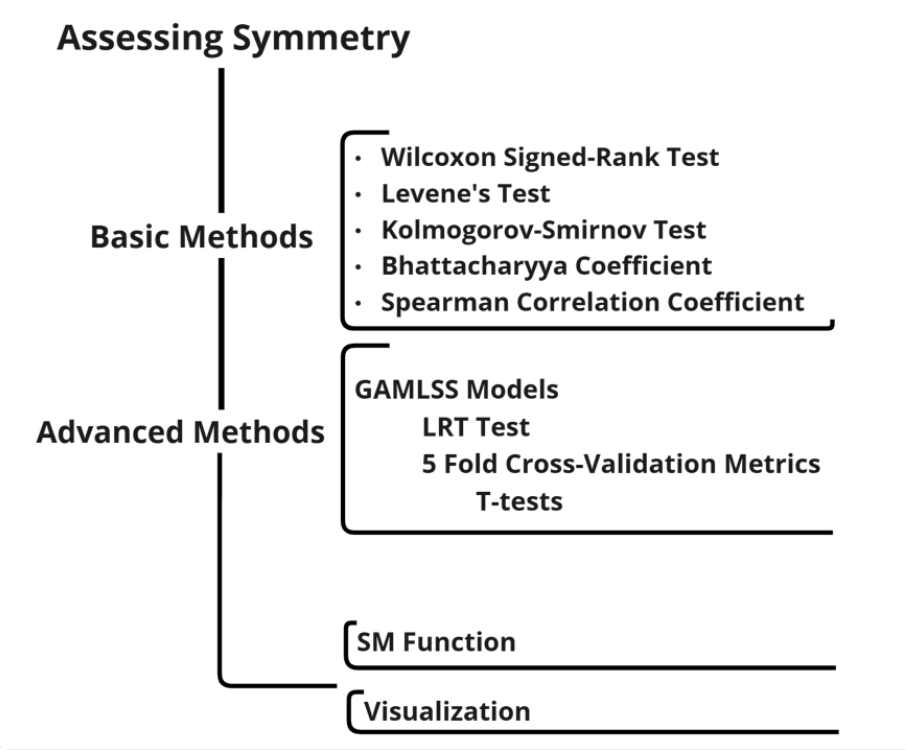
A totally different approach is the quantification of symmetry, grounded in the fuzzy symmetry framework. This method involves calculating the Symmetry Measure (SM), which provides a nuanced quantification of symmetry beyond traditional hypothesis testing.

We assess symmetry using two different approaches, designated as basic and advanced methods.

The basic methods, commonly used in the literature, rely on comparison tests, each comparing different features of the data and testing the hypothesis of equality.

The advanced methods evaluate the magnitude of the effect size of the side for each distribution parameter and their respective significances. The overall significance of the effect of "Side" is assessed with the LRT. The practical impact of "Side" on the model is evaluated using k-fold cross-validation, with Root Mean Square Error Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics.

Figure 3.1



Representation Symmetry Assessment Methodology

3.2.1 Basic Methods

Wilcoxon Signed-Rank Test (for Median Comparison)

In symmetry assessment, the Wilcoxon Signed-Rank Test Wilcoxon, 1945 is used to compare the medians of paired samples. Being a nonparametric statistical testIt is particularly useful when the assumptions of the paired t-test, such as normality, are not met. In this context the Wilcoxon Signed-Rank Test is used to compare the PPD values between two contralateral teeth sites to determine if there is a significant difference in their medians.

The Wilcoxon Signed-Rank Test evaluates the following hypotheses:

- **Null Hypothesis** (H_0): The median difference between the paired samples is zero, indicating no difference in PPD values between the two sites.
- **Alternative Hypothesis** (H_a): The median difference between the paired samples is not zero, indicating a difference in PPD values between the two sites.

Test Procedure

1. **Calculate Differences**: For each pair of observations, calculate the difference $d_i = x_i - y_i$, where x_i and y_i are the PPD values of the two contralateral sites for the i-th pair.

2. **Exclude Zero Differences**: Exclude any pairs where the difference d_i is zero.

3. **Rank Absolute Differences**: Rank the absolute differences $|d_i|$ from smallest to largest, assigning average ranks in the case of ties.

4. **Assign Signs to Ranks**: Assign the signs of the original differences d_i to the corresponding ranks.

5. **Calculate Test Statistic**: Sum the ranks of the positive differences W^+ and the ranks of the negative differences W^-. The test statistic W is the smaller of these two sums:

$$W = \min(W^+, W^-) \tag{3.1}$$

Test Statistic Distribution

Under the null hypothesis, the test statistic W follows a specific distribution, which can be approximated by a normal distribution for large sample sizes. The mean μ_W and standard deviation σ_W of W under the null hypothesis are given by:

$$\mu_W = \frac{n(n+1)}{4} \tag{3.2}$$

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}} \tag{3.3}$$

where n is the number of non-zero differences.

Calculation Steps

1. **Compute the Test Statistic**: Calculate the Wilcoxon test statistic W.

2. **Determine the p-value**: Using the normal approximation, the p-value is computed as:

$$Z = \frac{W - \mu_W}{\sigma_W} \tag{3.4}$$

Compare the calculated Z-value to the critical values of the standard normal distribution to determine the p-value.

Decision Rule

- **Reject** H_0 if the p-value is less than the significance level (0.05), indicating that there is a significant difference in the medians of the PPD values between the two sites.
- **Fail to reject** H_0 if the p-value is greater than the significance level, indicating that there is no significant difference in the medians of the PPD values between the two sites.

The Wilcoxon Signed-Rank Test (Wilcoxon, 1945) is a robust nonparametric test that does not require the assumption of normality, making it ideal for analyzing PPD values, which may not follow a normal distribution. This test provides a reliable method for comparing the medians of paired samples in periodontal research.(Oliveira, 2004)

Levene's Test for Variance Comparison

Levene's Test (Levene, 1960) is used to assess the equality of variances for a variable calculated for two or more groups. It is particularly useful when the assumption of homogeneity of variances, a key assumption in many parametric tests such as ANOVA, is in question. In the context of periodontal research, Levene's test can be applied to compare the variances of PPD values between different groups or contralateral teeth sites.

Levene's test evaluates the following hypotheses:

- **Null Hypothesis** (H_0): The variances of the PPD values are equal across the groups.
- **Alternative Hypothesis** (H_a): The variances of the PPD values are not equal across the groups.

Test Procedure

1. **Compute Group Means or Medians**: Depending on the variant of Levene's test being used, compute the mean or median PPD value for each group.

2. **Calculate Deviations**: For each observation, calculate the absolute deviation from the group mean or median. Let $Z_{ij} = |Y_{ij} - \bar{Y}_i|$ or $Z_{ij} = |Y_{ij} - \tilde{Y}_i|$, where Y_{ij} is the PPD value for the j-th observation in the i-th group, $\bar{Y}_i$ is the group mean, and $\tilde{Y}_i$ is the group median.

3. **Run ANOVA on Deviations**: Perform a one-way ANOVA on the absolute deviations Z_{ij} to test for significant differences among group variances.

Levene's test statistic W is computed as follows:

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{\cdot i} - Z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{\cdot i})^2} \tag{3.5}$$

where:

- N is the total number of observations,
- k is the number of groups,
- N_i is the number of observations in the i-th group,
- $Z_{\cdot i}$ is the mean of the absolute deviations in the i-th group,
- $Z_{\cdot\cdot}$ is the overall mean of the absolute deviations.

Calculation Steps

1. **Compute Absolute Deviations**: Calculate the absolute deviations from the group mean or median.

2. **Perform ANOVA**: Run a one-way ANOVA on these deviations to compute the test statistic W.

3. **Determine the p-value**: Compare the test statistic W to the critical value from the F-distribution with $k-1$ and $N-k$ degrees of freedom to determine the p-value.

Decision Rule

- **Reject** H_0 if the p-value is less than the significance level (commonly 0.05), indicating that there is a significant difference in the variances of the PPD values across the groups.
- **Fail to reject** H_0 if the p-value is greater than the significance level, indicating that there is no significant difference in the variances of the PPD values across the groups.

Levene's test is robust to departures from normality and is particularly valuable for testing the assumption of equal variances in parametric tests. By ensuring homogeneity of variances, it helps validate the use of parametric methods in the analysis of PPD data in periodontal research.

Kolmogorov-Smirnov (K-S) Test

The Kolmogorov-Smirnov (K-S) test (Massey Jr, 1951) is a nonparametric statistical test that evaluates the goodness of fit between a sample distribution and a reference probability distribution (one-sample K-S test) or between two sample distributions (two-sample K-S test). It is widely used to determine if a dataset follows a specific distribution or to compare the distributions of two independent samples.

The one-sample K-S test assesses how well a sample matches a specified theoretical distribution. It is particularly useful for testing the normality of data.

For a given cumulative distribution function (CDF) $F(x)$ and an empirical distribution function (EDF) $F_n(x)$ based on n observations, the K-S statistic D_n is:

$$D_n = \sup_x |F_n(x) - F(x)| \tag{3.6}$$

where $\sup_x$ denotes the supremum over all points x.

The two-sample K-S test compares two independent samples to determine if they come from the same distribution, without assuming normality.

Given two empirical distribution functions $F_{n1}(x)$ and $F_{n2}(x)$ from two samples of sizes n_1 and n_2, the K-S statistic $D_{n1,n2}$ is:

$$D_{n1,n2} = \sup_x |F_{n1}(x) - F_{n2}(x)| \tag{3.7}$$

Calculation Steps:

1. **Calculate the Empirical Distribution Function (EDF)**: For each sample, compute the EDF, the proportion of values less than or equal to each value x.
2. **Compute the Maximum Difference**: Identify the maximum absolute difference between the theoretical CDF and the EDF (one-sample) or between the two EDFs (two-sample).
3. **Critical Values and P-value**: Compare the K-S statistic D to the critical values from the K-S distribution to assess significance. The p-value can be calculated from the K-S distribution for the test statistic.

Hypotheses

- **Null Hypothesis (H_0)**: The data follows a specified distribution (one-sample) or there is no difference between the two distributions (two-sample).
- **Alternative Hypothesis (H_a)**: The data does not follow the specified distribution (one-sample) or there is a difference between the two distributions (two-sample).

Decision Rule

- **Reject** H_0 if the K-S statistic is greater than the critical value corresponding to the desired significance level, or if the p-value is less than the significance level (commonly 0.05).
- **Fail to reject** H_0 if the K-S statistic is less than or equal to the critical value, or the p-value is greater than the significance level.

The Kolmogorov-Smirnov test is powerful for its nonparametric nature, allowing it to be used with various distributions without assuming a specific distributional shape. This makes it invaluable in exploratory data analysis and hypothesis testing where the distribution of the data is unknown.

Bhattacharyya Coefficient

The Bhattacharyya Coefficient (BC) (Bhattacharyya, 1943) measures the similarity between two probability distributions, ranging from 0 (no overlap) to 1 (identical dis-

tributions). It is widely used in statistics and machine learning for comparing distributions.

For two discrete probability distributions p and q over the same domain X, the BC is defined as:

$$BC(p, q) = \sum_{x \in X} \sqrt{p(x)q(x)} \tag{3.8}$$

For continuous distributions, it is defined using the integral:

$$BC(p, q) = \int \sqrt{p(x)q(x)}\, dx \tag{3.9}$$

where $p(x)$ and $q(x)$ are the probability density functions of the two distributions.

A higher BC indicates greater overlap and similarity between the distributions. It is extensively used in pattern recognition, classification algorithms, and image processing to evaluate decision boundaries and feature distribution overlaps.

Spearman Correlation Coefficient

The Spearman correlation coefficient (denoted as ρ or r_s) is a nonparametric measure of rank correlation, assessing the statistical dependence between the rankings of two variables (Spearman, 1904). It evaluates how well the relationship between two variables can be described using a monotonic function, with values ranging from -1.0 to 1.0. A correlation of 1.0 indicates a perfect positive correlation of ranks, -1.0 indicates a perfect negative correlation of ranks, and 0.0 indicates no correlation between the ranks.

The formula for the Spearman correlation coefficient is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{3.10}$$

where d_i is the difference between the ranks of each pair of observations, and n is the number of observations (Zar, 2005).

Properties of Spearman's ρ

- **Unit-less**: The correlation coefficient is dimensionless.
- **Symmetry**: $\rho_{xy} = \rho_{yx}$, indicating that the correlation between x and y is the same in either order (Zar, 2005).

- **Range**: $-1 \leq \rho \leq 1$. A value of 1 implies a perfect positive rank correlation, -1 implies a perfect negative rank correlation, and 0 implies no rank correlation.

In this dissertation, we use Spearman's correlation coefficient because it does not require the data to be normally distributed and can be used with ordinal or non-parametric data (Zar, 2005). Given that periodontal pocket depth (PPD) can be defined both as continuous and ordinal data, Spearman's correlation is particularly suitable. Moreover, we focus solely on positive correlations, as negative correlations are not biologically plausible, and the data supports this focus.

Spearman's correlation measures the strength and direction of a monotonic relationship, whether it is always increasing or always decreasing, but not necessarily at a constant rate (Zar, 2005). It is less sensitive to outliers since it uses rank values instead of actual data values. This makes it appropriate for analyzing non-linear relationships and handling non-parametric and ordinal data effectively. Thus, it is well-suited for our analysis of periodontal data, ensuring robustness and reliability in our findings.

3.2.2 Advanced Methods

Assuming symmetry we create one variable that includes both sides observed PPD values together with a dictomic variable indicating the side of the mouth ("D1" for the upper right side and "D2" for the upper left side). To verify if the effect size of "side" affect the probabilistic distribution of periodontal disease indicator we proceed with a method that includes GAMLSS. (Rigby & Stasinopoulos, 2005)

Determination of the Probability Distribution of PPD

Before start fitting the GAMLSS models we determine the underlying probability distribution of PPD, using the `fitDist` function. This process is essential in statistical modeling to understand the data's underlying distribution, which is crucial for further analysis, predictions, accurate GAMLSS models, or model validation. This function is applied to a variable representing PPD values from both contralateral sites. The distribution type specified is for real positive values (`realplus`), controlled by parameters provided in `ctrl`, allowing for a maximum number of 500 iterations for the fitting algorithm to try and converge to a solution. This parameter ensures that the fitting process has enough iterations to adequately converge, but also prevents it from running indefinitely if convergence is not achieved within the specified number of iter-

ations. The output of `fitDist` function is a sequence of more adequate probability distributions ordered from the lower to the higher Akaike Information Criteria (AIC).

The `fitDist` function employs Maximum Likelihood Estimation (MLE) to identify the parameters of a given statistical distribution that maximize the likelihood function. The likelihood function $L(\theta; x)$ measures the probability of observing the data x given the parameters θ (Casella & Berger, 2002; Murphy, 2012).

Given observed data points $x = \{x_1, x_2, \ldots, x_n\}$ and a probability density function Probability Density Function (PDF) $f(x; \theta)$ parameterized by θ, the likelihood function is:

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) \tag{3.11}$$

For computational convenience, the log-likelihood function is used:

$$\ell(\theta; x) = \log L(\theta; x) = \sum_{i=1}^{n} \log f(x_i; \theta) \tag{3.12}$$

The goal of MLE is to find the parameter values $\hat{\theta}$ that maximize the log-likelihood function:

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta; x) \tag{3.13}$$

The optimization process involves:

1. **Choosing an Initial Guess**: Start with an initial guess for the parameters θ_0.
2. **Iterative Update**: Use numerical optimization techniques such as Newton-Raphson (Burden & Faires, 2010) to iteratively update the parameter estimates to increase the log-likelihood function.
3. **Convergence**: Continue until convergence, where subsequent updates result in negligible changes in the log-likelihood value.

The `fitDist` function utilizes MLE by:

1. Defining the log-likelihood function based on the chosen distribution.
2. Using numerical optimization methods to find the parameter values that maximize this function.
3. Returning the parameter estimates that best fit the data according to the specified distribution.

The PPD has an Ex-Gaussian distribution, which is characterized by three parameters: μ, σ, and τ. The PDF is given by:

$$f(x;\mu,\sigma,\tau) = \frac{1}{\tau}\exp\left(\frac{1}{\tau}(\mu - x + \frac{\sigma^2}{2\tau})\right)\Phi\left(\frac{x-\mu-\frac{\sigma^2}{\tau}}{\sigma}\right) \tag{3.14}$$

where $\Phi(\cdot)$ is the Cumulative Distribution Function (CDF) of the standard normal distribution (Evans et al., 1999).

Given observed data points $x = \{x_1, x_2, \ldots, x_n\}$, the log-likelihood function is:

$$\ell(\mu,\sigma,\tau;x) = \sum_{i=1}^{n}\log f(x_i;\mu,\sigma,\tau) \tag{3.15}$$

Substituting the PDF of the Ex-Gaussian distribution, we get:

$$\ell(\mu,\sigma,\tau;x) = \sum_{i=1}^{n}\left[\log\left(\frac{1}{\tau}\right) + \frac{1}{\tau}\left(\mu - x_i + \frac{\sigma^2}{2\tau}\right) + \log\Phi\left(\frac{x_i-\mu-\frac{\sigma^2}{\tau}}{\sigma}\right)\right] \tag{3.16}$$

The MLE estimates $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\tau}$ are obtained by maximizing the log-likelihood function:

$$\hat{\mu},\hat{\sigma},\hat{\tau} = \arg\max_{\mu,\sigma,\tau}\ell(\mu,\sigma,\tau;x) \tag{3.17}$$

The numerical optimization method used to maximize $\ell(\mu,\sigma,\tau;x)$ was the Newton-Raphson method (Burden & Faires, 2010).

Evaluation of Size Effects of Side With GAMLSS Models

The Generalized Additive Models for Location, Scale, and Shape is a flexible framework that extends Generalized Linear Models Generalized Linear Models (GLM)s and Generalised Additive Models (GAM)s by allowing not just the mean (or location) but also other parameters of the distribution (such as scale and shape) to be modeled as functions of predictors. This flexibility makes GAMLSS particularly powerful for modeling complex data structures that are not adequately described by traditional statistical models.

The general form of a GAMLSS for each parameter can be expressed as:

$$y_i \sim F(\mu_i,\sigma_i,\nu_i,\tau_i) \tag{3.18}$$

$$g_k(\theta_k) = X_k\beta_k + \sum_{j=1}^{J_k} s_{kj}(x_{kj}), \tag{3.19}$$

where:

- y_i are the response variables observed for each subject or experimental unit.
- F represents the family of distributions used, which could be normal, binomial, Poisson, or any other suitable distribution, depending on the nature of the data and the specific requirements of the analysis.
- g_k is a link function
- X_k is the design matrix
- β_k are coefficients
- s_{kj} are smooth functions of the predictors
- $\mu_i, \sigma_i, \nu_i, \tau_i$ are the parameters of the distribution that are modeled. Each parameter can be linked to a set of predictors x_{ij} through a link function g_k.
- β_{kj} are coefficients to be estimated, indicating the influence of predictors on each parameter of the distribution.

This model structure allows each parameter of the distribution to depend on the predictors, providing a comprehensive way to model the conditional distribution of the response variable.

GAMLSS models - Performance and comparison with LRT and 5-Fold cross validation

The LRT is a statistical test used to compare the goodness of fit between two competing statistical models (Wilks, 1938). It is defined by the test statistic:

$$D = -2\log\left(\frac{L(\text{null model})}{L(\text{alternative model})}\right) \tag{3.20}$$

where L(null model) is the likelihood of the null model (assuming no effect of the predictors) and L(alternative model) is the likelihood of the alternative model (which includes additional predictors or different parameterizations). The LRT statistic, D, follows a chi-square distribution under the null hypothesis, allowing us to assess whether

the inclusion of additional predictors significantly improves the model fit. The LRT develops as follows:

1. Compute the MLEs for both the null and alternative models.
2. Calculate the likelihoods of these models at their respective MLEs.
3. Use the formula for D to determine the likelihood ratio, which quantifies how much more likely the data is under one model compared to the other.
4. Refer to the chi-square distribution to find the p-value associated with D, assessing the statistical significance of the observed ratio.

K-fold cross-validation was performed to evaluate the predictive performance of both models, ensuring that the inclusion of the side factor improves the model's robustness. The data was split into k subsets, and the model was trained on $k-1$ subsets while the remaining subset was used for validation. This process was repeated k times (Arlot & Celisse, 2010).

Within this framework, we compute the following metrics:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.21}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.22}$$

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} \tag{3.23}$$

where:

- y_i are the observed values.
- $\hat{y}_i$ are the predicted values by the model.
- $\overline{d}$ represents the mean difference in RMSE or MAE between the null model and the alternative model across the validation folds.
- s_d is the standard deviation of these differences.
- n is the number of folds in the cross-validation.

The k-fold cross-validation develops as follows:

1. Split the data into n folds for cross-validation.
2. For each fold, fit the model to the $n-1$ training folds and calculate the RMSE and MAE on the validation fold.
3. Aggregate the RMSE and MAE across all folds to assess overall model performance.
4. Perform a paired t-test to compare the mean RMSE and MAE values obtained from the null and alternative models across the folds. This test evaluates whether the differences in performance metrics are statistically significant, providing insight into the predictive accuracy improvements gained by the alternative model.

3.2.3 Symmetry Measure Function

This section introduces a novel method for evaluating the symmetry of periodontal probing depths. The function $SM = f(A, A')$ is exponential decay type function designed to calculate the Symmetry Measure SM for a pair of measurements, A and A'.

This function is conceptualized as a membership function within the fuzzy symmetry framework. Membership functions in fuzzy logic signify the extent to which a pair A and A' is a member of a particular set, typically ranging between 0 and 1. The exponential component of the function ensures outputs within this spectrum. Consequently, it offers a quantification of the symmetry between A and A', where values nearing 1 suggest pronounced symmetry.

The function SM is defined as:

$$SM = f(A, A') = \exp\left(-\frac{|A - A'|}{\alpha + \beta \cdot \frac{A+A'}{2}}\right) \tag{3.24}$$

where:

$\exp(x)$ is the exponential function, e^x.

$|A - A'|$ represents the absolute difference between the values A and A'.

α and β are parameters that scale and influence the function.

$\frac{A+A'}{2}$ is the arithmetic mean of A and A'.

α – A baseline scaling factor, ensuring the denominator is never zero.

β – Controls the influence of the average value of A and A' on the normalization of the difference.

The function's dynamics are influenced by the parameters α and β. Manipulating these parameters allows for the calibration of the function's sensitivity and adaptability to the disparities between A and A', facilitating a refined symmetry assessment.

The parameter α is a constant that scales the sensitivity of the SM to differences between A and A', ensuring these differences are reflected in the SM. Its magnitude depends on the scale of A and A'. A smaller α, like 0.1, appropriately scales the sensitivity for small differences between A and A', ensuring these differences are reflected in the SM.

The mathematical influence of the parameters α and β can be described in three key points. First, α ensures the denominator never approaches zero, preventing the SM from becoming infinitely large and maintaining bounded similarity scores. Second, α provides a predetermined baseline sensitivity to differences between A and A', governed primarily by α when β is low or zero. Finally, with a fixed α, β play a moderation role by adjusting the function's sensitivity to the scale of A and A', making its influence more transparent and interpretable.

Setting parameter *alpha* to one

Setting α to a constant value of 1 has specific mathematical and practical implications:

- Comparability: Standardizes the base level of similarity across different pairs, making measures more comparable.
- Simplicity in Interpretation: Simplifies interpretation of changes in the SM, attributing adjustments to β and the values of A and A'.
- Moderate Sensitivity: Ensures small differences are neither exaggerated nor ignored, beneficial for detecting small but meaningful differences.
- Flexibility in Scaling: Allows β to flexibly scale the function's sensitivity to the relative magnitude of A and A'.

Parameter *beta*

The formula incorporates the average magnitude of A and A' through the term $\frac{A+A'}{2}$. As the parameter β increases, the influence of the average magnitude of the measurements on the similarity measure (SM) becomes more pronounced. This implies that with higher values of β, the formula takes the magnitudes of A and A' into account more significantly. In the context of oral health, an increasing SM with higher β suggests that strong correlations in measurements, such as periodontal pocket depths, lead to differences being considered less impactful on the symmetry measure. This can be particularly relevant for assessing bilateral symmetry in periodontal health, where a higher β value compensates for measurement differences, suggesting a form of "functional" symmetry despite the presence of minor variations.

Continuous versus discrete

To adapt this formula for continuous distributions, we consider the probability density functions (pdfs) of the continuous random variables X and Y, denoted as $f_X(x)$ and $f_Y(y)$ respectively. We replace the scalar values A and A' with their expected values $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, and modify the formula accordingly.

First, we compute the expected values of the distributions X and Y:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \tag{3.25}$$

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy \tag{3.26}$$

Then, we substitute the expected values into the similarity measure formula:

$$SM = \exp\left(-\frac{|\mathbb{E}[X] - \mathbb{E}[Y]|}{\alpha + \beta \cdot \frac{\mathbb{E}[X]+\mathbb{E}[Y]}{2}}\right) \tag{3.27}$$

To adapt this formula for continuous distributions, we consider the probability density functions (pdfs) of the continuous random variables X and Y, denoted as $f_X(x)$ and $f_Y(y)$ respectively. We replace the scalar values A and A' with their expected values $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, and modify the formula accordingly.

First, we compute the expected values of the distributions X and Y:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx \tag{3.28}$$

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy \tag{3.29}$$

Then, we substitute the expected values into the similarity measure formula:

$$SM = \exp\left(-\frac{|\mathbb{E}[X] - \mathbb{E}[Y]|}{\alpha + \beta \cdot \frac{\mathbb{E}[X]+\mathbb{E}[Y]}{2}}\right) \tag{3.30}$$

The function SM can be viewed as a continuous function $f(A, A')$ defined over the real numbers, where A and A' are continuous variables. In this continuous context, $f(A, A')$ describes a smooth surface over the (A, A')-plane. This smoothness implies that small changes in A or A' result in small changes in the value of $f(A, A')$.

- If A is close to A', $|A - A'|$ is small, making SM close to $\exp(0) = 1$, indicating high symmetry.
- If A is far from A', $|A - A'|$ is large, making the exponent more negative and SM smaller, indicating low symmetry.

Analysing specific values of A and A' involves sampling this continuous function at discrete points:

$$SM = f(A_i, A_j) \tag{3.31}$$

where A_i and A_j are specific discrete values. This sampling process converts the continuous function into a set of discrete values while maintaining the underlying properties of the continuous function.

The SM function expression (3.24) exhibits behaviour similar to a Gaussian (normal) distribution with the exponential decay ensuring that as $|A - A'|$ increases, SM decreases rapidly, similar to the tails of a Gaussian distribution and when $A \approx A'$, $|A - A'| \approx 0$, resulting in $SM \approx 1$, resembling the peak of a Gaussian distribution.

This behavior is maintained both in the continuous interpretation and in the discrete analysis, ensuring that the function preserves its Gaussian-like characteristics even when evaluated at discrete points.

The SM measure accounts for the relative size of A and A', making the symmetry measure dimensionless and thus comparable across different magnitudes. The parameters α and β control the scaling of this normalisation, ensuring the measure remains comparable regardless of the range of A and A'.

Gamma Parameter (γ)

The gamma parameter (γ) was introduced to capture the *directionality* of asymmetry between two paired measurements, denoted A and A', such as probing depths from contralateral periodontal sites. Rather than measuring the magnitude of the difference, this function identifies its sign, and is defined as:

$$\gamma = \frac{|A - A'|}{A - A'}$$

This expression evaluates to:

- $\gamma = +1$ if $A > A'$ (indicating asymmetry toward the right),
- $\gamma = -1$ if $A < A'$ (indicating asymmetry toward the left),
- γ is undefined when $A = A'$, a condition that was handled in computation by assignment of 0.

By isolating the direction of discrepancy, the gamma function complements other symmetry measures that quantify magnitude. It provides a binary encoding of spatial orientation, which can be leveraged as an explanatory feature in predictive models—particularly those designed to learn from directional patterns in biological structures.

3.3 Visualization of Similarity - Kernel Density Estimates

Throughout this work, Kernel Density Estimates (KDEs) were used to compare distributions of PPD variables in different contexts, specifically for assessing symmetry and similarity, where the symmetry is a particular case of similarity.

The first application of KDE plots was in assessing the symmetry of two contralateral PPD variables. The second application involved using KDE to assess the performance of H-D imputation during data preparation for Mother-Daughter Imputation Method (MoDau) imputation. Additionally, KDE was used to compare original data with synthetic data to evaluate dissimilarity, and to assess the results of MoDau imputation by comparing pre- and post-imputation distributions for similarity.

In the first three applications, the PPD was treated as discrete, with a small smoothing parameter, using the default settings of the R density function. In the last applica-

tion, the PPD was treated as continuous, and a higher smoothing parameter was used, through optimised KDE parameters.

The KDE for discrete data provides the probability that a discrete random variable equals a specific value, aiding in understanding the distribution of discrete periodontal measurements like PPD. Complementarily, KDE for continuous data, estimates the probability density function of a continuous random variable, offering a smooth estimate of the data's distribution.

Combining discrete KDE and continuous KDE allows for a comprehensive analysis of PPD data. The discrete KDE facilitates categorical analysis of discrete variable distributions, while the continuous KDE provides a smooth, continuous perspective, highlighting overall trends and patterns.

3.3.1 Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric method used to estimate the PDF of a random variable. Unlike parametric methods, KDE does not assume any specific underlying distribution for the data. Instead, it builds the density estimate based on the observed data points.

Mathematically, the KDE of a random variable X at a point x is given by:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{3.32}$$

where:

n is the number of data points,

h is the bandwidth parameter that controls the smoothness of the resulting density curve,

K is the kernel function, typically a symmetric and non-negative function that integrates to one,

x_i are the observed data points.

The choice of the kernel function K and the bandwidth h are crucial for the quality of the density estimate. Common kernel functions include the Gaussian, Epanechnikov, and uniform kernels. The bandwidth h determines the width of the kernel and thus the level of smoothing applied to the data.

A larger bandwidth results in a smoother density estimate by averaging over a larger number of data points, potentially oversmoothing the data and missing impor-

tant features. Conversely, a smaller bandwidth produces a more detailed and variable density estimate by focusing on a narrower range of data points, which can lead to overfitting and capturing noise in the data.

Kernel Density Estimation is widely used in various fields such as statistics, machine learning, and data analysis for visualizing the underlying distribution of data, identifying modes, and detecting outliers. The KDE parameters include the kernel function, the bandwidth, and the adjustment parameter.

Kernel Function

The kernel function plays a crucial role in KDE by determining how data points influence the estimated density. It affects the smoothness, shape, and local weighting of the density estimate. Specifically, the kernel function in KDE serves as a weighting function that shapes the influence of each data point on the estimated density at a given point. The kernel function assigns weights to data points based on their distance from the point at which the density is being estimated. Points closer to the estimation point receive higher weights, while those farther away receive lower weights. The kernel function helps smooth the density estimate by spreading the influence of each data point over a local neighbourhood, producing a continuous and smooth estimate of the underlying probability density function. Different kernel functions can produce slightly different shapes for the density estimate, influencing the fine details of the estimate. The kernel functions considered for selection include:

The **Gaussian Kernel** is widely used due to its smooth and symmetric shape. It places more weight on points closer to the estimation point, with weights decreasing exponentially as the distance increases. Is represented as:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{3.33}$$

The **Epanechnikov Kernel** is considered optimal in a mean square error sense. It is parabolic in shape and assigns zero weight to points beyond a certain distance. Is represented as:

$$K(u) = \frac{3}{4}(1 - u^2) \quad \text{for} \quad |u| \leq 1 \tag{3.34}$$

The **rectangular (uniform) kernel** gives equal weight to all points within a certain distance and zero weight to points outside this range. It produces a blocky, less

smooth density estimate. Is represented as:

$$K(u) = \frac{1}{2} \quad \text{for} \quad |u| \leq 1 \tag{3.35}$$

The **triangular kernel** linearly decreases the weight as the distance increases, producing a somewhat smoother estimate than the rectangular kernel. Is represented as:

$$K(u) = (1 - |u|) \quad \text{for} \quad |u| \leq 1 \tag{3.36}$$

The **biweight (quartic) kernel** provides a smooth estimate and assigns zero weight to points beyond a certain distance. Is represented as:

$$K(u) = \frac{15}{16}(1 - u^2)^2 \quad \text{for} \quad |u| \leq 1 \tag{3.37}$$

The **cosine kernel** assigns weights based on the cosine function and is known for its smoothness. Is represented as:

$$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right) \quad \text{for} \quad |u| \leq 1 \tag{3.38}$$

$$K(x, y) = \frac{x^T y}{\|x\|\|y\|} = \frac{x}{\|x\|} \cdot \frac{y}{\|y\|} \tag{3.39}$$

The **optcosine kernel** is similar to the cosine kernel but optimized for certain properties. Is represented as:

$$K(u) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right) \quad \text{for} \quad |u| \leq 1 \tag{3.40}$$

The choice of kernel function depends on the specific characteristics of the data, the context, the requirements of the analysis, and the desired properties of the density estimate.

Bandwidth

The bandwidth h (see equation 3.32 determines the width of the kernel function and directly affects the smoothness of the density estimate. A larger bandwidth results in a smoother density estimate by averaging over a larger number of data points, while a smaller bandwidth provides a more detailed and variable density estimate by focusing on a narrower range of data points.

The choice of bandwidth is critical as it balances the trade-off between bias and variance in the density estimate. A larger h reduces variance but increases bias, resulting in a smoother but potentially oversimplified density estimate. Conversely, a smaller h reduces bias but increases variance, resulting in a more detailed and potentially overfit density estimate. Selecting an appropriate bandwidth is essential for achieving a reliable and accurate density estimation.

Adjustment

The adjustment parameter (a) is a multiplicative factor applied to the bandwidth (h), which scales the width of the kernel function and thereby influences the smoothness of the density estimate. By adjusting the bandwidth, the adjustment parameter allows fine-tuning of the density estimate. Mathematically, if the original bandwidth is h_0, the adjusted bandwidth h is given by:

$$h = a \cdot h_0,$$

where a is the adjustment parameter.

A larger adjustment parameter ($a > 1$) increases the bandwidth, resulting in a smoother density estimate by averaging over a larger number of data points. Conversely, a smaller adjustment parameter ($a < 1$) decreases the bandwidth, producing a more detailed and variable density estimate by focusing on a narrower range of data points.

Number of points (n)

The `n` parameter sets the number of points at which the density is estimated. Setting `n = n` means that the density estimate will be calculated at `n` equally spaced points over the range of the data. This affects the resolution of the density plot. A higher `n` results in a finer resolution and a more detailed density estimate, while a lower `n` would produce a coarser estimate.

KDE Parameters

When PPD is treated as discrete we adopt the default parameters for the kernel density estimation (KDE) of `R density()` function. which are the folowing:

1. Kernel: The default kernel is Gaussian. The Gaussian kernel uses the Gaussian

(normal) distribution as the kernel function is symmetric and smooth, making it suitable for most data distributions. It ensures that the estimated density function is smooth and continuous.

2. Bandwidth selection method: The bandwidth, which determines the width of the kernel function used to estimate the density, was calculated using the rule of thumb for the standard deviation of the data, also known as "Silverman's Rule of Thumb"(`nrd0`). It calculates the bandwidth as

$$h = 0.9 \cdot \min(\sigma, IQR/1.34) \cdot n^{-1/5}$$

where σ is the standard deviation of the data, IQR is the interquartile range, and n is the number of data points (Silverman, 1986). This method provides a reasonable default bandwidth for many datasets.

3. Adjustment parameter: The default adjustment parameter is 1, meaning that the bandwidth is used as calculated by the `"nrd0"` method without any additional scaling.

4. Number of points (`n`): The default number of points at which the density is estimated is 512. This determines the resolution of the density plot. A higher `n` results in a finer resolution and a more detailed density estimate, while a lower `n` would produce a coarser estimate.

This setup provides a smooth and reliable estimate of the data's underlying density distribution for most datasets.

3.3.2 Optimization of the Kernel Parameters

When PPD is treated as continuous , in the context of comparing simulated with MoDau imputed values, our goal is to check how the overall trends and patterns match. This approach was used in the validation of MoDau to compare PPD pre- and post-imputation. To achieve this, we optimized the kernel parameters. Firstly, we decided on a range of potential bandwidths, kernel functions, and adjustments to optimize the KDE. These parameters were systematically combined to create a comprehensive parameter grid using the expand.grid function.

The grid search included bandwidths ranging from 0.1 to 2.5, incremented by 0.1. The kernel functions included Gaussian, Epanechnikov, Rectangular, Triangular, Bi-

weight, Cosine, and Optcosine. Adjustments ranged from 0.5 to 2.5, incremented by 0.1. The best kernel parameter combinations included the Biweight or Cosine kernel functions (more often the first one), with a consistent bandwidth of 2 and adjustment of 2 across all comparisons. This procedure was used for every site comparison.

3.3.2.1 Boostrap Method

Behind the visual evaluation, we also employ the bootstrap method to compare KDE across different noise levels and assess the similarity between the Daughter model's predictions and the simulated data.

The bootstrap methods, introduced by Efron and Tibshirani, 1994, are a class of statistical techniques that involve resampling with replacement from a dataset to estimate the sampling distribution of a statistic. These methods are particularly useful when the theoretical distribution of the statistic is complex or unknown. From an original sample of size n, B bootstrap samples of size n are generated by randomly drawing observations with replacement. For each bootstrap sample, the statistic of interest (e.g., mean, variance) is calculated. This process is repeated a large number of times—in our case, 10,000 times—to build a distribution of the statistic.

Each replicate is generated by randomly sampling rows from the original data matrix, allowing for the possibility that some rows may be repeated within a replicate while others may be omitted. This random selection process means that the composition of each bootstrap sample will likely vary, leading to different sets of observations being included in different replicates.

The bootstrap distribution of the statistic can be used to estimate confidence intervals. Common methods include the percentile method, the bias-corrected and accelerated method, and the standard error method. In our work, the confidence intervals for the difference in densities are estimated using the percentile method. The percentile method constructs confidence intervals based on the percentiles of the bootstrap distribution of the statistic of interest.

Bootstrap Procedure

Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ representing PPD measurements, the bootstrap procedure involves the following steps:

1. **Resampling**: Generate B bootstrap samples from the original dataset. Each bootstrap sample X^* is obtained by sampling n observations with replacement from

X.

2. **Kernel Density Estimation**: For each bootstrap sample X^*, estimate the kernel density function. The kernel density estimate $\hat{f}(x)$ is given by equation 3.32

where $K(\cdot)$ is the kernel function and h is the bandwidth parameter .

3. **Mean Difference Calculation**: Compute the mean difference in densities for each noise level by comparing the KDE from the original sample and the bootstrap samples. For two KDE $\hat{f}_1(x)$ and $\hat{f}_2(x)$, the mean difference is calculated as:

$$\Delta = \frac{1}{n} \sum_{i=1}^{n} \left| \hat{f}_1(x_i) - \hat{f}_2(x_i) \right| \tag{3.41}$$

4. **Confidence Intervals**: Construct the 95% confidence interval for the mean difference in densities using the bootstrap samples. The percentile method is typically used, where the confidence interval is determined from the empirical distribution of the bootstrap estimates.

In addition to bootstrap kernel mean differences comparisons, we use the Kolmogorov-Smirnov (KS) test to compare both distributions, providing both the KS statistic and the p-values. The specific details of the kernels, bandwidths, adjustments, mean differences, confidence intervals, KS statistics, and p-values are summarized in Table 21.

3.4 Clinical Perception of Symmetry

A sample of ten experienced dentists and periodontists were invited to score twenty pairs of pocket probing depth (PPD) values on a scale from zero to ten. The PPD values were presented within a bar chart. The columns representing the left side coloured in blue and the right side in green, and their width was represented visually by coloured bars and numbers, with the bars width represented by a number.

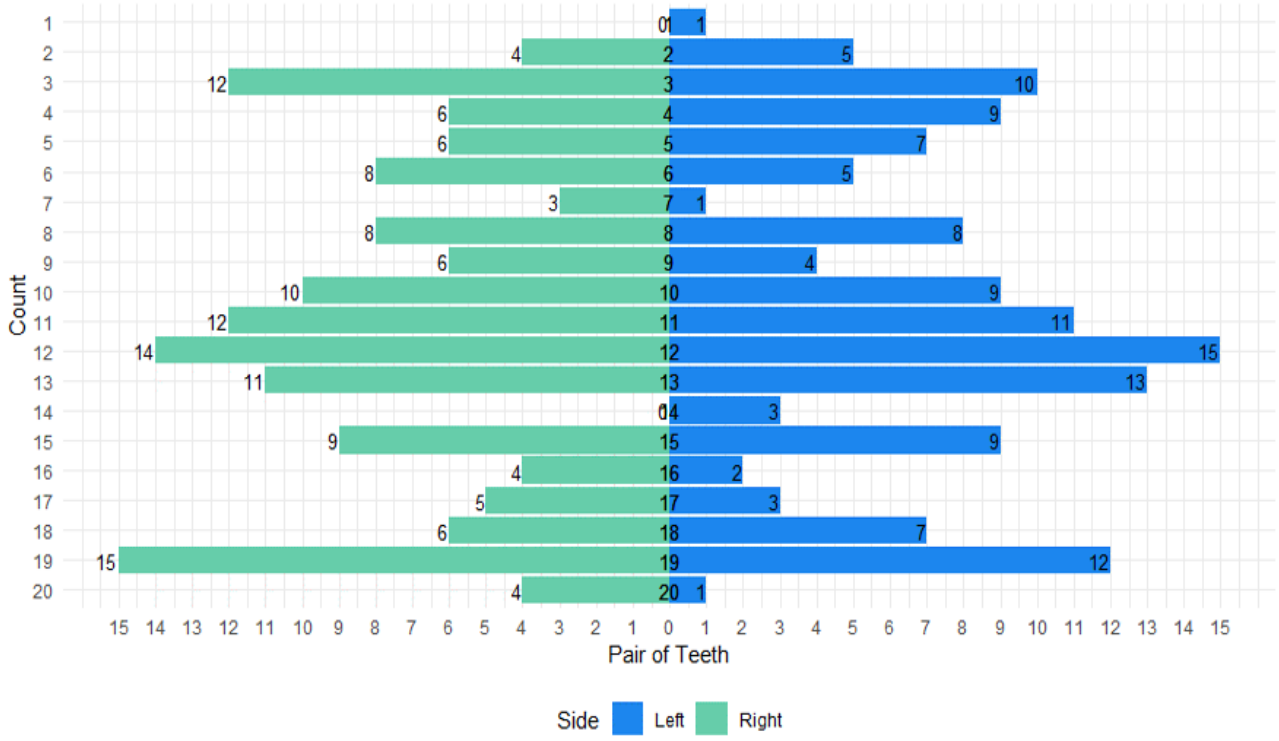
The goal is to correlate the scores assigned by dental professionals with those computed by SM function. The aim is to validate the reliability and accuracy of the computational method.

The data consists of probing depth measurements for pairs of teeth on both the left (blue) and right (green) sides of the mouth. Each pair of teeth is evaluated for symmetry by a group of dental professionals who assign a symmetry score ranging from 0 to 10, where 10 indicates perfect symmetry and 0 indicates no symmetry.

The bar plot in Figure 3.2 illustrates the PPD measurements for each pair of teeth,

scored by the dental professionals.

Figure 3.2



Buterfly Bar Plot for Symmetry Score

Having concluded the description of the symmetry assessment methods, we now proceed with MoDau imputation methodology.

3.5 Mother-Daughter Imputation Methodology

The primary objective of our research is to develop a robust method for imputing Periodontal Pocket Depth (PPD) values for the six sites of the tooth on the upper left teeth, using the values from the corresponding sites on the upper right teeth and the soft labels. This study leverages data from the NHANES 2011-2012, along with the Symmetry Measure SM.

To achieve this objective, we employed an eXtreme Gradient Boosting (XGBoost) regression modeling approach (Chen & Guestrin, 2016), tailored to the NHANES data

and directional symmetry measures (γSM). Our process can be broken down into the following steps:

1. **Preparation of the NHANES Data:** - We used Hot Deck imputation to create a complete dataset, ensuring no missing values in the initial NHANES data. - We evaluated the quality of the complete dataset to confirm the integrity and reliability of the imputed values.

2. **Training of the Mother Model:** - The mother model was trained using the complete NHANES dataset and the directional symmetry measures (γSM). - Soft labels were generated from this model to serve as initial predictive indicators for the PPD values.

3. **Generation of Simulated Data:** - A simulated dataset was created to test the MoDau method. This involved computing the respective SM and γSM for the simulated values. - The quality of the simulated data was evaluated to ensure it accurately represented the characteristics of the original dataset.

4. **Training the `New_Mother` Model:** - The `New_Mother` model was trained on the simulated dataset to generate additional soft labels. This step is crucial if the right-side PPD values of the new data differ substantially from the NHANES 2011-2012 dataset. - These `New_Mother` soft labels update the mother model to align with the new dataset. An arbitrary number of observations can be selected from the new dataset for training; in this study, we used five observations, the minimum number that does not cause errors in subsequent computations. - Alternatively, this computational path can be excluded by suppressing the `New_Mother` related features if deemed unnecessary.

5. Training the `Daughter_Model`: - The `Daughter_Model` was trained using the right-side PPD values of the new dataset, `Mother_Model` soft labels, and `New_Mother` soft labels (if they exist). - The `Daughter_Model` predictions, referred to as MoDau values, were generated to impute the missing PPD values.

6. **Evaluation of MoDau Imputation:** - The performance of the MoDau imputation was thoroughly evaluated to ensure accuracy and reliability in predicting the left-side PPD values based on the right-side data and generated soft labels.

Through this structured approach, we aimed to enhance the imputation accuracy of periodontal data, leveraging advanced regression techniques and comprehensive data preparation and evaluation methods.

3.5.1 Imputation with Hot Deck

Before proceed with the MoDau imputation of upper left side PPD values using the machine learning technique XGBoost, it is essential to address the issue of missing data. Proper imputation of missing data is crucial to avoid biases and inaccuracies in the analysis, thereby ensuring the robustness of the results.

In this study, the Hot Deck imputation method (Andridge & Little, 2010) was employed as an initial step to handle the missing data. This method involves replacing each missing value with an observed response from a similar unit, known as a "donor." The similarity between units is typically determined based on specific matching variables, ensuring that the imputed values closely resemble the true values.

The dataset, post-imputation, was then subjected to further analysis and modeling, including the application of the XGBoost for predictive modeling of the PPD values.

The subsequent sections will discuss the implementation of the XGBoost approach for imputing the upper left side PPD values and the analysis performed to evaluate the effectiveness of this method.

Algorithm for Hot Deck Imputation

The Hot Deck imputation algorithm can be summarized in the following steps:

Algorithm 1 Hot Deck Imputation Algorithm

Require: Dataset $\mathbf{X}$ with missing values
Ensure: Dataset $\mathbf{X}$ with imputed values
1: **for** each missing value x_{ij} in $\mathbf{X}$ **do**
2: Identify a set of donor units D that are similar to the unit with the missing value based on selected matching variables
3: Randomly select a donor value d_{ij} from the set D
4: Replace x_{ij} with d_{ij}
5: **end for**

Let $\mathbf{X} = \{x_{ij}\}$ be a data matrix where i denotes the unit (tooth site) and j denotes the variable (PPD). Suppose some values x_{ij} are missing.

For each missing value x_{ij}:

1. Define a set of matching variables $M \subset \{1, 2, \ldots, j-1, j+1, \ldots, p\}$, where p is the total number of variables.

2. Calculate the similarity measure $S(i,k)$ between the unit with the missing value (unit i) and all other units k based on the matching variables in M. This was done using Euclidean distance metric:

$$S(i,k) = \sqrt{\sum_{m \in M} (x_{im} - x_{km})^2} \tag{3.42}$$

3. Identify the set of donor units D that have the smallest distances $S(i,k)$ and do not have missing values for x_{ij}.

4. Randomly select a donor value d_{ij} from the set D.

5. Replace the missing value x_{ij} with the donor value d_{ij}.

The Hot Deck imputation was carried out sequentially for different variables across all dental sites using the `hotdeck()` function from `VIM` package.

Evaluation of Hot Deck imputation

The quality of data after "Hot Deck" imputation was assessed by comparing pre- and post-imputation statistics. This comparison included evaluating medians, variances, and proportions of unique PPD values, as well as grouped PPD values greater than 3mm. Additionally, both pre- and post-imputation distributions were compared using the Kolmogorov-Smirnov (K-S) test and by visually examining Kernel Density Estimate (KDE) plots. These assessments ensured that the imputed data maintained the integrity and distributional characteristics of the original dataset, providing confidence in the accuracy and reliability of the imputation process.

The proportion test was applied to compare the proportion of a specific probing pocket depth (PPD) value (k) from a particular tooth site, as well as the proportion of PPD values greater than 3 mm ($k > 3$), before and after imputation. The objective of the test is to ascertain whether there is a significant difference in the proportions of k between these two sets of PPD values.

The test is performed as follow:

- **Successes**: The number of times the value k appears in the original and imputed data

$$\text{successes\_BI} = \sum_{i=1}^{n} \mathbf{1}_{\{I\_O[i,\text{"Original"}]=k\}} \tag{3.43}$$

$$\text{successes\_AI} = \sum_{i=1}^{n} \mathbf{1}_{\{I\_O[i,\text{"Imputed"}]=k\}} \tag{3.44}$$

where I_O is the source of Original and Imputed PPD values, and **1** is the indicator function that equals 1 if the condition is true and 0 otherwise.

- **Trials**: The total number of non-missing values in the original and imputed data.

$$\text{trials\_BI} = \sum_{i=1}^{n} \mathbf{1}_{\{\textit{Original value is not NA}\}} \tag{3.45}$$

$$\text{trials\_AI} = \sum_{i=1}^{n} \mathbf{1}_{\{\textit{Imputed value is not NA}\}} \tag{3.46}$$

Performing the Proportion Test The null hypothesis H_0 is that the proportions of k in the original and imputed data are equal:

$$H_0 : p_{BI} = p_{AI}$$

The alternative hypothesis H_A is that the proportions are different:

$$H_A : p_{BI} \neq p_{AI}$$

The test statistic for comparing two proportions is calculated as:

$$z = \frac{(\hat{p}_{BI} - \hat{p}_{AI})}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_{BI}} + \frac{1}{n_{AI}}\right)}} \tag{3.47}$$

where:

- $\hat{p}_{BI} = \frac{\text{successes\_BI}}{\text{trials\_BI}}$ is the sample proportion in the original data.
- $\hat{p}_{AI} = \frac{\text{successes\_AI}}{\text{trials\_AI}}$ is the sample proportion in the imputed data.
- $\hat{p} = \frac{\text{successes\_BI+successes\_AI}}{\text{trials\_BI+trials\_AI}}$ is the combined proportion.

If the p-value is less 0.05, we reject the null hypothesis and conclude that there is a significant difference in the proportions of k between the original and imputed data. If the p-value is greater than the significance level, we fail to reject the null hypothesis, suggesting that there is no significant difference in the proportions.

3.6 Mother-Daughter Models

The Mother and Daughter models, named by "M" and "D" respectively, followed by the predicted tooth type and site (e.g., M21DV / D21DV, where 21 represents the left upper central incisor and DV the Distal Vestibular site), were developed to predict each tooth of the left side of the mouth. The Mother models use the data from NHANES 2011/2012 and the variable γSM engineered with PPD data from each pair of contralateral sites. The Daughter models use PPD from the right side teeth sites of the new data and the soft lables generated by the correspondent Mother model to predict the contralateral PPD values.

In this study, the Mother Models were developed to predict periodontal probing depth (PPD) values for specific sites (DV, V, MV, DL, L, ML) using data from the NHANES 2011/2012 dataset. The Directional Symmetry Measure (γSM) was computed from the original data and incorporated as a feature.

In this version of the MoDau imputation approach, we use the XGBoost modeling technique (other different modeling techniques, such as NN or any other non-parametric technique, could also be used) to fit the Mother and Daughter models. We start by approaching the XGBoost modeling.

MoDau imputation methodology algorithm is the following:

Algorithm 2 Algorithm of MoDau Imputation Method Using XGBoost Models

Require: NHANES dataset $\mathbf{X}$, directional symmetry measures γSM
Ensure: Predicted PPD values for upper left teeth

Initial Model - Mother Model - Training
for each site s of every upper right tooth t **do**
    Fit XGBoost regression model $M_{t,s}$ using $\mathbf{X}$ and γSM
    Generate soft labels $L_{t,s}$ from model $M_{t,s}$
end for
Store the soft labels $L_{t,s}$ for each site s and tooth t to use with New Data Sets (next steps)
*__Training New Mother Model with a random subset of New Data__
for each site s of every upper right tooth t **do**
    Fit new XGBoost regression model $M_{t,s}^{\text{new}}$ using $\mathbf{X}_{\text{syn}}$ and γSM
    Generate predictions from model $M_{t,s}^{\text{new}}$
end for
Increase the number of New Mother Model predictions by randomly sampling with replacement to match the total number of observations in the new dataset.
Daughter Model Training
for each site s of every upper right tooth t **do**
    Train XGBoost regression model $D_{t,s}$ using soft labels $L_{t,s}$ and data from upper right teeth in $\mathbf{X}_{\text{syn}}$ along with expanded New Mother Model predictions (if available)
end for
Model Validation and Performance
for each site s of every upper right tooth t **do**
    Validate model $D_{t,s}$ by comparing predicted values against known benchmarks
    Assess model's generalizability to new, unseen data
end for

* This step is not essential in the algorithm, ensures that in cases where the new data is significantly different from the NHANES data, the model remains robust. The predictions from this new mother model were also used as predictors in the daughter models.

3.6.1 Mother Models

The data preparation process involved selecting relevant features such as the Pocket Probing Depth (PPD) for a given site on a specific tooth and the corresponding γSM for the same site. These features were utilized to create a training matrix using the `xgb.DMatrix` function, which facilitates efficient data handling for XGBoost.

The mother models were trained using PPD data from the upper right side of the teeth in the NHANES 2011/2012 dataset, along with the respective γSM values. Their

soft labels, which are predicted values for the left side, were then used as predictors in the daughter models.

Common Characteristics Across Mother Models

All six models in any tooth share several common characteristics. They are all XGBoost regression models (xgb.Booster) designed for predicting values using the squared error loss function (reg). This objective function minimizes the squared differences between predicted and actual values, focusing on reducing the Mean Square Error (MSE) during training.

A comprehensive grid search, using the `expand.grid` function in R, was conducted across a range of hyperparameters using the `caret` package in R. The tuning grid included the following sequences for the selected parameters:

The learning rate parameter (`eta`) ranged from 0.01 to 0.3, with increments of 0.05, to scale the contribution of each tree.

The maximum tree depth parameter (`max_depth`) ranged from 1 to 6, with increments of 0.5, to control complexity.

The gamma parameter (`gamma`) ranged from 0 to 0.25, with increments of 0.05, to control the number of possible splits.

The parameter (`colsample_bytree`) ranged from 0.6 to 1, with increments of 0.2, to control the the number of features to be used during trainning.

The minimum child weight parameter (`min_child_weight`) ranged from 1 to 6, with increments of 1, to select the number of meaningful splits.

The subsample ratio parameter (`subsample`) ranged from 0.2 to 1, with increments of 0.1, to introduce randomness and reduce overfitting.

The `nrounds` parameter was set to 1000, specifies the optimal number of boosting rounds, determined from the cross-validation step with early stopping after 10 rounds of no improvement.

Cross-validation was performed using 5-fold cross-validation on the training data with up to 1000 boosting rounds. Early stopping was employed if the performance did not improve for 10 consecutive rounds, resulting in the determination of the optimal number of boosting rounds.

The parameter `watchlist = list(train = train_matrix)` allows monitoring of the model's performance on the training data during training. It provides feedback on how the model is performing and helps to ensure that it is learning correctly.

These hyperparameter ranges were chosen based on standard practices and preliminary experiments to cover a wide range of possible values, ensuring a thorough search for optimal combinations.

Once the parameter values were identified through the grid search, the `xgb.train` function was used to train the model on the full training set. Predictions were then generated for the original dataset, and performance metrics, including RMSE, MAE, and R-squared, were calculated to assess the model's accuracy and reliability.

The training process for each model involved two features: γSM and the original site-specific feature. The `cb.evaluation.log()` callback function was used to log the models' performance metrics at each iteration, helping monitor the training process and allowing for an understanding of how the models' performance evolved over time.

Distinctive Characteristics Across Mother Models

The size of each model in kilobytes (Kb) reflects the memory consumption of the trained model, which includes the structure of the trees, the parameters, and the booster configuration. The number of iterations (`niter`) denotes the total number of iterations the model underwent during the training phase. Each iteration corresponds to an update of the model's parameters based on a subset of the training data, allowing the model to progressively refine its predictions.

Initial and final training RMSE metrics indicate the Root Mean Squared Error (RMSE) at the beginning and end of the training process, respectively. The initial RMSE shows the error when the model starts learning, while the final RMSE shows the accuracy after all iterations.

3.6.2 Validation of Mother-Daughter Imputation

To validate the MoDau method, we initiated the process by creating a simulated dataset containing 5000 observations. The MoDau method was subsequently applied to impute the missing values on the left side. Performance metrics were then computed by comparing the predictions of the Daughter model with the actual values of the left

side of the simulated dataset. These metrics provide a quantitative assessment of the MoDau method's accuracy and reliability. The detailed methodology employed in the subsequent steps is thoroughly described in the following sections.

Simulated Data Set

To rigorously assess the performance of the MoDau method, we generated a synthetic dataset comprising 5000 observations. This synthetic dataset was designed to mimic real-world data characteristics, ensuring that the evaluation of the MoDau method would be both relevant and robust. By using this controlled environment, we can systematically analyze the effectiveness of the MoDau method in imputing missing values and ensuring the integrity of the data.

The synthetic data was created using the `add_noise` function, designed to introduce controlled random noise into a dataset. This function generates normally distributed random noise using the `rnorm` function in R. It takes two parameters: `data`, representing the original dataset, and `noise_level`, specifying the standard deviation of the noise as a proportion of the original data's standard deviation.

The `rnorm` function generates random noise from a normal distribution with a mean of 0 and a standard deviation equal to the product of `noise_level` and the standard deviation of the original data (`sd(data)`). This noise is added to the original data on an element-by-element basis, resulting in a new dataset with a similar distribution but with the added variability introduced by the noise.

The noise is scaled by the standard deviation of the input data multiplied by the specified noise level, ensuring that the magnitude of the noise is proportional to the variability of the original data. The function then returns the original data with the added noise.

To create the new dataset, the original data is sampled with replacement increase the number of observations from 3320 to 5000, randomly selecting indices from the original dataset. This step ensures that the new dataset retains the characteristics of the original dataset but with the added variability from the resampling. Subsequently, the values for the new simulated variables are generated by adding noise to the sampled data using the `add_noise` function. This approach ensures that the simulated dataset mimics the structure of the original data while introducing controlled variability, making it suitable for validating the performance of the MoDau method.

We generate 10 datasets with noise levels ranging from 0.1 to 1 to select a synthetic dataset with an adequate noise level that retains the main features of an ex-Gaussian

distribution while being different enough to test the MoDau imputation method. The rationale for deciding on the best noise level was to obtain a plausible distribution with moderate variation and controlled uncertainty, and to ensure a good level of empirical evidence, avoiding both overfitting and underfitting. This topic will be further discussed in the discussion chapter.

KDE Bootstrap Difference The distributions across all noise levels and the original distribution KDEs were compared. The computed differences between each noise level and the original distribution, along with their respective confidence intervals, were calculated using bootstrap methods.

3.6.3 Assessing Daughter Models Imputation Results

The chi-squared test was employed to compare the proportions of periodontal probing depth (PPD) values between the original simulated data and the predicted values generated by the Daughter models. This analysis was conducted for six sites of the upper left central incisor: DV, V, MV, DL, L, and ML. The chi-squared test statistic and corresponding p-values were calculated for each PPD value category, as well as for the aggregated PPD values greater than or equal to 4 mm. The purpose was to assess the accuracy and reliability of the daughter models' predictions in replicating the original data distribution.

1. **Initial Model - Mother Model - Training**: We began by fitting individual XGBoost regression models to the NHANES data for each site of every upper right tooth. This involved six models per tooth, ensuring a detailed and site-specific approach. The models incorporated both the PPD values and the γSM, which quantifies directional symmetry.

2. **Generation of Soft Labels**: The output from these initial models were soft labels, which are probabilistic predictions rather than discrete labels. These soft labels provide a more nuanced prediction, reflecting the uncertainty and variability inherent in the data.

3. **Creation of a Simulated Dataset**: To assess the robustness and generalizability of our method, we generated a simulated dataset. This was accomplished by introducing a controlled amount of noise to the original NHANES data, sim-

ulating variations that could be expected in real-world scenarios. This step helps in evaluating the method's performance under different conditions.

4. **Training the New Mother Model**: An identical new mother model was then fit to the simulated dataset. In this process, we used five observations from the 5000 simulated observations, though this number can be adjusted as needed. The respective γSM values were computed on the simulated dataset. This step ensures that the model can adapt to variations in the data and maintain its predictive accuracy.

5. **Daughter Model Training**: Utilizing the soft labels as predictors, we trained a second generation of XGBoost regression models, referred to as daughter models. These models included data from the upper right teeth of the simulated dataset, along with five new soft labels generated from the new mother model fitted to the simulated data. The γSM values of the simulated data were also incorporated, ensuring that the models captured the directional symmetry.

6. **Model Validation and Performance**: The performance of the daughter models was validated to ensure their accuracy and reliability in predicting PPD values for the upper left teeth. This validation involved comparing the probabilistic distributions of empirical versus predicted values using optimal kernel density estimates (KDE) and the Kolmogorov-Smirnov (KS) test. Bootstrap methods were employed to assess the method's ability to generalize to new, unseen data, ensuring robustness and reliability in diverse scenarios.

While step 4 is not essential to the algorithm, it ensures that the model remains robust in cases where the new data significantly differs from the NHANES data. The predictions from this new mother model were also used as predictors in the daughter models, enhancing the model's adaptability and accuracy.

The the Figure 3.3 shows the schema of the MoDau imputation method, where the dashed line is the

Figure 3.3

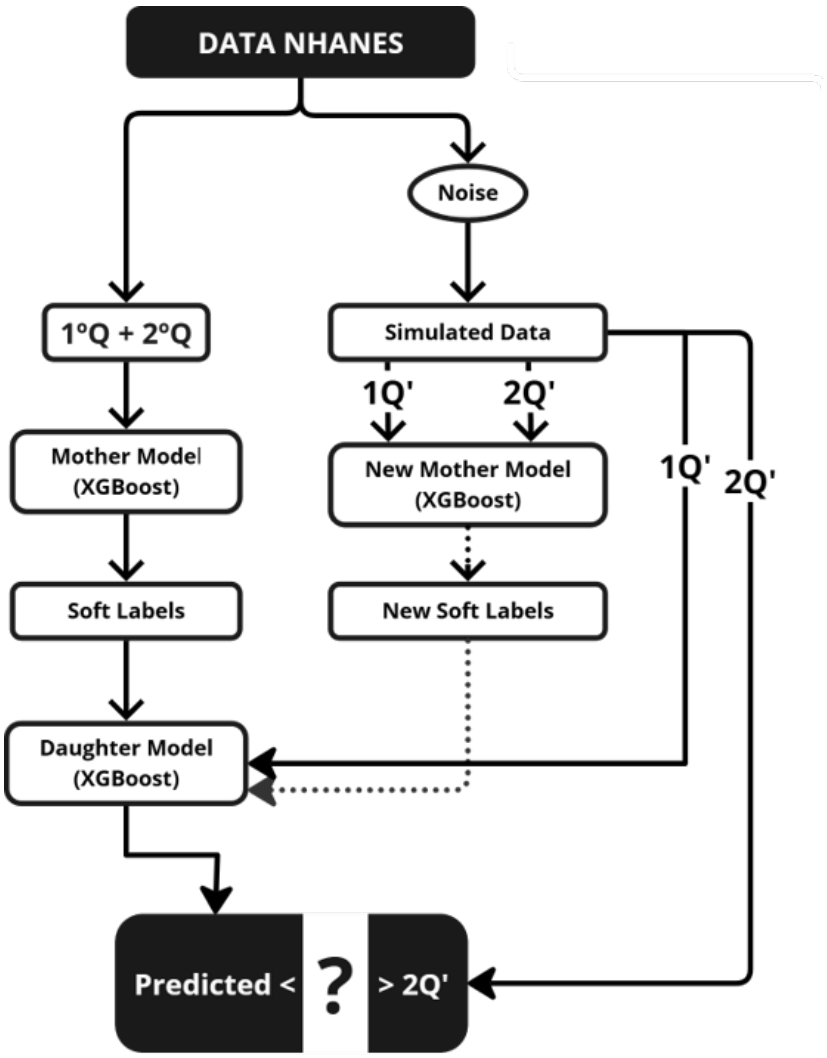


Diagram of MoDau

System Specifications

The main characteristics of our computer are: processor Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz, with 16 cores and a RAM of 64.0 GB (63.8 GB), which runs an operating system Windows 11 Pro of 64 bits, version 23H2, installed on 02/10/2022.

Chapter 4

Results

4.1 Symmetry Assessment

In this section, we present the results of the symmetry assessment for upper central incisors by site, accompanied by the respective tables. These tables are also repeated in Appendix I. For all other pairs of teeth, the corresponding tables can be found in Appendix I.

4.1.1 Upper Central Incisors (11,21)

Basic Methods

The results of the statistical tests comparing the probing pocket depth PPD medians and variances across six dental sites between the upper right central incisor (tooth 11) and the upper left central incisor (tooth 21) are summarized in Table 4.1. These tests aim to verify the symmetry of the contralateral sites of teeth 11 and 21.

For the DV sites, the median PPD values were identical for both teeth, with a non-significant Wilcoxon rank-sum test result (W = 3589510, p = 0.638), indicating no significant difference in medians. The variance comparison also showed no significant difference (F = 0.266, p = 0.606). The Kolmogorov-Smirnov (K-S) test confirmed no significant difference in the distribution shapes (D = 0.007, p = 1), suggesting symmetry at the DV sites.

In contrast, the V sites exhibited a significant difference in median PPD values (W = 3474941, p = 3.02e-03), though the variances remained similar (F = 0.065, p = 0.799). The K-S test indicated a minor distributional difference (D = 0.028, p =

0.247), suggesting a discrepancy in the central tendency but not in the variance.

The MV sites revealed highly significant differences in both median (W = 3354774, p = 5.33e-09) and variance (F = 18.218, p = 2.00e-05), supported by the K-S test (D = 0.079, p = 9.96e-08). These findings indicate substantial asymmetry in both the central tendency and variability of PPD between the contralateral MV sites.

At the DL sites, significant differences were observed in both the median (W = 3258241, p = 2.35e-04) and variance (F = 11.600, p = 6.65e-04), with the K-S test confirming a difference in distribution shapes (D = 0.050, p = 3.18e-03). These results suggest asymmetry in both the central tendency and variance at the DL sites.

For the L sites, the median and variance comparisons showed no significant differences (W = 3591491, p = 0.267; F = 0.230, p = 0.647). The K-S test corroborated these findings (D = 0.013, p = 0.9772), indicating symmetry at the L sites.

Finally, the ML sites demonstrated significant differences in both the median (W = 3405239, p = 5.38e-05) and variance (F = 7.425, p = 6.45e-03), as supported by the K-S test (D = 0.055, p = 5.21e-04). These results indicate substantial asymmetry at the ML sites.

Overall, these findings highlight notable asymmetries at specific sites (V, MV, DL, and ML) in both the central tendency and variability of PPD between teeth 11 and 21, as detailed in Table 4.1.

The assessment of probing pocket depth (PPD) symmetry between contralateral dental sites involves evaluating various statistical measures, including the Bhattacharyya coefficient and the correlation coefficient. These measures provide insight into the similarity of PPD distributions between corresponding sites on opposing teeth.

In Table 4.1, the Bhattacharyya coefficients for sites DV, V, MV, DL, L, and ML range from 0.997 to 1.000, indicating high similarity between the contralateral sites. The correlation coefficients, which measure the linear relationship between the PPD values at contralateral sites, range from 0.58 to 0.73, further supporting the similarity between the sites.

Table 4.1: Summary of statistical tests comparing central incisors 11 and 21 PPD medians and variances across six dental sites; distances between distributions.

Site	Stats	Teeth		Test Results		K-S	Bhat.	Corr.
		11	21	Stats	p		Coef.	Coef.
DV	Median	1	1	W = 3589510	0.638	D = 0.007	1.000	0.58
	Variance	0.685	0.761	F = 0.266	0.606	p = 1		
V	Median	1	1	W = 3474941	**3.02e-03**	D = 0.028	0.999	0.64
	Variance	0.538	0.578	F = 0.065	0.799	p = 0.247		
MV	Median	1	1	W = 3354774	**5.33e-09**	D = 0.079	0.997	0.66
	Variance	0.639	0.702	F = 18.218	**2.00e-05**	p = **9.96e-08**		
DL	Median	1	1	W = 3258241	**2.35e-04**	D = 0.050	0.998	0.62
	Variance	0.729	0.807	F = 11.600	**6.65e-04**	p = **3.18e-03**		
L	Median	1	1	W = 3591491	0.267	D = 0.013	0.999	0.65
	Variance	0.731	0.726	F = 0.230	0.647	p = 0.9772		
ML	Median	2	2	W = 3405239	**5.38e-05**	D = 0.055	0.998	0.73
	Variance	0.826	0.779	F = 7.425	**6.45e-03**	p = **5.21e-04**		

Abbreviations: 11 – Upper right central incisor, 21 – Upper left central incisor; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

Advanced Methods

Distal Vestibular Sites

The analysis was conducted using GAMLSS to assess the effect of the "Side" variable on PPD at distal vestibular sites of the upper central incisors (teeth 11 and 21).

Two models were compared: a null model (PPD$\sim$ 1) and a model including the "Side" predictor (PPD$\sim$ Side). The results are summarized in Table 4.2. The parameter estimates for the location (μ), scale (σ), and skewness (ν) in both models are nearly identical. The inclusion of the "Side" predictor in the model showed no significant impact on the PPD distribution at these sites. This was evidenced by the LRT result, which yielded a χ^2 value of 0.975 with a p-value of 0.807, indicating no significant difference between the two models. Cross-validation metrics, as shown in Table 4.3, further support this finding, with no substantial differences in RMSE and MAE between the models.

Table 4.2: Assessing Side effect on PPD at Upper Central Incisors Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.682	**<2e-16**	0.681	**<2e-16**	0.975	0.807
(identity)	21DV	•	•	0.002	0.91		
σ	Intercept	-1.112	**<2e-16**	-1.105	**<2e-16**		
(log)	21DV	•	•	-0.015	0.698		
ν	Intercept	-0.324	**<2e-16**	-0.337	**<2e-16**		
(log)	21DV	•	•	0.025	0.483		
	n	5172		5172			
	D. F.	3		6			
	Res. D. F.	5169		5166			
	G. D.	10948.65		10947.68			
	AIC	10954.65		10959.68			
	SBC	10974.3		10998.98			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table 4.3: Statistical Comparison of Cross Validation Metrics for Models: PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.115	1.115
Paired t-test for RMSE		
t-value		2.164
Degrees of Freedom		4
p-value		0.096
95% CI		[-2.2e-05, 1.78e-04]
Mean Difference		7.8e-05
Mean MAE	0.793	0.793
Paired t-test for MAE		
t-value		2.839
Degrees of Freedom		4
p-value		0.047
95% CI		**[3e-06, 2.47e-04]**
Mean Difference		1.25e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

Vestibular Sites

A similar approach was applied to vestibular sites. The comparison of the null model and the model including the "Side" predictor (PPD∼ Side) is detailed in Table 4.4. In contrast to the distal vestibular sites, the inclusion of the "Side" predictor significantly affected the PPD distribution at vestibular sites, as indicated by the LRT result (χ^2 = 8.947, p = 0.03).

Table 4.4: Assessing Side effect on PPD at Upper Central Incisors Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD ∼ Side Estimate	PPD ∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.363	< **2e-16**	0.347	< **2e-16**	8.947	**0.03**
(identity)	21V	•	•	0.033	0.273		
σ	Intercept	-0.693	< **2e-16**	-0.702	< **2e-16**		
(log)	21V	•	•	0.014	0.715		
ν	Intercept	-0.630	< **2e-16**	-0.658	< **2e-16**		
(log)	21V	•	•	0.052	0.365		
	n	5170		5170			
	D. F.	3		6			
	Res. D. F.	5167		5164			
	G. D.	11017.96		11009.01			
	AIC	11023.96		11021.01			
	SBC	11043.61		11060.32			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Cross-validation results presented in Table 4.5 demonstrate slight improvements in model performance metrics when the "Side" predictor is included. Both RMSE and MAE showed minor reductions, suggesting enhanced predictive accuracy and model fit.

Mesial Vestibular Sites

For mesial vestibular sites, the GAMLSS results in Table 4.5 indicate a significant effect of the "Side" predictor on PPD distribution. The LRT result yielded a χ^2 value of 25.989 with a highly significant p-value (p = 9.59e-06). Parameter estimates for the location, scale, and skewness parameters changed considerably when the "Side"

Table 4.5: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	0.921	0.919
Paired t-test for RMSE		
t-value		4.637
Degrees of Freedom		4
p-value		**9.76e-03**
95% CI		[**4.8e-04**, **1.92e-03**]
Mean Difference		1.2e-03
Mean MAE	0.743	0.742
Paired t-test for MAE		
t-value		4.045
Degrees of Freedom		4
p-value		**0.016**
95% CI		[**3.3e-04**, **1.78e-03**]
Mean Difference		1.06e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

predictor was included, suggesting a notable impact.

Table 4.6: Assessing Side effect on PPD at Upper Central Incisors Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1		PPD∼ Side		LRT	
		Estimate	Pr(> |t|)	Estimate	Pr(> |t|)	χ^2	Pr(> χ^2)
μ	Intercept	0.671	< **2e-16**	0.663	< **2e-16**	25.989	**9.59e-06**
(identity)	21MV	•	•	0.021	0.27		
σ	Intercept	-1.088	< **2e-16**	-1.105	< **2e-16**		
(log)	21MV	•	•	0.039	0.295		
ν	Intercept	-0.378	< **2e-16**	-0.446	< **2e-16**		
(log)	21MV	•	•	0.125	**5.93e-05**		
	n	5168		5168			
	D. F.	3		6			
	Res. D. F.	5165		5162			
	G. D.	10667.79		10641.8			
	AIC	10673.79		10653.8			
	SBC	10693.44		10693.1			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

The cross-validation metrics in Table 4.6 reveal significant improvements in RMSE and MAE, further supporting the relevance of the "Side" variable in predicting PPD at these sites. The paired t-tests for RMSE and MAE also showed significant differences, indicating that the inclusion of the "Side" predictor enhances the model's predictive capabilities.

Table 4.7: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD $\sim$ 1 and PPD $\sim$ Side

	PPD $\sim$ 1	PPD $\sim$ Side
Mean RMSE	1.070	1.067
Paired t-test for RMSE		
t-value		6.538
Degrees of Freedom		4
p-value		2.8e-003
95% CI		[**1.2e-03**, **3.0e-03**]
Mean Difference		2.1e-03
Mean MAE	0.763	0.761
Paired t-test for MAE		
t-value		7.388
Degrees of Freedom		4
p-value		1.8e-03
95% CI		[**1.5e-03**, **3.2e-03**]
Mean Difference		2.3e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

Distal Lingual Sites

At distal lingual sites, the results shown in Table 4.8 indicate a significant effect of the "Side" predictor on PPD distribution. The LRT result was significant (χ^2 = 12.879, p = 4.91e-03), suggesting that the inclusion of the "Side" variable significantly improves the model.

Table 4.8: Assessing Side effect on PPD at Upper Central Incisors Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ (identity)	Intercept	0.703	< **2e-16**	0.696	< **2e-16**	12.879	**4.91e-03**
	21DL	•	•	0.018	0.422		
σ (log)	Intercept	-1.048	< **2e-16**	-1.067	< **2e-16**		
	21DL	•	•	0.043	0.36		
ν (log)	Intercept	-0.216	< **2e-16**	-0.261	< **2e-16**		
	21DL	•	•	0.083	**0.024**		
	n	5238		5238			
	D. F.	3		6			
	Res. D. F.	5235		5232			
	G. D.	12025.03		12012.15			
	AIC	12031.03		12024.15			
	SBC	12050.72		12063.53			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Cross-validation metrics in Table 4.9 also reflect improvements in model performance. Both RMSE and MAE decreased with the inclusion of the "Side" predictor, and the paired t-tests confirmed significant differences, reinforcing the importance of the "Side" variable in these models.

Table 4.9: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD $\sim$ 1 and PPD $\sim$ Side

	PPD $\sim$ 1	PPD $\sim$ Side
Mean RMSE	1.188	1.187
Paired t-test for RMSE		
t-value		12.242
Degrees of Freedom		4
p-value		**2.6e-04**
95% CI		[**1.2e-03**, **1.9e-03**]
Mean Difference		1.5e-03
Mean MAE	0.871	0.870
Paired t-test for MAE		
t-value		11.323
Degrees of Freedom		4
p-value		**3.5e-04**
95% CI		[**1.3e-03**, **2.1e-03**]
Mean Difference		1.7e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

Lingual Sites

The GAMLSS results for lingual sites, presented in Table 4.10, show no significant effect of the "Side" predictor on PPD distribution.

The LRT result ($\chi^2 = 1.381$, p = 0.710) indicates no significant difference between the models with and without the "Side" predictor. However, cross-validation metrics (4.11) show slight improvements in RMSE and MAE with the inclusion of the "Side" predictor. Although the paired t-tests yielded significant p-values for these metrics, the differences are relatively minor, suggesting a marginal impact of the "Side" variable on model performance at these sites.

Table 4.10: Assessing Side effect on PPD at Upper Central Incisors Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.479	< **2e-16**	0.469	< **2e-16**	1.381	0.710
(identity)	21L	•	•	0.019	0.471		
σ	Intercept	-0.752	< **2e-16**	-0.758	< **2e-16**		
(log)	21L	•	•	0.013	0.729		
ν	Intercept	-0.416	< **2e-16**	-0.420	< **2e-16**		
(log)	21L	•	•	0.008	0.866		
	n	5168		5168			
	D. F.	3		6			
	Res. D. F.	5165		5162			
	G. D.	11798.54		11797.16			
	AIC	11804.54		11809.16			
	SBC	11824.19		11848.46			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table 4.11: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.077	1.077
Paired t-test for RMSE		
t-value		2.776
Degrees of Freedom		4
p-value		**0.048**
95% CI		[**7.5e-07**, **7.9e-04**]
Mean Difference		3.94e-04
Mean MAE	0.828	0.828
Paired t-test for MAE		
t-value		3.782
Degrees of Freedom		4
p-value		**0.019**
95% CI		[**9.52e-05**, **6.21e-04**]
Mean Difference		3.58e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

Mesial Lingual Sites

Table 4.12: Assessing Side Effect on PPD at Upper Central Incisors ML Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and LRT with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD ∼ Side Estimate	PPD ∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.805	< **2e-16**	0.787	< **2e-16**		
(identity)	21ML	•	•	0.228	**2.72e-10**		
σ	Intercept	-1.365	< **2e-16**	-1.290	< **2e-16**	18.96	**2.79e-04**
(log)	21ML	•	•	0.454	**4.85e-10**		
ν	Intercept	-0.037	0.025	-0.060	**0.011**		
(log)	21ML	•	•	-0.172	**7.86e-04**		
	n	5175		5175			
	D. F.	3		6			
	Res. D. F.	5172		5169			
	G. D.	12381.25		12400.21			
	AIC	12387.25		12412.21			
	SBC	12406.91		12451.52			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

For L site, "Side" predictor on PPD distribution, the LRT result (χ^2 = 18.96, p = 2.79e-04) confirms the significance of the "Side" variable in the model.

Cross-validation results (Table 4.13) show marked improvements in both RMSE and MAE when the "Side" predictor is included. The paired t-tests for these metrics indicate significant differences, highlighting the enhanced predictive accuracy and model fit with the inclusion of the "Side" variable.

Table 4.13: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.317	1.260
Paired t-test for RMSE		
t-value		4.125
Degrees of Freedom		4
p-value		**0.015**
95% CI		[**0.019**, **0.095**]
Mean Difference		0.057
Mean MAE	0.986	0.914
Paired t-test for MAE		
t-value		4.173
Degrees of Freedom		4
p-value		**0.014**
95% CI		[**0.024**, **0.120**]
Mean Difference		0.072

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

In summary the effect of the "Side" predictor on PPD distribution varies across different tooth sites of central incisors. While significant effects were observed at vestibular, mesial vestibular, distal lingual, and mesial lingual sites, no significant impact was found at distal vestibular and lingual sites. Cross-validation metrics generally supported these findings, demonstrating improved model performance with the inclusion of the "Side" predictor at the sites where significant effects were observed.

4.1.2 Upper Lateral Incisors (12,22)

Basic Methods

The results of the statistical tests comparing the PPD medians and variances across different dental sites between the right side (tooth 12) and the left side (tooth 22) are summarized in Table I.14. These tests aim to verify the symmetry of the contralateral sites of teeth 12 and 22.

At the distal vestibular (DV) sites, a significant difference was found in the median PPD values (W = 2979273, p = 8.61e-03). The variance comparison, however, did not show a significant difference (F = 0.489, p = 0.485), though the Kolmogorov-Smirnov (K-S) test indicated a significant difference in the distribution shapes (D = 0.040, p = 0.036). This suggests a difference in central tendency but not in variability.

For the vestibular (V) sites, both the median (W = 2922411, p = 1.45e-04) and variance (F = 10.961, p = 9.37e-04) comparisons revealed significant differences, supported by the K-S test (D = 0.059, p = 3.38e-04). These findings indicate asymmetry in both the central tendency and variability of PPD between the contralateral V sites.

In contrast, the mesial vestibular (MV) sites exhibited no significant differences in both median (W = 3088533, p = 0.883) and variance (F = 1.856, p = 0.173), with the K-S test also confirming no significant distributional difference (D = 0.010, p = 1). These results suggest symmetry at the MV sites.

The distal lingual (DL) sites showed no significant differences in median (W = 3103952, p = 0.847) and variance (F = 0.010, p = 0.921), as confirmed by the K-S test (D = 8.44e-03, p = 1), indicating symmetry at the DL sites.

For the lingual (L) sites, the median and variance comparisons also showed no significant differences (W = 3051359, p = 0.329; F = 0.379, p = 0.538), with the K-S test corroborating these findings (D = 0.010, p = 1). This suggests symmetry at the L sites.

Finally, the mesial lingual (ML) sites demonstrated no significant differences in both median (W = 3065477, p = 0.515) and variance (F = 0.024, p = 0.877), supported by the K-S test (D = 0.008, p = 1). These results indicate symmetry at the ML sites.

Overall, these findings highlight significant asymmetries at the DV and V sites in both central tendency and variability of PPD between teeth 12 and 22, as detailed in Table I.14. However, the MV, DL, L, and ML sites demonstrated symmetry, suggesting consistent PPD distributions across these contralateral sites.

Table I.14 reports the Bhattacharyya coefficients, which range from 0.998 to 1.000, demonstrating high similarity between the distributions of PPD at contralateral sites. The correlation coefficients range from 0.55 to 0.64, indicating a moderate to high linear relationship between the PPD values at the contralateral sites.

Advanced Methods

Distal Vestibular Sites

To evaluate the effect of the "site" predictor on the PPD at the distal vestibular sites of the upper lateral incisors (teeth 12 and 22), GAMLSS models were employed. Two models were compared: a null model (PPD$\sim$ 1) and a model including the "site" predictor (PPD$\sim$ Side). The results are detailed in Table I.15.

The parameter estimates for location (μ), scale (σ), and skewness (ν) showed min-

imal differences between the two models. Specifically, the inclusion of the "site" predictor did not significantly alter the PPD distribution at these sites. The Likelihood Ratio Test (LRT) resulted in a χ^2 value of 3.46 with a p-value of 0.326, indicating no significant improvement in the model fit by including the "site" variable.

The cross-validation metrics provided in Table I.16 further support these findings. The Mean Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were almost identical between the two models. The paired t-tests for both RMSE and MAE showed no significant differences, with p-values of 0.084 and 0.129, respectively. This suggests that the inclusion of the "site" predictor does not significantly enhance the predictive performance of the model for distal vestibular sites.

Vestibular Sites

For vestibular sites, the GAMLSS models were also employed to assess the impact of the "site" predictor on PPD. The results are summarized in Table I.17. In this case, the inclusion of the "site" predictor had a significant effect on the PPD distribution, as evidenced by the LRT result ($\chi^2 = 21.351$, p = 8.9e-05).

Cross-validation metrics in Table I.20 indicated slight improvements in the model performance with the inclusion of the "site" predictor. The Mean RMSE and MAE both showed minor reductions, and the paired t-tests confirmed significant differences, particularly for MAE (p = 0.013). These results suggest that the "site" variable is a relevant predictor for PPD at vestibular sites, enhancing the model's predictive accuracy.

Mesial Vestibular Sites

At mesial vestibular sites, the effect of the "site" predictor on PPD was also assessed using GAMLSS models. The findings are presented in Table I.19. The inclusion of the "site" predictor showed a borderline significant effect on the PPD distribution, with an LRT result of $\chi^2 = 7.70$ and a p-value of 0.052.

Cross-validation metrics, as shown in Table I.20, did not indicate substantial improvements in model performance. Both RMSE and MAE metrics remained largely unchanged, and the paired t-tests for these metrics did not show significant differences, with p-values of 0.617 and 0.607, respectively. These results imply that the inclusion of the "site" predictor does not markedly improve the model's predictive performance at mesial vestibular sites.

Distal Lingual Sites

For distal lingual sites, the analysis using GAMLSS models is summarized in Table I.21. The inclusion of the "site" predictor did not significantly impact the PPD distribution, as indicated by the LRT result ($\chi^2 = 0.299$, p = 0.960).

The cross-validation metrics presented in Table I.22 further support this conclusion. Both RMSE and MAE metrics showed no significant differences between the models, with p-values of 0.193 and 0.140, respectively. This suggests that the "site" predictor does not contribute significantly to the model's predictive capabilities at distal lingual sites.

Lingual Sites

The GAMLSS analysis for lingual sites is detailed in Table I.23. Similar to distal vestibular sites, the inclusion of the "site" predictor did not significantly affect the PPD distribution, as indicated by the LRT result ($\chi^2 = 1.583$, p = 0.663).

Cross-validation metrics in Table I.24 show that the RMSE and MAE values remained virtually unchanged with the inclusion of the "site" predictor. The paired t-tests for these metrics yielded p-values of 0.559 and 0.856, respectively, suggesting that the "site" variable does not enhance the model's predictive performance at lingual sites.

Mesial Lingual Sites

The effect of the "site" predictor on PPD at mesial lingual sites was also evaluated using GAMLSS models. The results are summarized in Table I.25. The inclusion of the "site" predictor did not significantly influence the PPD distribution, as indicated by the LRT result ($\chi^2 = 0.316$, p = 0.957).

Cross-validation metrics in Table I.26 further confirm this finding, with no significant differences in RMSE and MAE between the models. The paired t-tests for these metrics yielded p-values of 0.108 and 0.085, respectively. These results indicate that the "site" predictor does not significantly improve the model's predictive performance at mesial lingual sites.

In summary, the "site" predictor had varying levels of impact on the PPD distribution across 12, 22 tooth sites. Significant effects were observed at vestibular sites, suggesting the importance of the "site" variable in predicting PPD in these areas. However, no significant effects were found at distal vestibular, distal lingual, lingual,

and mesial lingual sites, indicating that the inclusion of the "site" predictor does not enhance model performance at these locations. Cross-validation metrics generally support these findings, highlighting the variability in the relevance of the "site" predictor across different tooth sites.

4.1.3 Upper Canines (13,23)

Basic Methods

The results of the statistical tests comparing the PPD means, medians, and variances across different dental sites between the right side (tooth 13) and the left side (tooth 23) are summarized in Table I.27. These tests aim to verify the symmetry of the contralateral sites of teeth 13 and 23.

At the distal vestibular (DV) sites, significant differences were found in both the median PPD values (W = 2979273, p = 8.61e-03) and the distribution shapes as indicated by the Kolmogorov-Smirnov (K-S) test (D = 0.040, p = 0.036). However, the variance comparison did not show a significant difference (F = 0.489, p = 0.485). This suggests asymmetry in the central tendency but not in the variability of PPD between the contralateral DV sites.

For the vestibular (V) sites, significant differences were observed in the median PPD values (W = 2922411, p = 1.45e-04) and variance (F = 10.961, p = 9.37e-04), supported by the K-S test (D = 0.059, p = 3.38e-04). These findings indicate asymmetry in both central tendency and variability of PPD between the contralateral V sites.

In contrast, the mesial vestibular (MV) sites exhibited no significant differences in both median (W = 3088533, p = 0.883) and variance (F = 1.856, p = 0.173), with the K-S test also confirming no significant distributional difference (D = 0.010, p = 1). These results suggest symmetry at the MV sites.

The distal lingual (DL) sites showed no significant differences in both median (W = 3103952, p = 0.847) and variance (F = 0.010, p = 0.921), as confirmed by the K-S test (D = 0.008, p = 1). This suggests symmetry at the DL sites.

For the lingual (L) sites, the median and variance comparisons also showed no significant differences (W = 3051359, p = 0.329; F = 0.379, p = 0.538), with the K-S test corroborating these findings (D = 0.010, p = 1). These results indicate symmetry at the L sites.

Finally, the mesial lingual (ML) sites demonstrated no significant differences in both

median (W = 3065477, p = 0.515) and variance (F = 0.024, p = 0.877), supported by the K-S test (D = 0.008, p = 1). These results suggest symmetry at the ML sites.

Overall, the findings highlight significant asymmetries at the DV and V sites in both the central tendency and variability of PPD between teeth 13 and 23, as detailed in Table I.27. However, the MV, DL, L, and ML sites demonstrated symmetry, suggesting consistent PPD distributions across these contralateral sites.

In Table I.27, the Bhattacharyya coefficients range from 0.998 to 1.000, suggesting high similarity between the PPD distributions at contralateral sites. The correlation coefficients vary from 0.45 to 0.56, indicating a moderate linear relationship between the PPD values at these sites.

Advanced Methods

Distal Vestibular Sites

The assessment of the "Side" effect on PPD at upper canines distal vestibular sites was conducted using GAMLSS models. The inclusion of the "Side" predictor showed a significant impact on the PPD distribution, as indicated by the likelihood ratio test (LRT) result (χ^2 = 47.78, p = 2.37e-10). The location parameter (μ) remained consistent with an intercept estimate of 0.776, while the scale (σ) and skewness (ν) parameters were also significantly affected (Table I.28). The cross-validation metrics, presented in Table I.29, indicate a reduction in both RMSE and MAE when including the "Side" predictor, with significant paired t-test results ($p < 0.001$), suggesting that the model's predictive accuracy improved with the inclusion of "Side".

Vestibular Sites

For vestibular sites, the "Side" predictor significantly influenced the PPD distribution, as shown by the LRT result (χ^2 = 35.66, p = 8.85e-08). The estimate for the location parameter (μ) changed from 0.314 to 0.261, indicating a notable shift. The scale (σ) and skewness (ν) parameters, however, did not show significant changes except for the location parameter (Table I.30). Cross-validation results in Table I.31 show a minor but statistically significant reduction in MAE (p = 2.89e-03), highlighting the functional improvement of the model with the "Side" predictor.

Mesial Vestibular Sites

The effect of the "Side" predictor on mesial vestibular sites was not significant, as indicated by the LRT result ($\chi^2 = 0.299$, p = 0.960) (Table I.32). The parameter estimates for location (μ), scale (σ), and skewness (ν) remained largely unchanged. Cross-validation metrics (Table I.33) support this finding, showing no significant differences in RMSE and MAE between the null model and the model with the "Side" predictor.

Distal Lingual Sites

At distal lingual sites (Table: I.34), the inclusion of the "Side" predictor significantly affected the PPD distribution (LRT $\chi^2 = 32.075$, p = 5.05e-07). The location parameter (μ) showed a slight change from 0.774 to 0.785, while the scale (σ) and skewness (ν) parameters also indicated significant variations. The cross-validation results in Table I.35 demonstrate significant improvements in both RMSE and MAE, with paired t-test results ($p < 0.01$), underscoring the enhanced predictive performance of the model when including the "Side" predictor.

Lingual Sites

For lingual sites (Table: I.36), the "Side" predictor did not significantly affect the PPD distribution, as shown by the LRT result ($\chi^2 = 1.588$, p = 0.662). The estimates for the location (μ), scale (σ), and skewness (ν) parameters remained consistent between the models. Cross-validation metrics (Table I.37) reveal no significant differences in RMSE and MAE, indicating that the inclusion of the "Side" predictor does not improve the model's predictive accuracy at these sites.

Mesial Lingual Sites

The inclusion of the "Side" predictor at mesial lingual sites significantly influenced the PPD distribution (LRT $\chi^2 = 24.169$, p = 2.3e-05) (Table: I.38). The location parameter (μ) slightly changed from 0.746 to 0.751. The scale (σ) and skewness (ν) parameters showed significant changes, highlighting the impact of the "Side" predictor. Cross-validation metrics in Table I.39 indicate significant reductions in RMSE and MAE, with paired t-test results ($p < 0.05$), suggesting improved predictive performance with the inclusion of the "Side" predictor.

4.1.4 Upper First Pre Molars (14, 24)

Basic Methods

Table I.40 shows that the Bhattacharyya coefficients for the various sites range from 0.991 to 1.000, indicating high distributional similarity. The correlation coefficients, ranging from 0.58 to 0.73, suggest a moderate to high linear relationship between the PPD values at the contralateral sites.

Advanced Methods

Distal Vestibular Sites

The analysis of the distal vestibular sites using GAMLSS models, as shown in Table I.41, indicates a significant effect of the "Side" predictor on the PPD distribution. The likelihood ratio test (LRT) revealed a χ^2 value of 79.809 with a p-value less than 2.2e-16, demonstrating the substantial impact of the "Side" variable. The estimates for the location parameter (μ) slightly increased from 0.748 to 0.756 with the inclusion of the "Side" predictor. Additionally, significant changes were observed in the scale (σ) and skewness (ν) parameters with p-values of 6.0e-06 and 1.2e-10, respectively. Cross-validation results presented in Table I.42 demonstrate a significant reduction in both RMSE and MAE when including the "Side" predictor, with paired t-test results showing p-values of 1.77e-05 and 4.50e-05, respectively. This indicates an improvement in the model's predictive accuracy with the inclusion of the "Side" variable.

Vestibular Sites

At the vestibular sites, the "Side" predictor also had a significant impact on the PPD distribution, as evidenced by the LRT results ($\chi^2 = 27.614$, p = 4.38e-06). The location parameter (μ) estimate increased from 0.440 to 0.460. The scale (σ) and skewness (ν) parameters showed significant changes, particularly in the skewness parameter with a p-value of 9.26e-05 (Table I.43). The cross-validation metrics, shown in Table I.44, indicate a significant reduction in MAE with a paired t-test p-value of 0.030, reflecting the functional improvement of the model with the "Side" predictor.

Mesial Vestibular Sites

For the mesial vestibular sites, the GAMLSS models indicated no significant effect of the "Side" predictor on the PPD distribution, as demonstrated by the LRT results (χ^2 = 1.479, p = 0.687). The estimates for the location parameter (μ) remained unchanged at 0.769, and no significant changes were observed in the scale (σ) and skewness (ν) parameters (Table I.45). The cross-validation metrics, shown in Table I.46, did not indicate significant improvements in RMSE and MAE, suggesting that the inclusion of the "Side" predictor does not enhance the model's functional properties for these sites.

Distal Lingual Sites

The distal lingual sites exhibited a significant effect of the "Side" predictor on the PPD distribution. The LRT results, as presented in Table I.47, showed a χ^2 value of 89.332 with a p-value less than 2.2e-16. The location parameter (μ) estimate increased from 0.837 to 1.111, and significant changes were observed in the scale (σ) and skewness (ν) parameters. Cross-validation results in Table I.48 demonstrate substantial reductions in RMSE and MAE, with paired t-test results showing p-values of 6.55e-07 and 1.10e-7, respectively, indicating significant improvement in the model's predictive accuracy with the inclusion of the "Side" predictor.

Lingual Sites

The inclusion of the "Side" predictor at lingual sites also showed a significant effect on the PPD distribution, with LRT results (χ^2 = 9.091, p = 0.028). The location parameter (μ) estimate changed slightly from 0.589 to 0.578. While the scale (σ) parameter did not show significant changes, the skewness (ν) parameter was affected (Table I.49). The cross-validation metrics in Table I.50 reveal a significant reduction in RMSE with a paired t-test p-value of 0.016 and in MAE with a p-value of 0.036, indicating the model's improved predictive performance when including the "Side" predictor.

Mesial Lingual Sites

For the mesial lingual sites, the "Side" predictor had a significant impact on the PPD distribution, as shown by the LRT results (χ^2 = 55.471, p = 5.45e-12). The location parameter (μ) estimate increased from 0.761 to 0.788, and the scale (σ) and skewness (ν) parameters also showed significant changes (Table I.51). The cross-validation results in

Table I.52 indicate significant reductions in both RMSE and MAE, with paired t-test results showing p-values of 0.006 and 0.011, respectively, demonstrating the enhanced predictive accuracy of the model with the inclusion of the "Side" predictor.

4.1.5 Upper Second Premolars (15, 25)

Basic Methods

The results of the statistical tests comparing the PPD means, medians, and variances across different dental sites between the upper second premolar (tooth 15) and the contralateral upper second premolar (tooth 25) are summarized in Table I.53. These tests aim to verify the symmetry of the contralateral sites of teeth 15 and 25.

At the distal vestibular (DV) sites, significant differences were found in both the median PPD values (W = 2066818, p < 2.2e-16) and the distribution shapes as indicated by the Kolmogorov-Smirnov (K-S) test (D = 0.165, p < 2.2e-16). The variance comparison also showed a significant difference (F = 114.920, p < 2.2e-16). These results indicate asymmetry in both the central tendency and the variability of PPD between the contralateral DV sites.

For the vestibular (V) sites, significant differences were observed in the median PPD values (W = 2293838, p = 4.87e-07) and variance (F = 0.946, p = 0.331), supported by the K-S test (D = 0.048, p = 1.29e-02). These findings suggest asymmetry in both the central tendency and the variability of PPD between the contralateral V sites.

In contrast, the mesial vestibular (MV) sites exhibited significant differences in both median (W = 2359115, p = 2.28e-03) and variance (F = 4.914, p = 0.027), with the K-S test also confirming significant distributional differences (D = 0.053, p = 3.45e-03). These results indicate asymmetry at the MV sites.

The distal lingual (DL) sites showed significant differences in both median (W = 2825090, p < 2.2e-16) and variance (F = 6.902, p = 8.64e-03), supported by the K-S test (D = 0.127, p = 4.44e-16). These findings suggest asymmetry in both the central tendency and the variability at the DL sites.

For the lingual (L) sites, no significant differences were observed in either median (W = 2445321, p = 0.391) or variance (F = 1.906, p = 0.167), with the K-S test also indicating no significant distributional differences (D = 0.026, p = 0.437). These results suggest symmetry at the L sites.

Finally, the mesial lingual (ML) sites demonstrated significant differences in both

median (W = 2883271, p < 2.2e-16) and variance (F = 3.579, p = 0.057), supported by the K-S test (D = 0.135, p < 2.2e-16). These results indicate asymmetry at the ML sites.

Overall, the findings highlight significant asymmetries at the DV, V, MV, DL, and ML sites in both the central tendency and variability of PPD between teeth 15 and 25, as detailed in Table I.53. However, the L site demonstrated symmetry, suggesting consistent PPD distributions across these contralateral sites.

In Table I.53, the Bhattacharyya coefficients range from 0.980 to 1.000, indicating high similarity in the PPD distributions at the contralateral sites. The correlation coefficients, ranging from 0.55 to 0.62, support a moderate to high linear relationship between the PPD values at these sites.

Advanced Methods

Distal Vestibular Sites

The GAMLSS models revealed a significant effect of the "Side" predictor on the PPD distribution at the distal vestibular sites of upper second premolars. The likelihood ratio test (LRT) yielded a χ^2 value of 147.22 with a p-value less than 2.2e-16, indicating a substantial impact. The location parameter (μ) estimate slightly increased from 0.752 to 0.764 with the inclusion of the "Side" predictor. Additionally, the scale (σ) and skewness (ν) parameters were significantly influenced by "Side," with p-values of 1.81e-08 and less than 2e-16, respectively (Table I.54). The cross-validation metrics, shown in Table I.55, demonstrated significant reductions in both RMSE and MAE, with paired t-test p-values of 2.83e-04 and 3.08e-04, respectively, indicating improved predictive accuracy with the inclusion of the "Side" variable.

Vestibular Sites

At the vestibular sites, the "Side" predictor significantly affected the PPD distribution, as indicated by the LRT results (χ^2 = 32.167, p = 4.84e-07). The location parameter (μ) estimate remained relatively stable, with a minor decrease from 0.469 to 0.464. The scale (σ) and skewness (ν) parameters also exhibited significant changes, particularly the skewness parameter with a p-value of 0.002 (Table I.56). The cross-validation results in Table I.57 revealed significant reductions in RMSE and MAE, with paired t-test p-values of 1.33e-04 and 9.47e-05, respectively, reflecting the enhanced predictive performance of the model when incorporating the "Side" predictor.

Mesial Vestibular Sites

The inclusion of the "Side" predictor significantly impacted the PPD distribution at the mesial vestibular sites. The LRT results indicated a significant effect with a χ^2 value of 23.899 and a p-value of 3.0e-05. The location parameter (μ) estimate decreased slightly from 0.785 to 0.772, while the scale (σ) and skewness (ν) parameters showed notable changes (Table I.58). Cross-validation metrics presented in Table I.59 revealed significant improvements in RMSE and MAE, with paired t-test p-values of 0.029 and 0.032, respectively, indicating the inclusion of the "Side" predictor enhances the model's functional properties.

Distal Lingual Sites

The distal lingual sites also exhibited a significant effect of the "Side" predictor on the PPD distribution. The LRT results showed a χ^2 value of 8.384 with a p-value of 0.039. The location parameter (μ) estimate increased from 0.821 to 1.123, while significant changes were observed in the scale (σ) and skewness (ν) parameters (Table I.60). The cross-validation results in Table I.61 demonstrated substantial reductions in RMSE and MAE, with paired t-test p-values of 1.81e-06 and 1.08e-07, respectively, indicating a marked improvement in predictive accuracy with the inclusion of the "Side" variable.

Lingual Sites

For the lingual sites, the "Side" predictor significantly affected the PPD distribution, as evidenced by the LRT results ($\chi^2 = 10.575$, $p = 0.014$). The location parameter (μ) estimate decreased slightly from 0.647 to 0.637. Significant changes were observed in the scale (σ) parameter with a p-value of 0.002, while the skewness (ν) parameter did not exhibit a significant change (Table I.62). The cross-validation metrics in Table I.63 revealed significant differences in RMSE and MAE, with paired t-test p-values of 3.86e-05 and 3.29e-06, respectively, indicating that the inclusion of the "Side" predictor enhances the model's functional properties.

Mesial Lingual Sites

The inclusion of the "Side" predictor had a significant impact on the PPD distribution at the mesial lingual sites. The LRT results indicated a significant effect with a χ^2 value of 180.48 and a p-value less than 2e-16. The location parameter (μ) estimate increased from 1.101 to 1.215, and significant changes were observed in the scale (σ)

and skewness (ν) parameters (Table I.64). The cross-validation results in Table I.65 showed significant reductions in both RMSE and MAE, with paired t-test p-values of 3.86e-05 and 3.29e-06, respectively, indicating enhanced predictive accuracy of the model with the inclusion of the "Side" predictor.

4.1.6 Upper First Molars (16, 26)

Basic Methods

The results of the statistical tests comparing the PPD means, medians, and variances across different dental sites between the upper first molar (tooth 16) and the contralateral upper first molar (tooth 26) are summarized in Table I.66. These tests aim to verify the symmetry of the contralateral sites of teeth 16 and 26.

At the distal vestibular (DV) sites, significant differences were found in both the median PPD values (W = 1661474, p < 2.2e-16) and the distribution shapes as indicated by the Kolmogorov-Smirnov (K-S) test (D = 0.178, p < 2.2e-16). The variance comparison also showed a significant difference (F = 21.548, p = 3.56e-06). These results indicate asymmetry in both the central tendency and the variability of PPD between the contralateral DV sites.

For the vestibular (V) sites, significant differences were observed in the median PPD values (W = 1882430, p = 2.95e-06) and variance (F = 9.084, p = 2.60e-03), supported by the K-S test (D = 0.061, p = 1.01e-03). These findings suggest asymmetry in both the central tendency and the variability of PPD between the contralateral V sites.

The mesial vestibular (MV) sites exhibited significant differences in both median (W = 1874751, p = 4.02e-06) and variance (F = 4.038, p = 4.46e-02), with the K-S test also confirming significant distributional differences (D = 0.069, p = 1.37e-04). These results indicate asymmetry at the MV sites.

The distal lingual (DL) sites showed significant differences in both median (W = 2317438, p < 2.2e-16) and variance (F = 4.150, p = 4.17e-02), supported by the K-S test (D = 0.128, p = 1.17e-14). These findings suggest asymmetry in both the central tendency and the variability at the DL sites.

For the lingual (L) sites, no significant differences were observed in either median (W = 2089879, p = 0.062) or variance (F = 0.01, p = 0.921), with the K-S test also indicating no significant distributional differences (D = 0.022, p = 0.696). These results suggest symmetry at the L sites.

Finally, the mesial lingual (ML) sites demonstrated significant differences in both median (W = 2487157, p < 2.2e-16) and variance (F = 16.52, p = 4.89e-05), supported by the K-S test (D = 0.204, p < 2.2e-16). These results indicate asymmetry at the ML sites.

Overall, the findings highlight significant asymmetries at the DV, V, MV, DL, and ML sites in both the central tendency and variability of PPD between teeth 16 and 26, as detailed in Table I.66. However, the L site demonstrated symmetry, suggesting consistent PPD distributions across these contralateral sites.

Table I.66 reports Bhattacharyya coefficients ranging from 0.963 to 1.000, showing high distributional similarity between the contralateral sites. The correlation coefficients vary from 0.43 to 0.50, indicating a moderate linear relationship between the PPD values at these sites.

Advanced Methods

Distal Vestibular Sites

The analysis of the distal vestibular sites using GAMLSS models, as shown in Table I.67, indicates a significant effect of the "Side" predictor on the PPD distribution. The likelihood ratio test (LRT) revealed a χ^2 value of 182.27 with a p-value less than $2.2e-16$, demonstrating the substantial impact of the "Side" variable. The estimates for the location parameter (μ) slightly decreased from 0.748 to 0.737 with the inclusion of the "Side" predictor. Additionally, significant changes were observed in the scale (σ) and skewness (ν) parameters with p-values of less than 2e-16 and 0.034, respectively. Cross-validation metrics, presented in Table I.68, showed significant reductions in both RMSE and MAE, indicating that the inclusion of the "Side" predictor improves the model's predictive accuracy.

Vestibular Sites

At the vestibular sites, the "Side" predictor significantly influenced the PPD distribution, as evidenced by the LRT results (χ^2 = 28.104, p = 3.45e-06). The location parameter (μ) estimate showed a slight decrease from 0.582 to 0.578 with the inclusion of the "Side" predictor. Significant changes were observed in the skewness (ν) parameter with a p-value of 0.002 (Table I.69). The cross-validation results in Table I.70 revealed significant differences in RMSE and MAE, further confirming the enhanced predictive performance of the model when incorporating the "Side" predictor.

Mesial Vestibular Sites

For the mesial vestibular sites, the GAMLSS models indicated a significant effect of the "Side" predictor on the PPD distribution, as demonstrated by the LRT results (χ^2 = 38.831, p = 1.88e-08). The estimates for the location parameter (μ) slightly decreased from 0.787 to 0.781 with the inclusion of the "Side" predictor. Significant changes were observed in the scale (σ) and skewness (ν) parameters with p-values of 1.52e-12 and less than 2e-16, respectively (Table I.71). The cross-validation metrics, shown in Table I.72, indicated significant improvements in RMSE and MAE, suggesting that the inclusion of the "Side" predictor enhances the model's functional properties for these sites.

Distal Lingual Sites

The distal lingual sites exhibited a significant effect of the "Side" predictor on the PPD distribution. The LRT results, as presented in Table I.73, showed a χ^2 value of 89.999 with a p-value less than 2.2e-16. The location parameter (μ) estimate increased from 1.140 to 1.273, and significant changes were observed in the scale (σ) parameter with a p-value of 0.117. The cross-validation metrics in Table I.74 did not show significant improvements in RMSE and MAE, indicating that the inclusion of the "Side" variable does not significantly enhance the model's predictive accuracy for these sites.

Lingual Sites

The GAMLSS models for the lingual sites revealed a significant effect of the "Side" predictor on the PPD distribution. The LRT results indicated a χ^2 value of 20.209 with a p-value of 1.5e-04. The location parameter (μ) estimate remained stable at 0.734, and no significant changes were observed in the scale (σ) and skewness (ν) parameters (Table I.75). The cross-validation results, as shown in Table I.76, revealed no significant differences in RMSE and MAE, suggesting that the inclusion of the "Side" predictor does not significantly enhance the model's functional properties for these sites.

Mesial Lingual Sites

The mesial lingual sites demonstrated a significant effect of the "Side" predictor on the PPD distribution, as shown by the LRT results (χ^2 = 289.46, p = less than 2.20e-16). The location parameter (μ) estimate increased from 1.131 to 1.322, and significant changes were observed in the scale (σ) and skewness (ν) parameters (Table I.77). Cross-

validation metrics in Table I.78 showed significant reductions in RMSE and MAE, confirming the enhanced predictive performance of the model with the inclusion of the "Side" predictor for these sites.

4.1.7 Upper Second Molars (17, 27)

Basic Methods

The results of the statistical tests comparing the PPD means, medians, and variances across different dental sites between the upper second molar (tooth 17) and the contralateral upper second molar (tooth 27) are summarized in Table I.79. These tests aim to verify the symmetry of the contralateral sites of teeth 17 and 27.

At the distal vestibular (DV) sites, significant differences were observed in both the median PPD values (W = 1850993, p = 6.36e-10) and the distribution shapes as indicated by the Kolmogorov-Smirnov (K-S) test (D = 0.084, p = 1.31e-06). The variance comparison also showed a significant difference (F = 21.365, p = 3.91e-06), indicating asymmetry in both the central tendency and the variability of PPD between the contralateral DV sites.

For the vestibular (V) sites, significant differences were found in the median PPD values (W = 1925833, p = 4.10e-05) and variance (F = 7.675, p = 5.63e-03), supported by the K-S test (D = 0.058, p = 2.36e-03). These results suggest asymmetry in both the central tendency and the variability of PPD between the contralateral V sites.

The mesial vestibular (MV) sites exhibited significant differences in the median (W = 1757368, p < 2.2e-16) and variance (p = 1.52e-12), with the K-S test also confirming significant distributional differences (D = 0.117, p = 1.241). These findings indicate asymmetry at the MV sites.

The distal lingual (DL) sites showed significant differences in the median (W = 2343649, p = 6.78e-16) and variance (p = 1.21e-11), supported by the K-S test (D = 0.115). These results suggest asymmetry in both the central tendency and the variability at the DL sites.

For the lingual (L) sites, no significant differences were observed in either median (W = 2115555, p = 0.105) or variance (F = 3.020, p = 0.082), with the K-S test also indicating no significant distributional differences (D = 0.020, p = 0.826). These results suggest symmetry at the L sites.

Finally, the mesial lingual (ML) sites demonstrated significant differences in the

median (W = 2382157, p < 2.2e-16) and variance (p = 3.33e-16), supported by the K-S test (D = 0.134). These results indicate asymmetry at the ML sites.

Overall, the findings highlight significant asymmetries at the DV, V, MV, DL, and ML sites in both the central tendency and variability of PPD between teeth 17 and 27, as detailed in Table I.79. However, the L site demonstrated symmetry, suggesting consistent PPD distributions across these contralateral sites.

In Table I.79, the Bhattacharyya coefficients range from 0.982 to 1.000, indicating high similarity between the PPD distributions at the contralateral sites. The correlation coefficients range from 0.46 to 0.52, indicating a moderate linear relationship between the PPD values at these sites.

Advanced Methods

Distal Vestibular Sites

The analysis of distal vestibular sites using GAMLSS models, as shown in Table I.80, indicates a significant effect of the "Side" predictor on the PPD distribution. The likelihood ratio test (LRT) revealed a χ^2 value of 71.466 with a p-value of 2.07e-15, demonstrating the substantial impact of the "Side" variable. The estimates for the location parameter (μ) increased slightly from 0.806 to 0.821 with the inclusion of the "Side" predictor. Significant changes were observed in the scale (σ) and skewness (ν) parameters with p-values of 1.01e-06 and 7.17e-12, respectively. Cross-validation metrics, presented in Table I.81, showed significant reductions in both RMSE and MAE, indicating that the inclusion of the "Side" predictor improves the model's predictive accuracy.

Vestibular Sites

At the vestibular sites, the "Side" predictor significantly influenced the PPD distribution, as evidenced by the LRT results (χ^2 = 16.90, p = 7.40e-05). The location parameter (μ) estimate decreased slightly from 0.681 to 0.661 with the inclusion of the "Side" predictor. Significant changes were observed in the skewness (ν) parameter with a p-value of 0.137 (Table I.82). The cross-validation results in Table I.83 revealed significant differences in RMSE and MAE, further confirming the enhanced predictive performance of the model when incorporating the "Side" predictor.

Mesial Vestibular Sites

For the mesial vestibular sites, the GAMLSS models indicated a significant effect of the "Side" predictor on the PPD distribution, as demonstrated by the LRT results (χ^2 = 31.65, p = 6.20e-07). The estimates for the location parameter (μ) slightly decreased from 0.804 to 0.794 with the inclusion of the "Side" predictor. Significant changes were observed in the scale (σ) and skewness (ν) parameters with p-values of 1.24e-08 and 7.44e-05, respectively (Table I.84). The cross-validation metrics, shown in Table I.85, indicated significant improvements in RMSE and MAE, suggesting that the inclusion of the "Side" predictor enhances the model's functional properties for these sites.

Distal Lingual Sites

The distal lingual sites exhibited a significant effect of the "Side" predictor on the PPD distribution. The LRT results, as presented in Table I.86, showed a χ^2 value of 103.57 with a p-value less than 2.2e-16. The location parameter (μ) estimate increased from 1.105 to 1.216, and significant changes were observed in the scale (σ) parameter with a p-value less than 2e-16. The cross-validation metrics in Table I.87 did not show significant improvements in RMSE and MAE, indicating that the inclusion of the "Side" variable does not significantly enhance the model's predictive accuracy for these sites.

Lingual Sites

The GAMLSS models for the lingual sites revealed a significant effect of the "Side" predictor on the PPD distribution. The LRT results indicated a χ^2 value of 5.853 with a p-value of 0.119. The location parameter (μ) estimate remained stable at 0.811, and significant changes were observed in the scale (σ) and skewness (ν) parameters (Table I.88). The cross-validation results, as shown in Table I.89, revealed no significant differences in RMSE and MAE, suggesting that the inclusion of the "Side" predictor does not significantly enhance the model's functional properties for these sites.

Mesial Lingual Sites

The mesial lingual sites demonstrated a significant effect of the "Side" predictor on the PPD distribution, as shown by the LRT results (χ^2 = 107.71, p = less than 2e-16). The location parameter (μ) estimate increased from 1.315 to 1.434, and significant changes were observed in the scale (σ) and skewness (ν) parameters (Table I.90). Cross-

validation metrics in Table I.91 showed significant reductions in RMSE and MAE, confirming the enhanced predictive performance of the model with the inclusion of the "Side" predictor for these sites.

4.2 Symmetry Measure

4.2.1 Numerical Assessment

The analysis of symmetry measures (SM) for pairs of contralateral upper teeth provides insights into the degree of symmetry present across different dental sites. The SM values, where an SM of 1 denotes perfect symmetry and 0 indicates complete asymmetry, were computed for pairs of contralateral teeth across six specific dental sites. The Table 4.14 details the summary statistics of SM values for upper contralateral teeth pairs on sites DV, V, MV, DL, L, and ML.

The mean SM values for the DV site (Table 4.14) were consistently high, with an average of 0.85 for the central incisors (11, 21), lateral incisors (12, 22), and canines (13, 23). This value slightly decreased for the first and second premolars (14, 24; 15, 25) and first and second molars (16, 26; 17, 27). Median values for all pairs remained at 1, indicating that at least half of the teeth pairs exhibited perfect symmetry. The variance in SM values, although minimal, showed an incremental increase from the central incisors to the molars, with the highest variance observed in the second molars (17, 27).

Table 4.14: Summary of SM values for upper contralateral teeth pairs across six dental sites.

		Pairs of Contralateral Teeth						
Site	Parameter	(11,21)	(12,22)	(13,23)	(14,24)	(15,25)	(16,26)	(17,27)
DV	Mean	0.85	0.85	0.85	0.82	0.80	0.78	0.83
	Median	1	1	1	1	1	1	1
	Variance	0.042	0.042	0.042	0.048	0.055	0.055	0.046
V	Mean	0.86	0.84	0.84	0.84	0.83	0.83	0.81
	Median	1	1	1	1	1	1	1
	Variance	0.047	0.051	0.051	0.053	0.065	0.059	0.055
MV	Mean	0.89	0.84	0.84	0.82	0.80	0.80	0.78
	Median	1	1	1	1	1	1	1
	Variance	0.036	0.043	0.043	0.048	0.051	0.054	0.051
DL	Mean	0.85	0.84	0.84	0.81	0.78	0.77	0.76
	Median	1	1	1	1	0.7	0.7	0.7
	Variance	0.041	0.043	0.043	0.044	0.052	0.051	0.055
L	Mean	0.86	0.85	0.85	0.85	0.82	0.83	0.81
	Median	1	1	1	1	1	1	1
	Variance	0.046	0.048	0.048	0.048	0.051	0.058	0.046
ML	Mean	0.89	0.85	0.85	0.82	0.80	0.78	0.80
	Median	1	1	1	1	1	1	1
	Variance	0.032	0.041	0.041	0.045	0.050	0.055	0.052

At the V site (Table 4.14), the mean SM values ranged from 0.81 to 0.86, with the highest mean observed for the central incisors (11, 21) and the lowest for the second molars (17, 27). The median values again consistently showed perfect symmetry, while the variances indicated a slight increase from the central incisors to the molars, suggesting a gradual decrease in symmetry.

The MV site (Table 4.14) displayed a higher mean SM value of 0.89 for the central incisors (11, 21) but a noticeable decrease to 0.78 for the second molars (17, 27). Median values remained at 1, consistent with other sites. Variances followed a similar pattern to the other sites, with an increase from the central incisors to the molars.

For the DL site (Table 4.14), the mean SM values varied from 0.85 for the central incisors (11, 21) to 0.76 for the second molars (17, 27). Notably, the median values dropped to 0.7 for the first and second molars (16, 26; 17, 27), indicating a higher degree of asymmetry in these teeth pairs. Variance increased progressively from the central incisors to the molars.

In the L site (Table 4.14), mean SM values were relatively high, ranging from 0.81 to 0.86. The upper central incisors (11, 21) and first premolars (14, 24) had the highest

mean values, while the second molars (17, 27) had the lowest. Median values were consistently at 1, and variances showed a slight increase from the central incisors to the molars.

The ML site (Table 4.14) revealed mean SM values starting at 0.89 for the central incisors (11, 21) and dropping to 0.80 for the second molars (17, 27). Median values remained consistently at 1, while the variances showed an incremental rise from the central incisors to the molars.

4.2.2 Visual Assessment: Bar Plots and Kernel Density Estimates

The grade of symmetry is visually assessed through symmetry measure bar plots and kernel density estimates.

The symmetry measure bar plots depict the results for Symmetry Measure (SM) values across all pairs of contralateral upper teeth, which categorize the SM values rounded to one decimal place and display the percentage distribution on the y-axis. These figures collectively illustrate the degree of symmetry a pair of contralateral sites. SM varys from 1 to 0 where SM=1 represent perfect symmetry and SM=0 complete asymmetry.

The kernel density estimates depict the prevalence of PPD values at contralateral sites, represented by blue and red lines. These plots are used to assess symmetry by evaluating the overlap of the KDE curves. Complete overlap of the curves at a specific site suggests a high grade of symmetry. Similar to the symmetry measure bar plots, the y-axis in KDE plots represents the probability of each PPD value occurring, providing a detailed visualisation of symmetry and asymmetry among contralateral tooth pairs.

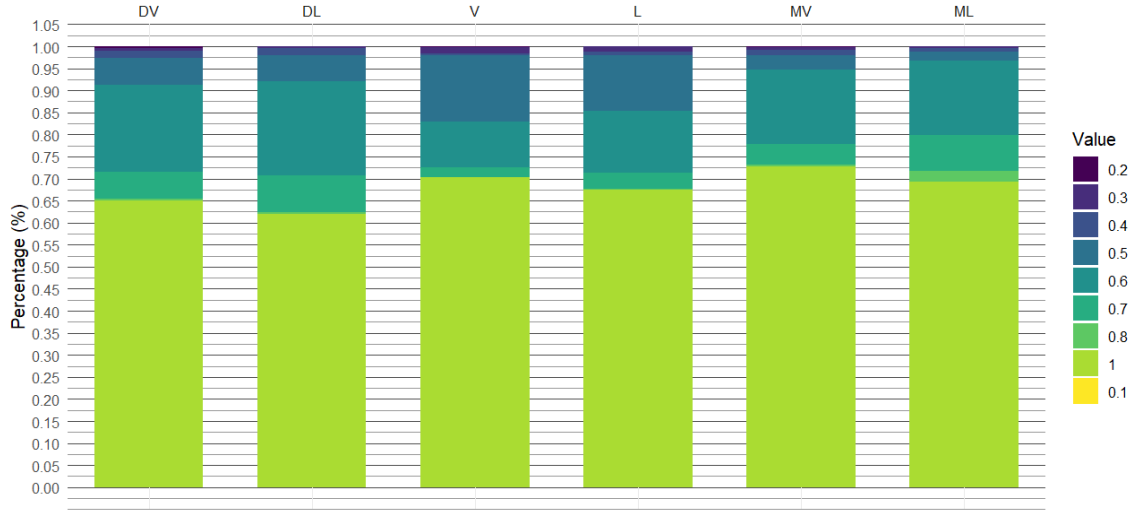
This section we present the bar plots and kernel density plots for the upper central incisors, all other figures regarding this results can be found in appendix I.

Upper Central Incisors (11, 21)

Bar Plots

The bar plots for the upper central incisors (Figure I.1) shows the SM values across all six sites (DV, V, MV, DL, L, ML). The bar plots indicate that the distribution is heavily concentrated in the higher SM categories, particularly around 0.9 and 1.0.

Figure 4.1

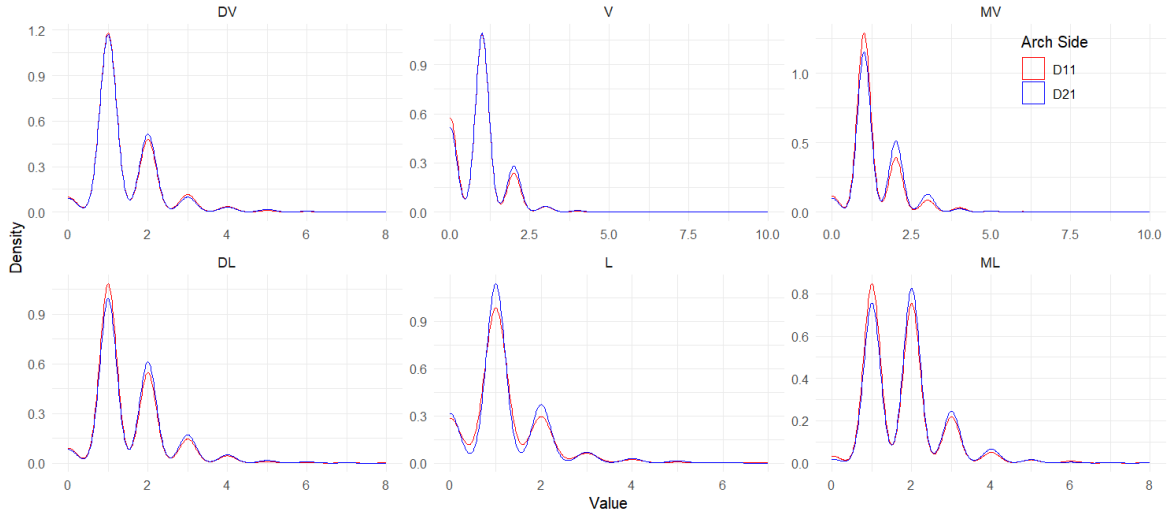


Upper Central Incisors Percentages of SM Values by Site

This suggests that most central incisor pairs exhibit a high degree of symmetry, with a significant portion of the sites showing nearly perfect symmetry. The sites DV, V, and MV, in particular, display a higher percentage of SM values at 1.0, indicating a predominant occurrence of perfect symmetry at these locations.

Kernel Density Estimates

Figure 4.2



Kernel Density Estimates of Upper Central Incisors by Site

Upper Lateral Incisors (12, 22)

Bar Plots

Figure I.3 illustrates the SM values for the upper lateral incisors. The bar plots reveal that the distribution remains skewed towards higher SM categories, similar to the central incisors. The percentages of SM values at 0.9 and 1.0 are slightly lower compared to the central incisors, indicating a marginally greater variability in symmetry. However, the sites DV, V, and MV still show a significant concentration of high SM values, reflecting a high degree of symmetry at these locations.

Upper Canines (13, 23)

Bar Plots

Figure I.5 depicts the SM values for the upper canines. The distribution shows a broader range of SM values, with a noticeable percentage of occurrences in the mid-range categories (0.5 to 0.8). Despite this broader distribution, the sites DV, V, and MV continue to exhibit higher percentages of SM values at 0.9 and 1.0, indicating that while there is an increase in asymmetry among canines, these sites still maintain a predominant occurrence of symmetry.

Upper First Premolars (14, 24)

Bar Plots

Figure I.7 presents the SM values for the upper first premolars. The bar plots indicate a further broadening of the distribution, with a significant portion of SM values falling below 1, mid-range SM categories. This broader spread suggests more noticeable variability in the symmetry of the first premolars. Sites such as DL and L show a higher occurrence of mid-range SM values, reflecting increased asymmetry compared to the anterior teeth.

Upper Second Premolars (15, 25)

Bar Plots

Figure Figure I.9 shows the SM values for the upper second premolars. The trend of increasing spread continues, with bar plots displaying significant percentages in both

the mid-range and lower SM categories. This indicates a further increase in variability and a higher occurrence of asymmetry compared to the first premolars. The site ML, in particular, shows a broader distribution with lower SM values, suggesting notable asymmetry at this location.

Upper First Molars (16, 26)

Bar Plots

Figure I.11 illustrates the SM values for the upper first molars. The bar plots demonstrate an even broader distribution, with significant percentages in the lower SM categories (0.2 to 0.6). This suggests a substantial increase in the variability of symmetry, with many first molar pairs exhibiting noticeable asymmetry. Sites DL and L show a particularly high occurrence of lower SM values, indicating greater asymmetry at these locations.

Upper Second Molars (17, 27)

Bar Plots

Figure I.13 depicts the SM values for the upper second molars. The distribution is the broadest among all tooth pairs, with bar plots showing a substantial portion of SM values in the lower categories (0.1 to 0.5). This indicates that the second molars exhibit the highest degree of asymmetry, with a significant percentage of occurrences falling well below perfect symmetry. Sites ML and L display particularly high occurrences of lower SM values, reflecting a predominant asymmetry at these locations.

Overall The bar plots across these figures This trend is visually represented by the concentration of SM values around 1 for the anterior teeth and a broader spread of SM values for the posterior teeth. These findings are consistent with the statistical summaries in the tables, reinforcing the observation that posterior teeth are more prone to asymmetry compared to their anterior counterparts. The figures collectively demonstrate a clear trend: symmetry is highest among the anterior teeth (central and lateral incisors) and decreases progressively as one moves posteriorly towards the molars. The bar plots provide a clear visual representation of this trend, highlighting the increasing variability and prevalence of asymmetry from the incisors to the molars. This pattern aligns with the statistical summaries provided in the tables, reinforcing

the observation that posterior teeth are more prone to asymmetry than their anterior counterparts.

Kernel Density Estimates Plots Analysis

The central incisors (Figure I.2) showed almost perfect matches in site distal vestibular and vestibular and the lingual site being the most discordant site. In the lateral incisors (Figure I.4) the most concordant sites were the distal lingual, mesial vestibular and mesial lingual and the most discordant the lingual site. The canines (Figure I.6) showed two concordant sites, the mesial vestibular and the lingual and the most discordant stes were the distal vestibular and distal lingual. The premolars (Figures I.8 and I.10) showed high concordance in mesial vestibular and distal vestibular in the first premolar and moderate concordance in the site lingual in second premolar with the other sites showing low concordance or even large discordance. In both molars (Figures I.12 and I.14) the lingual sites showed high concordance, with all the other sites showing low concordance or large discordance as in sites distal and mesial lingual of the second molar.

The grade of KDE lines overlap vary from site to site with the sites of central and lateral incisors, and canines presenting more overlapping lines then the posterior teeth: first and second premolars, and first and second molars in line with the mean SM values and respective variances (Table 4.14)

4.2.3 Clinical perception of symmetry

The correlation analysis revealed a Spearman correlation coefficient of 0.96, indicating a strong agreement between the scores assigned by the professionals and those computed by the SM function.

4.3 Data preparation for MoDau

4.3.1 Hot Deck Imputation

The Table 4.15 show the number of not planned missinng PPD values the would affect the performance of XGBoost models, reason why we decide to perfom H-D imputation.

Table 4.15: Missing Data by Tooth Site in Upper Arch

	Sites							Sites					
Tooth	**MV**	**V**	**DV**	**ML**	**L**	**DL**	**Tooth**	**MV**	**V**	**DV**	**ML**	**L**	**DL**
11	538	544	547	542	534	594	**21**	498	501	505	505	496	522
12	546	535	541	571	526	556	**22**	521	525	531	601	519	529
13	381	381	379	405	372	448	**23**	384	397	385	432	373	416
14	815	822	821	816	802	804	**24**	823	832	826	831	810	810
15	784	787	783	775	776	778	**25**	782	784	785	771	767	774
16	941	940	943	929	926	934	**26**	964	966	984	943	945	947
17	924	918	923	912	912	915	**27**	953	951	963	909	907	916

In this section, the results of H-D imputation for tooth 21 are presented, along with the respective tables and figure. The results for the other teeth are also included, with a focus on highlighting the most significant findings, and the respective tables figures can be consulted in Appendix II.

Upper Right Central Incisor - Tooth 11

The following sections present the detailed comparisons between the original and H-D imputed PPD values across different sites of tooth 11. The analysis is divided into three main subsections: comparison of PPD values, comparison of PPD > 3 proportions, and overall PPD statistics before and after imputation.

Original vs. H-D Imputed PPD Values by Site and Unique Value Categories

Table 4.16 provides a comparative analysis of the original and imputed PPD values across various sites for the upper right central incisor (Tooth 11). The table presents the unique value categories (k) for each site, along with corresponding statistics (D), p-values, confidence intervals, and the original and imputed values.

For site 11DV, significant differences were observed in categories $k = 1$ and $k = 2$, with p-values of 0.028 and 0.016, respectively. The confidence intervals for these categories do not contain zero, indicating a significant deviation between the original and imputed values. In contrast, other categories showed no significant differences.

Similarly, site 11V showed no significant differences across all categories, with the lowest p-value being 0.219 for $k = 0$. This suggests that the imputation process did not introduce significant deviations from the original data for this site.

For site 11MV, all categories indicated non-significant differences, with p-values well

above the significance threshold. The smallest p-value recorded was 0.649 for $k = 0$, again indicating a good agreement between the original and imputed values.

Site 11DL also exhibited no significant differences across all categories, with the lowest p-value of 0.121 for $k = 1$. This suggests a reliable imputation process for this site.

The comparison for site 11L showed non-significant differences across all categories, with p-values ranging from 0.633 to 0.778, indicating consistent imputed values with the original data.

Lastly, site 11ML displayed no significant differences across all categories, with p-values well above the significance threshold. This suggests that the imputation process was effective in maintaining the integrity of the original data.

Table 4.16: Comparison of Original and Imputed Periodontal Pocket Depth Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
11DV	0	6.661e-01	0.414	[-6.378e-03, 1.542e-02]	5.121e-02	4.669e-02
	1	4.856e+00	**0.028**	**[3.111e-03, 5.252e-02]**	6.113e-01	5.834e-01
	2	5.802e+00	**0.016**	**[-4.947e-02, -5.167e-03]**	2.492e-01	2.765e-01
	3	4.621e-01	0.497	[-1.641e-02, 7.939e-03]	6.022e-02	6.446e-02
	4	3.914e-01	0.532	[-9.357e-03, 4.812e-03]	1.911e-02	2.139e-02
	5	2.614e-01	0.609	[-2.723e-03, 4.624e-03]	5.770e-03	4.819e-03
	6	1.140e-01	0.736	[-2.018e-03, 2.850e-03]	2.524e-03	2.108e-03
	7	3.252e-02	0.857	[-1.183e-03, 1.421e-03]	7.212e-04	6.024e-04
11V	0	1.511e+00	0.219	[-8.529e-03, 3.714e-02]	2.950e-01	2.807e-01
	1	3.563e-03	0.952	[-2.426e-02, 2.579e-02]	5.598e-01	5.590e-01
	2	2.606e+00	0.106	[-3.085e-02, 2.916e-03]	1.225e-01	1.364e-01
	3	8.857e-02	0.766	[-7.746e-03, 5.700e-03]	1.765e-02	1.867e-02
	4	6.107e-02	0.805	[-3.312e-03, 2.567e-03]	3.242e-03	3.614e-03
	5	1.606e-02	0.899	[-8.611e-04, 9.792e-04]	3.602e-04	3.012e-04
	6	4.821e-02	0.826	[-1.416e-03, 1.770e-03]	1.081e-03	9.036e-04
11MV	0	2.076e-01	0.649	[-9.073e-03, 1.456e-02]	5.967e-02	5.693e-02
	1	1.498e-02	0.903	[-2.227e-02, 2.523e-02]	6.671e-01	6.657e-01
	2	1.366e-01	0.712	[-2.417e-02, 1.649e-02]	2.031e-01	2.069e-01
	3	9.612e-05	0.992	[-1.038e-02, 1.049e-02]	4.493e-02	4.488e-02
	4	2.351e-02	0.878	[-7.121e-03, 6.086e-03]	1.725e-02	1.777e-02
	5	8.479e-02	0.771	[-3.894e-03, 2.882e-03]	4.313e-03	4.819e-03
	6	1.099e-01	0.740	[-2.021e-03, 2.837e-03]	2.516e-03	2.108e-03
	7	3.136e-02	0.859	[-1.183e-03, 1.416e-03]	7.189e-04	6.024e-04
11DL	0	3.071e-01	0.580	[-7.590e-03, 1.355e-02]	4.696e-02	4.398e-02
	1	2.411e+00	0.121	[-5.219e-03, 4.514e-02]	5.627e-01	5.428e-01
	2	2.327e+00	0.127	[-4.103e-02, 5.076e-03]	2.850e-01	3.030e-01
	3	4.309e-01	0.512	[-1.811e-02, 9.002e-03]	7.557e-02	8.012e-02
	4	1.408e-02	0.906	[-7.064e-03, 7.974e-03]	2.274e-02	2.229e-02
	5	1.251e-01	0.724	[-4.248e-03, 2.943e-03]	4.769e-03	5.422e-03
	6	5.854e-02	0.809	[-1.414e-03, 1.807e-03]	1.101e-03	9.036e-04
	7	3.901e-02	0.843	[-1.184e-03, 1.446e-03]	7.337e-04	6.024e-04
11L	0	2.279e-01	0.633	[-1.427e-02, 2.345e-02]	1.709e-01	1.663e-01
	1	7.966e-02	0.778	[-2.119e-02, 2.832e-02]	5.933e-01	5.898e-01
	2	2.660e-01	0.606	[-2.446e-02, 1.426e-02]	1.780e-01	1.831e-01
	3	6.216e-02	0.803	[-1.084e-02, 8.390e-03]	3.733e-02	3.855e-02
	4	5.984e-01	0.439	[-8.374e-03, 3.610e-03]	1.328e-02	1.566e-02
	5	1.547e-01	0.694	[-2.322e-03, 3.477e-03]	3.589e-03	3.012e-03
	6	9.267e-02	0.761	[-1.901e-03, 2.594e-03]	2.154e-03	1.807e-03
	7	1.278e-01	0.721	[-2.387e-03, 1.644e-03]	1.436e-03	1.807e-03
11ML	0	1.525e-02	0.902	[-6.121e-03, 6.943e-03]	1.728e-02	1.687e-02
	1	8.910e-02	0.765	[-2.119e-02, 2.880e-02]	4.384e-01	4.346e-01
	2	4.602e-01	0.498	[-3.316e-02, 1.610e-02]	3.906e-01	3.991e-01
	3	1.284e-01	0.720	[-1.301e-02, 1.883e-02]	1.138e-01	1.108e-01
	4	5.752e-02	0.811	[-7.018e-03, 8.972e-03]	2.628e-02	2.530e-02
	5	8.622e-03	0.926	[-4.707e-03, 4.281e-03]	7.919e-03	8.133e-03
	6	7.348e-02	0.786	[-2.899e-03, 3.824e-03]	4.680e-03	4.217e-03
	7	1.593e-02	0.900	[-8.610e-04, 9.786e-04]	3.600e-04	3.012e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table 4.17 focuses on the comparison of proportions of PPD values greater than 3 between the original and imputed datasets across different sites.

Table 4.17: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
11DV	3.38e-02	0.854	[-9.18e-03, 7.60e-03]	2.81e-02	2.89e-02
11V	1.78e-03	0.966	[-3.66e-03, 3.51e-03]	5.04e-03	5.12e-03
11MV	1.19e-02	0.913	[-8.36e-03, 7.48e-03]	2.52e-02	2.56e-02
11DL	8.52e-03	0.926	[-9.03e-03, 8.22e-03]	2.97e-02	3.01e-02
11L	2.42e-01	0.623	[-9.100-03, 5.44e-03]	2.05e-02	2.23e-02
11ML	7.90e-02	0.779	[-8.39e-03, 1.12e-02]	3.10e-02	3.86e-02

For site 11DV, the p-value of 0.854 indicates no significant difference between the original and imputed proportions, with a confidence interval ranging from -0.009 to 0.008. This trend is consistent across other sites such as 11V, 11MV, 11DL, and 11ML, all of which have p-values significantly higher than the threshold, suggesting no substantial deviation between the original and imputed proportions.

Interestingly, site 11L showed a slightly lower p-value of 0.623, yet it still indicates no significant difference with a confidence interval from -0.009 to 0.005. Overall, the imputation method appears robust in maintaining the proportions of PPD values greater than 3 across all sites.

PPD Statistics Before and After H-D Imputation by Site

Table 4.18 presents the PPD statistics before and after H-D imputation across different sites, including median, variance, and the results of statistical tests.

For site DV, the median PPD values remained consistent before and after imputation. However, the Wilcoxon test result ($W = 4752302$, $p = 0.013$) suggests a statistically significant difference. The variance remained relatively stable (0.722 before and 0.717 after), with no significant difference indicated by the F-test ($p = 0.848$).

Site V showed no significant differences in median or variance, with the Wilcoxon test ($p = 0.088$) and F-test ($p = 0.882$) indicating a good agreement between the original and imputed values. This trend continues for site MV, with the Wilcoxon

test and F-test both indicating no significant differences (p-values of 0.640 and 0.945, respectively).

For site DL, while the median remained the same, the Wilcoxon test indicated no significant difference (p = 0.084). The variance showed a slight increase after imputation (from 0.748 to 0.771), but this change was not statistically significant (p = 0.410).

Table 4.18: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	**Stats**	**Tooth 11**		**Test Results**		**K-S**
		BI	**AI**	**Stats**	**p**	
DV	Median Variance	1 0.722	1 0.717	W = 4752302 F = 2.399	**1.291e-2** 0.122	D = 0.032 p = 0.848
V	Median Variance	1 0.549	1 0.555	W = 4712600 F = 0.008	8.775e-2 0.928	D = 0.015 p = 0.882
MV	Median Variance	1 0.675	1 0.666	W = 4644844 F = 0.005	0.640 0.945	D = 0.004 p = 1
DL	Median Variance	1 0.748	1 0.771	W = 4629765 F = 1.693	8.443e-2 0.193	D = 0.023 p = 0.410
L	Median Variance	1 0.757	1 0.773	W = 4673501 F = 0.195	0.422 0.661	D = 0.008 p = 1
ML	Median Variance	2 0.841	2 0.816	W = 4617429 F = 0.481	0.925 0.488	D = 0.004 p = 1

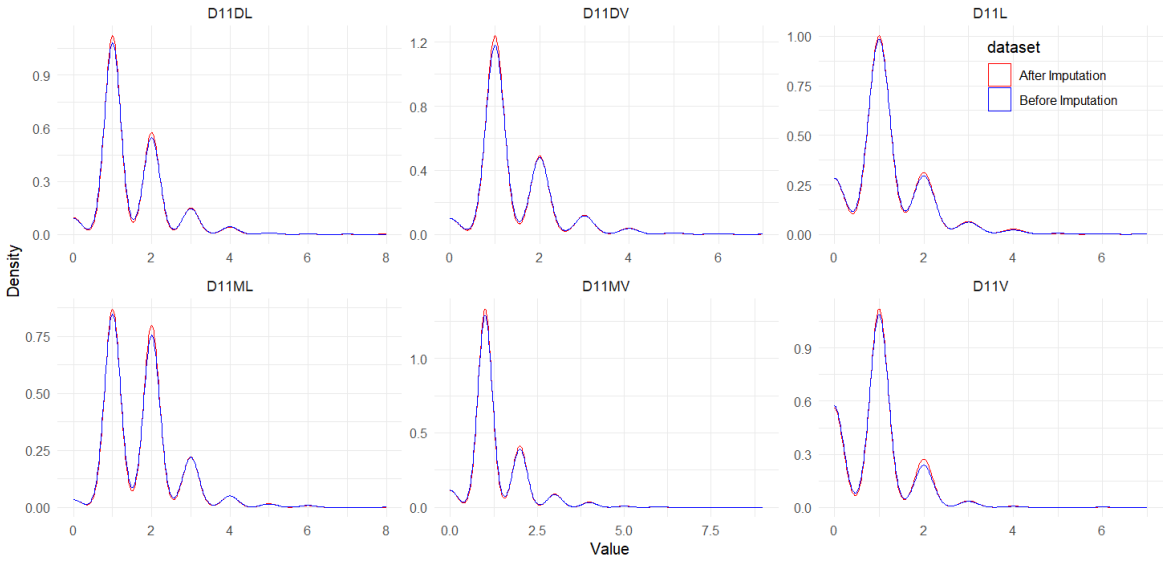
Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value; K-S – Kolmogorov-Smirnov; W – Wilcoxon Statistic; F – F-Statistic; D – Distance

Site L exhibited consistent median values and variances before and after imputation, with all tests indicating no significant differences (p-values of 0.422 for the Wilcoxon test and 1 for the F-test).

Lastly, site ML showed consistent median values before and after imputation. The Wilcoxon test (p = 0.925) and F-test (p = 0.488) confirmed no significant differences, indicating effective imputation.

Kernel Density Estimates of Imputed and Original

Figure 4.3



Kernel Density Estimates of Imputed and Original 11 by Site

In Figure 4.3, the KDE of the PPD distributions before and after imputation exhibit a high degree of overlap across all six sites. The minor discrepancies observed suggest only slight differences between the original and imputed data. These findings are consistent with the results of the K-S test (D statistic from 0.032 to 0.004), indicating a strong similarity between the two datasets.

Upper Right Lateral Incisor - Tooth 12

The subsequent sections present an in-depth analysis of the data quality assessment following H-D Imputation for the upper right lateral incisor (Tooth 12). The assessment is structured into three primary subsections: comparison of original versus imputed PPD values by site, comparison of PPD values greater than 3, and a comprehensive evaluation of PPD statistics before and after imputation.

Original vs. H-D Imputed PPD Values by Site

Table II.4 illustrates the comparison between original and imputed PPD values across various sites for Tooth 12. Significant differences were identified in several categories. Specifically, for site 12DV, the categories $k = 1$, $k = 2$, and $k = 7$ exhibited highly significant differences with p-values of 9.96×10^{-8}, 4.18×10^{-11}, and 0.004, respectively. These categories demonstrated substantial deviations in original versus imputed values, as reflected in their confidence intervals, which do not include zero.

For site 12V, no significant differences were detected across all categories, with the lowest p-value being 0.274 for $k = 2$, indicating a good concordance between the original and imputed datasets. Similarly, site 12MV showed non-significant differences for all categories, suggesting that the imputation maintained the integrity of the original data.

Site 12DL, however, displayed significant differences in categories $k = 1$, $k = 2$, and $k = 3$, with p-values of 0.013, 0.037, and 0.046, respectively. The confidence intervals for these categories excluded zero, indicating significant changes post-imputation. In contrast, site 12L showed no significant differences across all categories, with p-values ranging from 0.763 to 0.961, suggesting that the imputation process was effective.

Lastly, site 12ML did not exhibit any significant differences across all categories, confirming the robustness of the imputation method for this site.

Original vs. H-D Imputed PPD $>$ 3 Proportions by Site

Table II.5 focuses on the comparison of proportions of PPD values greater than 3 between the original and imputed datasets. The results indicate no significant differences across all sites. For instance, site 12DV showed a p-value of 0.845, with a confidence interval ranging from -0.009 to 0.011, suggesting that the imputation did not significantly alter the proportion of PPD values greater than 3.

Similar trends were observed for sites 12V, 12MV, 12DL, 12L, and 12ML, with all p-values well above the significance threshold. These results collectively indicate that the H-D imputation process preserved the proportions of PPD values greater than 3 across all assessed sites.

PPD Statistics Before and After H-D Imputation by Site

Table II.6 presents a comprehensive comparison of PPD statistics before and after H-D imputation. For site DV, the Wilcoxon test result (W = 4928995, $p = 1.65 \times 10^{-7}$) and the F-test for variance (F = 10.617, $p = 0.001$) indicate significant differences in

both median and variance, highlighting a notable change post-imputation.

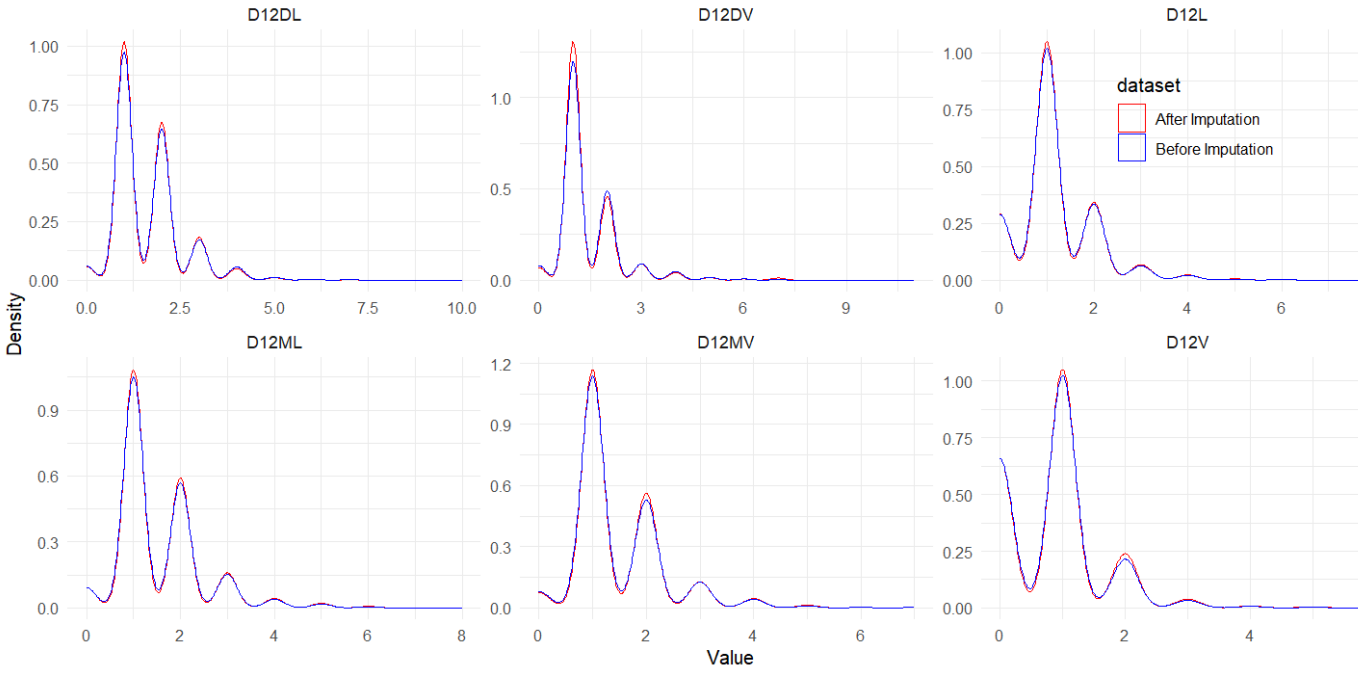
Conversely, site V showed no significant differences in median (W = 4703153, $p = 0.196$) or variance (F = 0.313, $p = 0.576$), indicating that the imputation did not significantly impact the original data distribution. This consistency extends to site MV, with non-significant p-values for both the Wilcoxon test and the F-test, suggesting a high fidelity of the imputed data.

Site DL exhibited significant differences in median values, as indicated by the Wilcoxon test (W = 4758760, $p = 0.006$), but the variance remained stable (F = 3.4863, $p = 0.062$). This implies that while the central tendency of the data changed post-imputation, the variability did not.

For site L, both the Wilcoxon test (W = 4659273, $p = 0.728$) and the F-test (F = 0.0339, $p = 0.854$) showed no significant differences, indicating effective imputation. Site ML also exhibited non-significant differences for both tests, affirming the reliability of the imputed values.

Kernel Density Estimates of Imputed and Original

Figure 4.4



Kernel Density Estimates of Imputed and Original 12 by Site

Upper Right Canine - Tooth 13

This section presents a comprehensive evaluation of the data quality after Hot Deck (H-D) Imputation for the Upper Right Canine (Tooth 13). The analysis includes a comparison of original versus imputed PPD values by site, the proportion of PPD values greater than 3, and overall PPD statistics before and after imputation.

Original vs. H-D Imputed PPD Values by Site

Table II.7 shows the comparative analysis of original and imputed PPD values across different sites for Tooth 13. The analysis reveals no significant differences across all sites and categories. For site 13DV, the highest p-value was observed at 0.8659 for $k = 0$, indicating no substantial deviation between original and imputed values. Similar results were observed for sites 13V, 13MV, 13DL, 13L, and 13ML, where all p-values exceeded the significance threshold, indicating consistent imputed values with the original data. These results demonstrate that the H-D imputation effectively maintained the integrity of the original PPD values across all assessed sites.

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.8 compares the proportions of PPD values greater than 3 between original and imputed datasets. The results indicate no significant differences across all sites. For instance, site 13DV showed a p-value of 0.813, with a confidence interval ranging from -0.008 to 0.006, suggesting no significant change in the proportion of PPD values greater than 3 post-imputation. Similar trends were observed for sites 13V, 13MV, 13DL, 13L, and 13ML, with all p-values well above the significance threshold. These findings suggest that the H-D imputation preserved the proportions of PPD values greater than 3 across all sites.

PPD Statistics Before and After H-D Imputation by Site

Table II.9 provides a detailed comparison of PPD statistics before and after H-D imputation. For site DV, the Wilcoxon test result (W = 4906766, p = 0.684) and the F-test for variance (F = 0.0618, p = 0.804) indicate no significant differences in median and variance values, suggesting that the imputation did not alter the original data distribution. Similar results were observed for sites V, MV, DL, L, and ML, where both the Wilcoxon test and the F-test indicated no significant differences in median and

variance values. These findings confirm that the H-D imputation method effectively maintained the original data characteristics across all assessed sites.

Kernel Density Estimates of Imputed and Original Values

Figure II.3 presents the kernel density estimates of imputed and original PPD values for Tooth 13 across different sites. The density plots show a high degree of overlap between the original and imputed distributions, further confirming the efficacy of the H-D imputation method in preserving the original data distribution. These visual representations align with the statistical tests, reinforcing the reliability and accuracy of the imputation process.

In summary, the results across all tables indicate that the H-D imputation method successfully maintained the integrity and distribution of the original PPD values for the Upper Right Canine (Tooth 13), with no significant differences observed across various statistical measures and sites.

Upper Right First Premolar - Tooth 14

This section presents a detailed analysis of the data quality assessment following Hot Deck (H-D) Imputation for the Upper Right First Premolar (Tooth 14). The evaluation includes a comparison of original versus imputed PPD values by site, the proportion of PPD values greater than 3, and overall PPD statistics before and after imputation.

Original vs. H-D Imputed PPD Values by Site

Table II.10 provides a comparison of original and imputed PPD values across different sites for Tooth 14. The analysis reveals no significant differences across most sites and categories. For site D14DV, the p-values indicate no substantial deviation between the original and imputed values, with the highest p-value being 0.991 for $k = 5$. Similarly, site D14V showed consistent results, with p-values indicating no significant differences in any category, the lowest being 0.971 for $k = 1$.

For site D14MV, the results showed no significant differences across all categories, with the highest p-value at 0.991 for $k = 6$. Site D14DL also displayed no significant differences, with the highest p-value at 0.978 for $k = 0$. Similar consistency was observed for site D14L, with no significant differences across all categories, and the highest p-value at 0.980 for $k = 0$.

Site D14ML presented no significant differences across most categories, except for

$k = 0$ where the p-value was 0.218, indicating a minor deviation. Overall, the H-D imputation method effectively maintained the integrity of the original PPD values across all assessed sites for Tooth 14.

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.11 compares the proportions of PPD values greater than 3 between the original and imputed datasets. The results indicate no significant differences across all sites. For site 14DV, the p-value of 0.630, with a confidence interval ranging from -0.006 to 0.010, suggests no significant change in the proportion of PPD values greater than 3 post-imputation. Similar trends were observed for sites 14V, 14MV, 14DL, 14L, and 14ML, with all p-values well above the significance threshold. These findings confirm that the H-D imputation preserved the proportions of PPD values greater than 3 across all sites.

PPD Statistics Before and After H-D Imputation by Site

Table II.12 presents a detailed comparison of PPD statistics before and after H-D imputation. For site DV, the Wilcoxon test result (W = 4146546, $p = 0.973$) and the F-test for variance (F = 0.023, $p = 0.880$) indicate no significant differences in median and variance values, suggesting that the imputation did not alter the original data distribution. Similar results were observed for sites V, MV, DL, L, and ML, where both the Wilcoxon test and the F-test indicated no significant differences in median and variance values. These findings confirm that the H-D imputation method effectively maintained the original data characteristics across all assessed sites for Tooth 14.

Kernel Density Estimates of Imputed and Original Values

Figure II.4 shows the kernel density plots of imputed and original PPD values for Tooth 14 across different sites. The density plots exhibit a high degree of overlap between the original and imputed distributions, further corroborating the statistical tests' results. These visual representations reinforce the effectiveness and reliability of the H-D imputation process in preserving the original data distribution.

In summary, the results from all tables indicate that the H-D imputation method successfully maintained the integrity and distribution of the original PPD values for the Upper Right First Premolar (Tooth 14), with no significant differences observed across various statistical measures and sites.

Upper Right Second Premolar - Tooth 15

This section presents an in-depth analysis of the data quality assessment following Hot Deck (H-D) Imputation for the Upper Right Second Premolar (Tooth 15). The evaluation is structured into three main subsections: comparison of original versus imputed PPD values by site, comparison of PPD values greater than 3, and a comprehensive assessment of PPD statistics before and after imputation.

Original vs. H-D Imputed PPD Values by Site

Table II.13 provides a detailed comparison of original and imputed PPD values across different sites for Tooth 15. The analysis indicates no significant differences across most sites and categories. For site DV, the p-values range from 0.611 to 0.849, suggesting no substantial deviations between original and imputed values. Similarly, site V showed consistent results, with the highest p-value being 0.9188 for $k = 3$, indicating effective imputation.

For site MV, the results demonstrate no significant differences across all categories, with p-values indicating no significant deviation. Site DL also exhibited no significant differences, with the highest p-value at 0.913 for $k = 0$. Consistency was observed for site L, with no significant differences across all categories, and the highest p-value at 0.927 for $k = 4$.

Site ML presented no significant differences across most categories, indicating a minor deviation. Overall, the H-D imputation method effectively maintained the integrity of the original PPD values across all assessed sites for Tooth 15.

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.14 compares the proportions of PPD values greater than 3 between the original and imputed datasets. The results indicate no significant differences across all sites. For site 15DV, the p-value of 0.630, with a confidence interval ranging from -0.006 to 0.010, suggests no significant change in the proportion of PPD values greater than 3 post-imputation. Similar trends were observed for sites 15V, 15MV, 15DL, 15L, and 15ML, with all p-values well above the significance threshold. These findings confirm that the H-D imputation preserved the proportions of PPD values greater than 3 across all sites.

PPD Statistics Before and After H-D Imputation by Site

Table II.15 presents a detailed comparison of PPD statistics before and after H-D imputation. For site DV, the Wilcoxon test result (W = 4241975, $p = 0.575$) and the F-test for variance (F = 0.164, $p = 0.686$) indicate no significant differences in median and variance values, suggesting that the imputation did not alter the original data distribution. Similar results were observed for sites V, MV, DL, L, and ML, where both the Wilcoxon test and the F-test indicated no significant differences in median and variance values. These findings confirm that the H-D imputation method effectively maintained the original data characteristics across all assessed sites for Tooth 15.

Kernel Density Estimates of Imputed and Original Values

Figure II.5 shows the kernel density plots of imputed and original PPD values for Tooth 15 across different sites. The density plots exhibit a high degree of overlap between the original and imputed distributions, further corroborating the statistical tests' results. These visual representations reinforce the effectiveness and reliability of the H-D imputation process in preserving the original data distribution.

In summary, the results from all tables indicate that the H-D imputation method successfully maintained the integrity and distribution of the original PPD values for the Upper Right Second Premolar (Tooth 15), with no significant differences observed across various statistical measures and sites.

Upper Right First Molar - Tooth 16

Original vs. H-D Imputed PPD Values by Site

Table II.16 presents the comparison between original and H-D imputed PPD values for the Upper Right First Molar (Tooth 16) across different sites. Analysis reveals no significant differences for most sites, with p-values generally above the significance threshold. For site DV, the p-values ranged from 0.3837 to 0.8508, indicating no substantial deviations between the original and imputed values. Notably, site DL exhibited a significant difference at $k = 4$ with a p-value of 0.0486 and a confidence interval indicating a significant discrepancy. Site MV and V showed consistent results, with no significant differences observed, as indicated by p-values well above 0.05. Overall, H-D imputation maintained the integrity of the original PPD values across most sites for Tooth 16, with minimal significant deviations.

Original vs. H-D Imputed PPD > 3 Proportions Values by Site

Table II.17 compares the proportions of PPD values greater than 3 between the original and imputed datasets. The results show no significant differences across most sites. Site 16DL was the exception, showing a significant difference with a p-value of 0.0294 and a confidence interval indicating a significant increase in the proportion of PPD values greater than 3 after imputation. Other sites, including 16DV, 16V, 16MV, 16L, and 16ML, did not exhibit significant differences, suggesting that the H-D imputation preserved the original proportions of PPD values greater than 3.

PPD Statistics Before and After H-D Imputation by Site

Table II.18 provides a comparison of PPD statistics before and after H-D imputation. For site DV, the Wilcoxon test (W = 3976425, $p = 0.580$) and the F-test for variance (F = 1.885, $p = 0.170$) indicate no significant differences in median and variance values, suggesting that imputation did not alter the original data distribution. Similar results were observed for sites V, MV, DL, L, and ML, where both the Wilcoxon test and the F-test indicated no significant differences in median and variance values. Site DL, however, showed a significant difference in variance (F = 4.334, $p = 0.037$), indicating a minor but notable change in data variability post-imputation.

Kernel Density Estimates of Imputed and Original Values

Figure II.6 illustrates the kernel density plots of imputed and original PPD values for Tooth 16 across different sites. The density plots exhibit a high degree of overlap between the original and imputed distributions, supporting the statistical findings. This visual confirmation further substantiates that the H-D imputation method effectively maintained the original data distribution for the Upper Right First Molar (Tooth 16).

Upper Right Second Molar - Tooth 17

Original vs. H-D Imputed PPD Values by Site

Table II.19 compares the original and H-D imputed PPD values for the Upper Right Second Molar (Tooth 17). The analysis indicates no significant differences for most sites. For site DV, p-values ranged from 0.570 to 0.930, suggesting no significant discrepancies between the original and imputed values. Notably, site DL exhibited a significant difference at $k = 1$ with a p-value of 0.030 and a confidence interval

indicating a significant discrepancy. For other sites such as V, MV, L, and ML, the results showed no significant differences, with p-values well above 0.05, confirming that H-D imputation preserved the original PPD values across most sites for Tooth 17.

Original vs. H-D Imputed PPD > 3 Proportions Values by Site

Table II.20 presents the comparison of PPD values greater than 3 between original and imputed datasets. The results indicate no significant differences across all sites. For site 17DV, the p-value of 0.865, with a confidence interval ranging from -0.008 to 0.010, suggests no significant change in the proportion of PPD values greater than 3 post-imputation. Similar trends were observed for sites 17V, 17MV, 17DL, 17L, and 17ML, with all p-values well above the significance threshold, confirming that H-D imputation preserved the proportions of PPD values greater than 3 across all sites.

PPD Statistics Before and After H-D Imputation by Site

Table II.21 shows the comparison of PPD statistics before and after H-D imputation for Tooth 17. For site DV, the Wilcoxon test result (W = 3962640, $p = 0.766$) and the F-test for variance (F = 0.0728, $p = 0.787$) indicate no significant differences in median and variance values, suggesting that the imputation did not alter the original data distribution. Similar results were observed for sites V, MV, L, and ML, where both the Wilcoxon test and the F-test indicated no significant differences in median and variance values. Site DL showed a significant difference in median (W = 4118499, $p = 0.030$), indicating a minor change in data distribution post-imputation.

Kernel Density Estimates of Imputed and Original Values

Figure II.7 depicts the kernel density plots of imputed and original PPD values for Tooth 17 across different sites. The density plots show a high degree of overlap between the original and imputed distributions, supporting the statistical findings. This visual confirmation further demonstrates that the H-D imputation method effectively maintained the original data distribution for the Upper Right Second Molar (Tooth 17).

In summary, the analysis for both Tooth 16 and Tooth 17 indicates that the H-D imputation method successfully preserved the integrity and distribution of the original PPD values, with minimal significant differences observed across various statistical measures and sites.

Upper Left Central Incisor - Tooth 21

Original vs. H-D Imputed PPD Values by Site

Table II.22 presents the comparison between original and H-D imputed PPD values for the Upper Left Central Incisor (Tooth 21) across different sites. For site DV, the analysis indicates no significant differences between the original and imputed values, with p-values ranging from 0.516 to 0.935, suggesting that the H-D imputation effectively preserved the original data distribution. Similarly, for site V, the p-values varied from 0.150 to 0.913, indicating no significant discrepancies. Sites MV, DL, L, and ML also showed no significant differences, with p-values well above the 0.05 threshold. The results demonstrate that the H-D imputation method maintained the integrity of the original PPD values across all evaluated sites for Tooth 21.

Original vs. H-D Imputed PPD > 3 Proportions Values by Site

Table II.23 compares the proportions of PPD values greater than 3 between the original and imputed datasets for Tooth 21. The analysis reveals no significant differences across all sites, with p-values ranging from 0.549 to 0.828. For instance, site 21DV had a p-value of 0.632, indicating no significant change in the proportion of PPD values greater than 3 after imputation. Sites such as 21V, 21MV, 21DL, 21L, and 21ML showed similar trends, confirming that the H-D imputation preserved the original proportions of PPD values greater than 3 across all sites.

PPD Statistics Before and After H-D Imputation by Site

Table II.24 provides a comparison of PPD statistics before and after H-D imputation for Tooth 21. For site DV, the Wilcoxon test result (W = 4690520, $p = 0.770$) and the F-test for variance (F = 0.2579, $p = 0.612$) indicate no significant differences in median and variance values, suggesting that imputation did not alter the original data distribution. Similar results were observed for sites V, MV, DL, L, and ML, where both the Wilcoxon test and the F-test indicated no significant differences in median and variance values. This consistency across different sites highlights the robustness of the H-D imputation method in maintaining the original data characteristics for Tooth 21.

Kernel Density Plot of Imputed and Original Values

Figure II.8 illustrates the kernel density plots of imputed and original PPD values for Tooth 21 across different sites. The density plots exhibit a high degree of overlap between the original and imputed distributions, corroborating the statistical findings. This visual confirmation further supports the conclusion that the H-D imputation method effectively preserved the original data distribution for the Upper Left Central Incisor (Tooth 21).

Overall, the analysis for Tooth 21 indicates that the H-D imputation method successfully maintained the integrity and distribution of the original PPD values, with minimal significant differences observed across various statistical measures and sites.

Upper Left Lateral Incisor - Tooth 22

Original vs. H-D Imputed PPD Values by Site

Table II.25 presents the comparison of original versus H-D imputed PPD values for the Upper Left Lateral Incisor (Tooth 22) by site. For site DV, the p-values ranged from 0.199 to 0.972, indicating no significant differences between the original and imputed values. Site V similarly showed no significant differences, with p-values ranging from 0.382 to 0.903. For site MV, the p-values varied from 0.426 to 0.983, suggesting that the imputed values closely matched the original data. Sites DL, L, and ML also demonstrated non-significant differences between the original and imputed values, with all p-values well above the 0.05 threshold. These findings suggest that the H-D imputation method preserved the integrity of the original PPD values for Tooth 22 across all sites.

Original vs. H-D Imputed PPD $>$ 3 Proportions Values by Site

The comparison of PPD values greater than 3 before and after H-D imputation is detailed in Table II.26. The p-values for all sites, including 22DV, 22V, 22MV, 22DL, 22L, and 22ML, were above the significance threshold, ranging from 0.526 to 0.908. For example, site 22DV had a p-value of 0.526, indicating no significant change in the proportion of PPD values greater than 3 after imputation. This consistency across various sites indicates that the imputation process did not significantly alter the proportion of PPD values greater than 3, thus maintaining the original data's distribution integrity.

PPD Statistics Before and After H-D Imputation by Site

Table II.27 compares PPD statistics before and after H-D imputation for Tooth 22. For site DV, the Wilcoxon test result (W = 4681564, $p = 0.393$) and the F-test for variance (F = 1.3269, $p = 0.249$) indicate no significant differences in the median and variance values. Similarly, for site V, the Wilcoxon test and F-test results (W = 4705147, $p = 0.283$; F = 0.0249, $p = 0.875$) show non-significant differences, suggesting the imputation preserved the original data characteristics. Sites MV, DL, L, and ML followed the same trend, with both the Wilcoxon test and F-test indicating no significant changes in the statistical properties of the data post-imputation. These results highlight the efficacy of the H-D imputation method in maintaining the original data distribution for Tooth 22.

Kernel Density Plot of Imputed and Original Values

Figure II.9 shows the kernel density plots of imputed and original PPD values for Tooth 22 across different sites. The density plots reveal a high degree of overlap between the original and imputed distributions, providing a visual confirmation of the statistical findings. This overlap further supports the conclusion that the H-D imputation method effectively preserved the original data distribution for the Upper Left Lateral Incisor (Tooth 22).

In summary, the analysis of Tooth 22 demonstrates that the H-D imputation method successfully maintained the original PPD values' integrity and distribution, with no significant differences observed across various statistical measures and sites.

Upper Left Canine - Tooth 23

Original vs. H-D Imputed PPD Values by Site

Table II.28 presents the comparison of original and H-D imputed PPD values for the Upper Left Canine (Tooth 23) by site. The results show no significant differences between the original and imputed values across all sites. For instance, site DV exhibited p-values ranging from 0.591 to 0.934, indicating that the imputed values closely matched the original data. Similar patterns were observed for sites V, MV, DL, L, and ML, with all p-values well above the significance threshold. This consistency suggests that the H-D imputation method preserved the original data distribution for PPD values in Tooth 23.

Original vs. H-D Imputed PPD > 3 Proportions Values by Site

Table II.29 compares the proportions of PPD values greater than 3 before and after H-D imputation by site. The p-values for all sites, including 23DV, 23V, 23MV, 23DL, 23L, and 23ML, were above 0.05, indicating no significant differences. For example, site 23DV had a p-value of 0.771, suggesting that the proportion of PPD values greater than 3 remained consistent after imputation. These results indicate that the H-D imputation method did not significantly alter the distribution of PPD values greater than 3, maintaining the integrity of the original data.

PPD Statistics Before and After H-D Imputation by Site

Table II.30 shows the comparison of PPD statistics before and after H-D imputation for Tooth 23. The Wilcoxon test results and F-tests for variance across all sites indicated no significant differences in median and variance values between the original and imputed datasets. For instance, site DV had a Wilcoxon test result of W = 4874719 and $p = 0.967$, and an F-test for variance with F = 0.0667 and $p = 0.796$. These findings were consistent across other sites, including V, MV, DL, L, and ML, confirming that the imputation method maintained the original statistical properties of the data.

Kernel Density Plot of Imputed and Original Values

The kernel density plots in Figure II.10 provide a visual comparison of the original and imputed PPD values for Tooth 23 by site. The plots demonstrate a high degree of overlap between the original and imputed data distributions, further supporting the statistical results. This visual evidence corroborates the conclusion that the H-D imputation method effectively preserved the original data distribution for the Upper Left Canine (Tooth 23).

In conclusion, the analysis of Tooth 23 demonstrates that the H-D imputation method successfully preserved the integrity and distribution of the original PPD values. No significant differences were observed across various statistical measures and sites, confirming the effectiveness of the imputation process.

Upper Left First Premolar - Tooth 24

Original vs. H-D Imputed PPD Values by Site

Table II.31 compares the original and H-D imputed PPD values for the Upper Left First Premolar (Tooth 24) by site. The p-values across all sites indicate no significant differences between the original and imputed values. For instance, at site DV, p-values ranged from 0.0711 to 0.9052, showing that the imputed values closely matched the original data. Similarly, for sites V, MV, DL, L, and ML, the p-values were consistently above the significance threshold, demonstrating the effectiveness of the H-D imputation method in preserving the original data distribution for PPD values in Tooth 24.

Original vs. H-D Imputed PPD $>$ 3 Proportions Values by Site

Table II.32 presents the comparison of proportions of PPD values greater than 3 before and after H-D imputation by site. The p-values for all sites were above 0.05, indicating no significant differences. For example, site D24DV had a p-value of 0.412, suggesting that the proportion of PPD values greater than 3 remained consistent after imputation. This consistency was observed across all other sites, including D24V, D24MV, D24DL, D24L, and D24ML. These results confirm that the H-D imputation method did not significantly alter the distribution of PPD values greater than 3, maintaining the integrity of the original data.

PPD Statistics Before and After H-D Imputation by Site

Table II.33 shows the comparison of PPD statistics before and after H-D imputation for Tooth 24. The Wilcoxon test results and F-tests for variance indicated no significant differences in median and variance values between the original and imputed datasets across all sites. For instance, at site DV, the Wilcoxon test result was $W = 4140321$ with $p = 0.996$, and the F-test for variance was $F = 0.300$ with $p = 0.584$. Similar results were observed for other sites, including V, MV, DL, L, and ML, with all p-values indicating non-significance. These findings confirm that the imputation method maintained the original statistical properties of the data for Tooth 24.

Kernel Density Estimates of Imputed and Original Values

The kernel density plots in Figure II.11 visually compare the original and imputed PPD values for Tooth 24 by site. The high degree of overlap between the original and

imputed data distributions further supports the statistical results. This visual evidence corroborates the conclusion that the H-D imputation method effectively preserved the original data distribution for the Upper Left First Pre Molar (Tooth 24).

In summary, the analysis of Tooth 24 demonstrates that the H-D imputation method successfully preserved the integrity and distribution of the original PPD values. The results across various statistical measures and sites confirm the effectiveness of the imputation process, with no significant differences observed between the original and imputed datasets.

Upper Left Second Premolar - Tooth 25

Original vs. H-D Imputed PPD Values by Site

Table II.34 compares the original and H-D imputed PPD values for the Upper Left Second Premolar (Tooth 25) across various sites. The p-values for most sites indicate no significant differences between the original and imputed values, confirming the effectiveness of the H-D imputation method. For example, at site DV, the p-value ranges from 0.527 to 0.970, indicating that the imputed values closely match the original data. Similar observations can be made for sites V, MV, DL, L, and ML. However, at site DL, a significant difference was observed at k = 8 (p-value = 0.033), suggesting a discrepancy in the imputed values at this specific site. This outlier should be further investigated to understand the underlying cause.

Original vs. H-D Imputed PPD $>$ 3 Proportions Values by Site

Table II.35 presents the comparison of the proportions of PPD values greater than 3 before and after H-D imputation by site. The p-values across all sites indicate no significant differences, suggesting that the imputation process preserved the proportion of PPD values greater than 3. For instance, at site DV, the p-value is 0.607, indicating that the proportion remains consistent post-imputation. This consistency is observed across other sites, including V, MV, DL, L, and ML, reinforcing the reliability of the imputation method in maintaining the integrity of the original data distribution.

PPD Statistics Before and After H-D Imputation by Site

Table II.36 compares the PPD statistics before and after H-D imputation for Tooth 25. The Wilcoxon test results and F-tests for variance indicate no significant differences

in median and variance values between the original and imputed datasets across most sites. For example, at site DV, the Wilcoxon test result is $W = 4166893$ with $p = 0.482$, and the F-test for variance is $F = 0.0439$ with $p = 0.834$. These results are consistent across sites MV, DL, L, and ML. However, at site V, a significant difference in the median values was observed with $W = 4342589$ and $p = 0.016$, indicating a potential discrepancy in the imputed data at this site. Further investigation is warranted to address this anomaly.

Kernel Density Estimates of Imputed and Original Values

The kernel density estimates in Figure II.12 provide a visual comparison of the original and imputed PPD values for Tooth 25 by site. The high degree of overlap between the original and imputed data distributions corroborates the statistical findings. This visual evidence supports the conclusion that the H-D imputation method effectively preserved the original data distribution for the upper left second premolar (Tooth 25).

In summary, the analysis of Tooth 25 demonstrates that the H-D imputation method successfully maintained the integrity and distribution of the original PPD values across most sites. While minor discrepancies were noted at specific sites, the overall effectiveness of the imputation process is evident from the consistent statistical properties observed in the original and imputed datasets.

Upper Left First Molar - Tooth 26

Original vs. H-D Imputed PPD Values by Site

Table II.37 presents a comparison between the original and H-D imputed PPD values for the upper left first molar (Tooth 26) across various sites. The results show no significant differences in most sites, indicating the efficacy of the imputation method. For instance, site DV exhibits a p-value of 0.808 at k = 0, with the confidence interval suggesting a minor deviation between the original and imputed values. Similar trends are observed across sites V, MV, DL, L, and ML. Notably, at site DV k = 4, a relatively low p-value of 0.082 hints at a slight discrepancy, though it remains above the threshold for statistical significance. Overall, these results underscore the reliability of the H-D imputation process in preserving the original data characteristics.

Original vs. H-D Imputed PPD > 3 Proportions Values by Site

Table II.38 details the comparison of proportions of PPD values greater than 3, before and after H-D imputation, by site. The p-values across the sites indicate no significant differences, affirming the consistency of the imputation process. For instance, at site DV, the p-value is 0.111, indicating that the proportion of PPD values greater than 3 remains stable post-imputation. This consistency is observed across all sites, including V, MV, DL, L, and ML, thus validating the imputation method's ability to maintain the original data's distributional properties.

PPD Statistics Before and After H-D Imputation by Site

Table II.39 compares the PPD statistics before and after H-D imputation for Tooth 26. The Wilcoxon test and F-tests for variance reveal no significant differences in median and variance values between the original and imputed datasets across all sites. For example, at site DV, the Wilcoxon test results in $W = 3960621$ with $p = 0.145$, and the F-test for variance yields $F = 0.8468$ with $p = 0.358$, indicating that the imputed values closely match the original data. This pattern is consistent across other sites, such as V, MV, DL, L, and ML, reinforcing the robustness of the H-D imputation method.

Kernel Density Plot of Imputed and Original Values

The kernel density plots in Figure II.13 visually compare the original and imputed PPD values for Tooth 26 by site. The high degree of overlap between the original and imputed data distributions corroborates the statistical findings, further illustrating that the H-D imputation method effectively preserves the original data distribution for the upper left first molar (Tooth 26).

In summary, the analysis of Tooth 26 demonstrates that the H-D imputation method successfully maintains the integrity and distribution of the original PPD values across all examined sites. The consistent statistical properties and visual confirmation through kernel density plots provide robust evidence of the method's effectiveness.

Upper Left Second Molar - Tooth 27

Original vs. H-D Imputed PPD Values by Site

Table II.40 compares the original PPD values with those imputed using the Hot Deck method for the Upper Left Second Molar (Tooth 27) across various sites. The results indicate that the differences between the original and imputed values are not statistically significant in most cases, as evidenced by the p-values. For instance, at site DV, the p-value is 0.546 at k = 0, suggesting no significant deviation between the original and imputed values. Similar trends are observed across sites V, MV, DL, L, and ML, indicating the robustness of the imputation method. However, some variability is noted, such as at site MV, k = 2, with a relatively low p-value of 0.393, indicating slight discrepancies that remain statistically non-significant.

Original vs. H-D Imputed PPD $>$ 3 Proportions Values by Site

Table II.41 presents the comparison of proportions of PPD values greater than 3 before and after H-D imputation by site. The p-values indicate no significant differences across all sites, affirming the consistency of the imputation process. For example, at site DV, the p-value is 0.569, showing stability in the proportion of PPD values greater than 3 post-imputation. This consistency is observed across other sites, such as V, MV, DL, L, and ML, confirming the imputation method's ability to maintain the original data's distributional properties.

PPD Statistics Before and After H-D Imputation by Site

Table II.42 compares the PPD statistics before and after H-D imputation for Tooth 27. The Wilcoxon test and F-tests for variance reveal no significant differences in median and variance values between the original and imputed datasets across all sites. For instance, at site DV, the Wilcoxon test results in $W = 3949964$ with $p = 0.508$, and the F-test for variance yields $F = 0.155$ with $p = 0.694$, indicating that the imputed values closely match the original data. Similar patterns are observed at other sites such as V, MV, DL, L, and ML, reinforcing the robustness of the H-D imputation method.

Kernel Density Plot of Imputed and Original Values

The kernel density plots in Figure II.14 visually compare the original and imputed PPD values for Tooth 27 by site. The high degree of overlap between the original and

imputed data distributions corroborates the statistical findings, further illustrating that the H-D imputation method effectively preserves the original data distribution for the Upper Left Second Molar (Tooth 27).

In conclusion, the analysis of Tooth 27 demonstrates that the H-D imputation method successfully maintains the integrity and distribution of the original PPD values across all examined sites. The consistent statistical properties and visual confirmation through kernel density plots provide robust evidence of the method's effectiveness.

4.4 Modau

In this section of the results chapter, we present the findings related to the MoDau imputation method. We begin by presenting the synthetic data generated to test the imputation models in the first section. Following this, we characterise the Mother and Daughters models. Finally, we conclude with the imputation results and their validation.

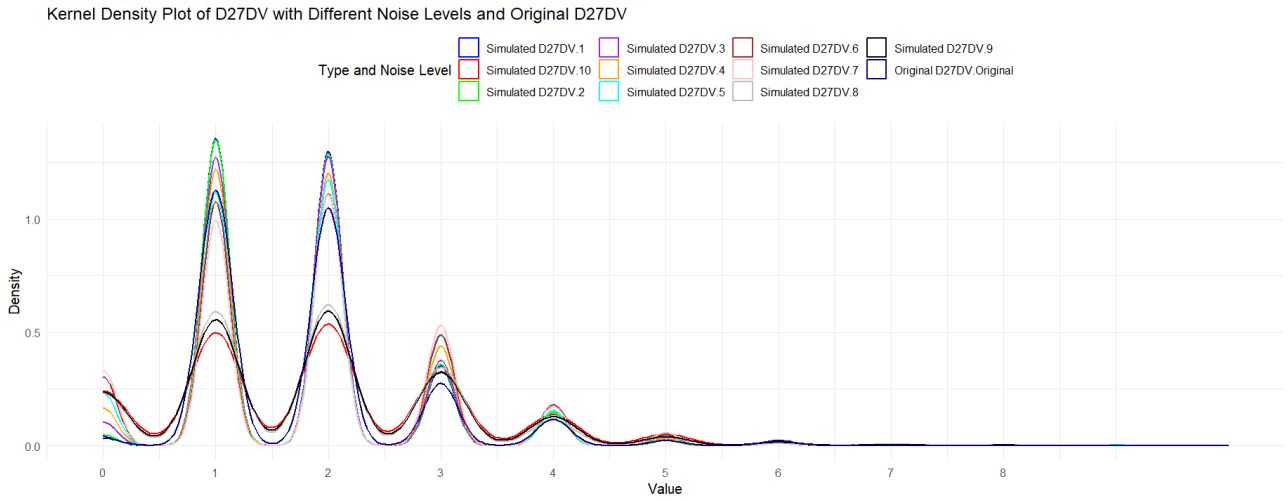
4.4.1 Simulated Data

Noise Levels

We generate ten data sets each one with a noise level from 0.1 to 10 and then compared them with KDE plots see Figure 4.5. the numerical and graphic to choose the noise level of 0.4 are presented bellow.

The figure 4.5 illustrates the variation in Kernel Density Estimates (KDE) as different levels of noise are added to the original distribution. This figure specifically refers to site 27DV, which was randomly selected to exemplify the effect of varying noise levels on the distribution of PPD. It can be observed that for noise levels 0.1, 0.2, and 0.4, the peaks at x=1 and x=2 are higher than in the original data, with the peak at PPD = 1 being higher than at PPD = 2, similar to the original data. This indicates a more similar distribution to the original data than at all other noise levels.

Figure 4.5



Kernel Density Plots by Noise level

The Table 4.19 presents the mean difference in densities and the corresponding 95% confidence intervals for noise levels ranging from 0.1 to 1.0. These results of the bootstrap test for KDE across the noise levels refer to site D27.

Notably, the small difference for the noise level 0.2 and the consistency of all other noise levels, with differences ranging from 0.11 to 0.15, are highlighted. Six out of ten differences are 0.14. It is also noteworthy that the lower limit of the confidence interval consistently departs from 0, with the lower limits for 0.1 and 0.3 being lower than any other and close to zero, suggesting a small difference between densities. Unlike the lower limit, the upper limit does not show a clear trend.

Table 4.19: Bootstrap Test for KDE Difference Across Noise Levels

Noise Level	**Mean Difference in Densities**	**95% CI**
0.1	0.11	[0.002, 0.135]
0.2	0.03	[0.002, 0.186]
0.3	0.14	[0.083, 0.155]
0.4	0.14	[0.082, 0.158]
0.5	0.14	[0.085, 0.149]
0.6	0.15	[0.082, 0.174]
0.7	0.14	[0.084, 0.172]
0.8	0.13	[0.104, 0.161]
0.9	0.14	[0.130, 0.142]
1.0	0.14	[0.131, 0.152]

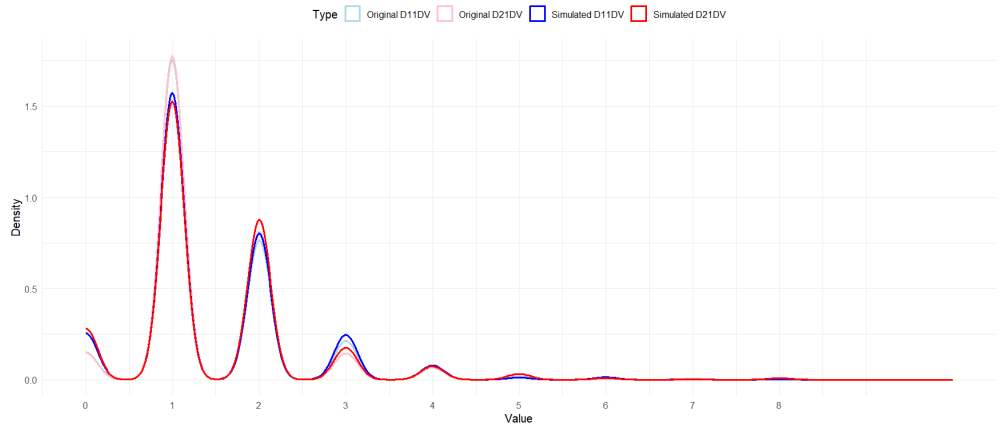
Notably, (Table 4.19) the small difference for the noise level 0.2 and the consistency of all other noise levels, with differences ranging from 0.11 to 0.15, are highlighted. Six out of ten differences are 0.14. It is also noteworthy that the lower limit of the confidence interval consistently departs from 0, with the lower limits for 0.1 and 0.3 being lower than any other and close to zero, suggesting a small difference between densities. Unlike the lower limit, the upper limit does not show a clear trend.

Comparing Simulated with Original site by Site

11, 21 Original KDE *versus* Simulated KDE

Site DV

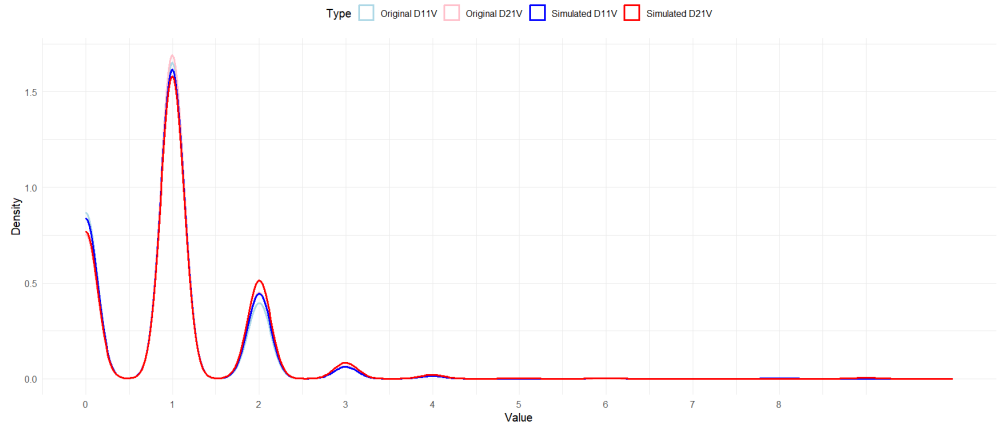
Figure 4.6



Kernel Density Plot of Simulated and Original 11DV, 21DV (noise level: 0.4)

Site V

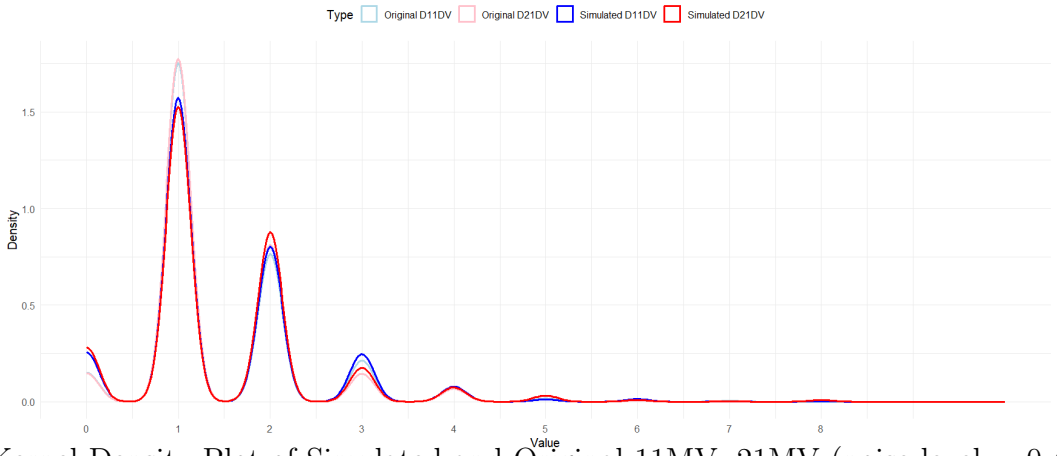
Figure 4.7



Kernel Density Plot of Simulated 11V, 21V (noise level: 0.4) and Original 11V, 21V

Site MV

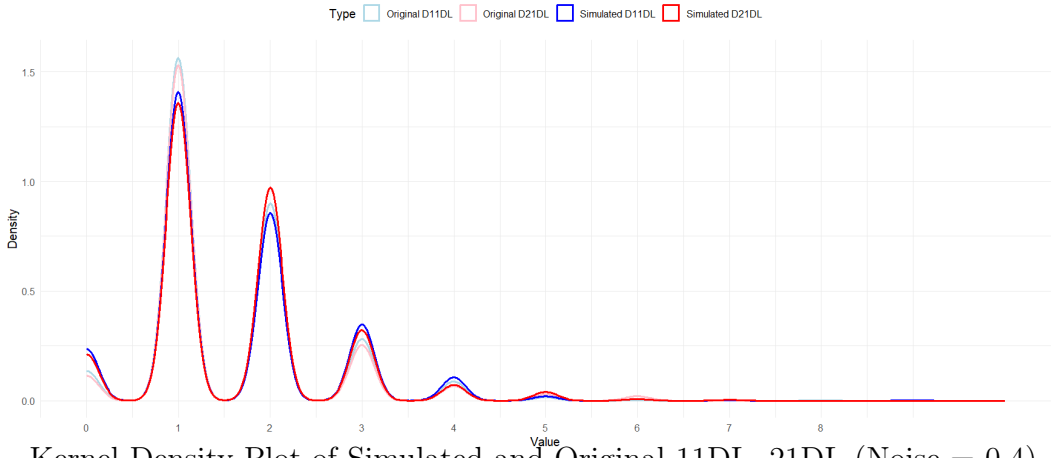
Figure 4.8



Kernel Density Plot of Simulated and Original 11MV, 21MV (noise level = 0.4)

Site DL

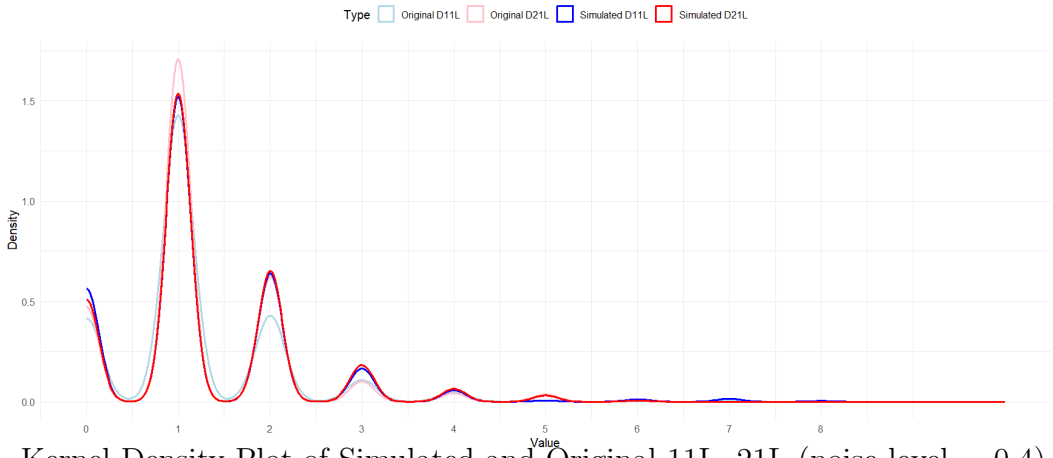
Figure 4.9



Kernel Density Plot of Simulated and Original 11DL, 21DL (Noise = 0.4)

Site L

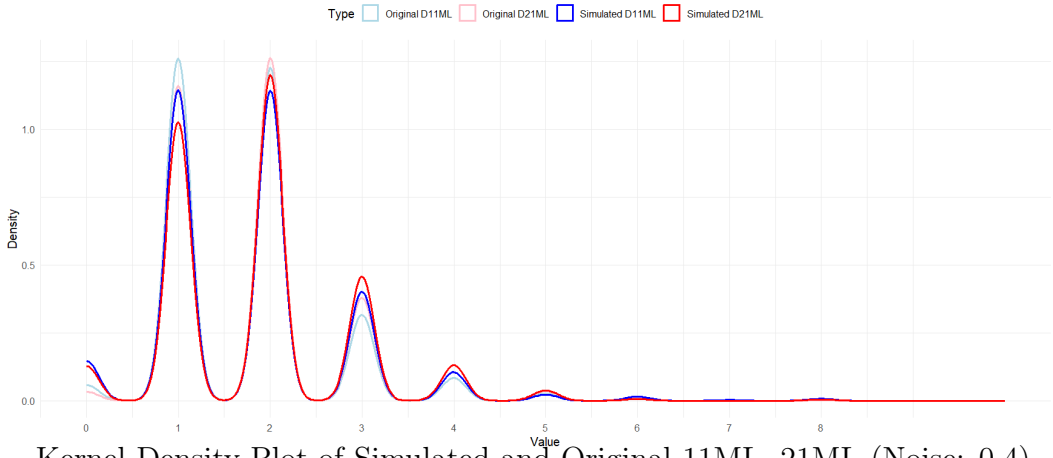
Figure 4.10



Kernel Density Plot of Simulated and Original 11L, 21L (noise level = 0.4)

Site ML

Figure 4.11



Kernel Density Plot of Simulated and Original 11ML, 21ML (Noise: 0.4)

4.4.1.1 Overall analysis of kernels depicting simulated (noise = 0.4) and original

The plots in analysis share the main significant features As observed, the peaks of the KDE plots align along the X-axis, suggesting a significant level of similarity between the original and simulated datasets. The plots demonstrate that the simulated data closely follows the density curves of the original data, indicating that the simulation effectively captures the essential characteristics of the original data. The observed deviations are due to the introduced noise level, which introduces variability as expected but does not drastically alter the overall distribution pattern.

4.5 Mother-Daughter Models Imputation

The final set of parameters values for the Mother models was the following: `objective = "reg:squarederror"`, `max_depth = 3`, `eta = 0.1`, `gamma = 0`, `colsample_bytree = 1.0`, `min_child_weight = 3`, `subsample = 0.7`, and the `nrounds =1000`. The results from the training process are the following.

4.5.1 Tooth 21 - MoDau Imputation

Mother Models M21

The characteristics of the M21 mother models (Table IV.1) revealed differences in model size, iterations, and RMSE values across various sites. For instance, the M21L model had the largest size at 366.00 Kb and required 344 iterations, whereas the M21V model was the smallest at 146.00 Kb with 134 iterations.

Table 4.20: Distinctive Characteristics of M21 Mother Models by Site

Metric	M21DV	M21V	M21MV	M21DL	M21L	M21ML
Mother Model Size (Kb)	274.80	146.00	211.80	261.00	366.00	258.80
N.Iter.	261	134	198	253	344	246
Init. Train. RMSE	1.193	0.868	1.152	1.305	1.062	1.476
Final Train. RMSE	1.82e-02	5.94e-02	3.39e-02	2.92e-02	1.60e-02	3.67e-02
Features	γSM.DV 11DV	γSM.V 11V	γSM.MV 11MV	γSM.DL 11DL	γSM.L 11L	γSM.ML 11ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table 4.21: Performance Metrics for the M21 Mother Models by Site

Metric	M21DV	M21V	M21MV	M21DL	M21L	M21ML
RMSE	0.018	0.059	0.034	0.029	0.016	0.037
MAE	0.003	0.007	0.004	0.005	0.002	0.005
MSE	3.30e-04	3.52e-03	1.15e-03	8.50e-04	2.60e-04	1.35e-03
R^2	99.96%	99.48%	99.84%	99.90%	99.97%	99.84%

Table 4.22: Mother Models M21 Features Importance Metrics by Site

Model	Feature	Gain	Cover	Frequency
M21DV	γSM.DV	0.612	0.509	0.615
	Original 11DV	0.388	0.491	0.385
M21V	γSM.V	0.599	0.538	0.651
	Original 11V	0.401	0.462	0.349
M21MV	γSM.MV	0.558	0.474	0.639
	Original 11MV	0.442	0.526	0.361
M21DL	γSM.DL	0.577	0.485	0.592
	Original 11DL	0.423	0.515	0.408
M21L	γSM.L	0.558	0.612	0.690
	Original 11L	0.442	0.388	0.310
M21ML	γSM.ML	0.525	0.512	0.638
	Original 11ML	0.475	0.488	0.362

Performance metrics for the M21 mother models, as shown in Table IV.2, included RMSE, MAE, MSE, and R^2 values. The models exhibited excellent performance with RMSE values ranging from 0.016 (M21L) to 0.059 (M21V). MAE values were consistently low across all sites, indicating minimal absolute errors in predictions. MSE values were also low, with the highest being 3.52e-03 for M21V. The R^2 values were exceptionally high, exceeding 99.48% for all models, indicating a very strong fit to the data.

Feature importance metrics, presented in Table IV.3, highlighted the significance of the directional symmetry measure (γSM) and original NHANES 2011/2012 sites. The gain, cover, and frequency metrics were used to quantify feature importance. For example, the γSM.L feature in the M21L model had the highest cover (0.612) and frequency (0.690), indicating its substantial contribution to the model's predictive power. Similarly, the γSM.DV feature in the M21DV model exhibited the highest gain (0.612), emphasizing its critical role in the model's accuracy.

Table 4.24: Performance Metrics by Site for D21

Metric	D21DV	D21V	D21MV	D21DL	D21L	D21ML
RMSE	0.181	0.142	0.177	0.162	0.161	0.156
MAE	0.143	0.112	0.140	0.128	0.128	0.123
MSE	0.033	0.020	0.031	0.026	0.026	0.024
R^2	96.63%	97.44%	96.58%	97.47%	97.37%	97.49%

Daughter Models D21

The daughters models were fitted to the predictors: upper right side PPD values from the simulated data, Mother Models predictions and to five predictions of the New Mother Models.

The characteristics of the D21 daughter models are summarized in Table IV.4. Model sizes ranged from 31.2 Mb to 32.5 Mb across the different sites, and each model required 30000 iterations for training. Initial training RMSE values varied, with D21V showing the lowest at 0.953 and D21ML the highest at 1.590. Final training RMSE values were substantially lower, ranging from 0.142 to 0.181, reflecting the models' improved accuracy after training.

Table 4.23: Distinctive Characteristics of the D21 by Site

Metric	D21DV	D21V	D21MV	D21DL	D21L	21ML
Model Size (Mb)	31.2	32.1	31.2	32.4	32.1	32.5
N.Iter.	30000	30000	30000	30000	30000	30000
Init. Train. RMSE	1.305	0.953	1.287	1.430	1.180	1.590
Final Train. RMSE	0.181	0.142	0.177	0.162	0.161	0.156

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Performance metrics for the D21 daughter models are presented in Table IV.5. The RMSE values ranged from 0.142 (D21V) to 0.181 (D21DV), demonstrating a high level of accuracy across all models. MAE values were consistently low, with the smallest errors observed in D21V (0.112) and the largest in D21DV (0.143). MSE values were also low, with D21V showing the lowest at 0.020 and D21DV the highest at 0.033. The R^2 values exceeded 96.58

Feature importance metrics, as shown in Table IV.6, highlight the contributions of different predictors to each model. For D21DV, the simulated 11DV feature had the highest cover (0.601) and frequency (0.582), indicating its significant role in the

Table 4.25: Feature Importance Metrics of D21 by Site

Model	Feature	Gain	Cover	Frequency
D21DV	Mother Predictions	0.389	0.273	0.324
	New Mother Predictions	0.332	0.127	0.094
	Simulated 11DV	0.279	0.601	0.582
D21V	Mother Predictions	0.442	0.301	0.300
	New Mother Predictions	0.274	0.088	0.072
	Simulated 11V	0.283	0.611	0.628
D21MV	Mother Predictions	0.380	0.304	0.346
	New Mother Predictions	0.335	0.088	0.066
	Simulated 11MV	0.285	0.608	0.587
D21DL	Mother Predictions	0.381	0.290	0.309
	New Mother Predictions	0.344	0.115	0.088
	Simulated 11DL	0.275	0.596	0.603
D21L	Mother Predictions	0.432	0.261	0.286
	New Mother Predictions	0.290	0.135	0.100
	Simulated 11L	0.278	0.604	0.614
D21ML	Mother Predictions	0.442	0.302	0.322
	New Mother Predictions	0.280	0.093	0.073
	Simulated 11ML	0.278	0.605	0.605

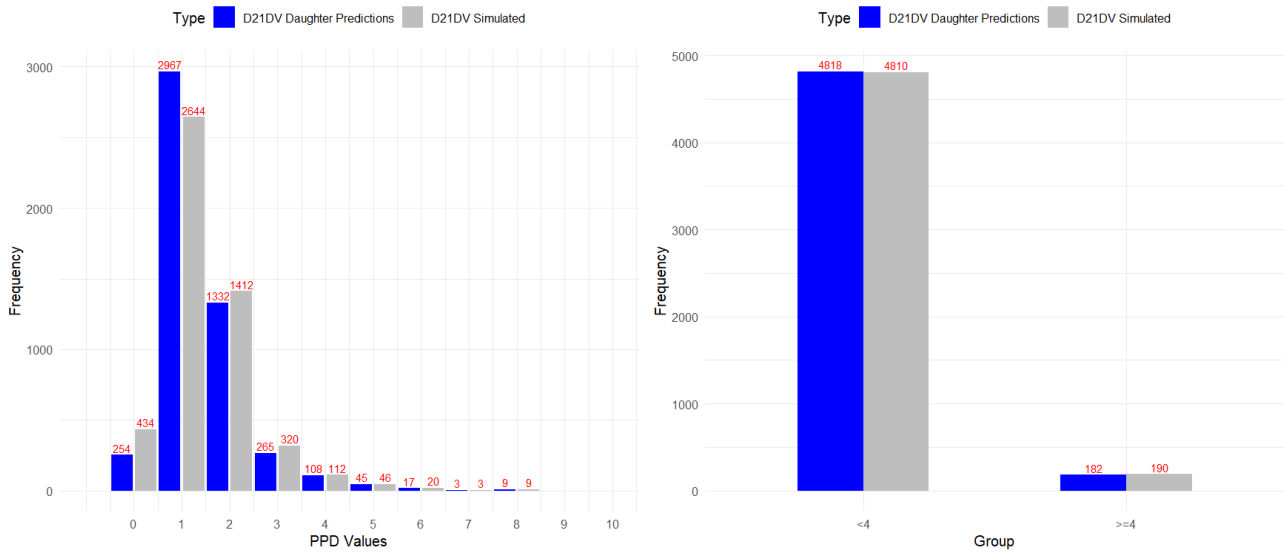
model. The mother predictions contributed the most gain (0.442) in the D21ML model, emphasizing their importance in prediction accuracy. New mother predictions were less influential, with lower cover and frequency values across all models.

Daughter models predictions by site

Site DV

The results for site DV, presented in Table IV.7, indicate that for PPD values of 0 and 1, the chi-squared statistics were 47.093 and 18.594, respectively, with p-values significantly less than 0.05, suggesting a substantial difference between the predicted and simulated counts. For PPD values of 2, 3, 4, 5, 6, and 7, the p-values were greater than 0.05, indicating no significant difference. Among these, 25% of the comparisons had p-values less than 0.05, 62.5% had p-values greater than 0.05 but less than 1, and 12.5% had p-values equal to 1. The aggregate PPD values (PPD $\geq$ 4) showed no significant difference, with a chi-squared statistic of 0.137 and a p-value of 0.711.

Figure 4.12



Histograms of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Table 4.26: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
21DV	0	254	434	47.093	**6.77e-12**
	1	2967	2644	18.594	**1.62e-05**
	2	1332	1412	2.332	0.127
	3	265	320	5.171	0.023
	4	108	112	0.073	0.787
	5	45	46	0.011	0.917
	6	17	20	0.243	0.622
	7	3	3	0	1
PPD ≥ 4		182	190	0.137	0.711

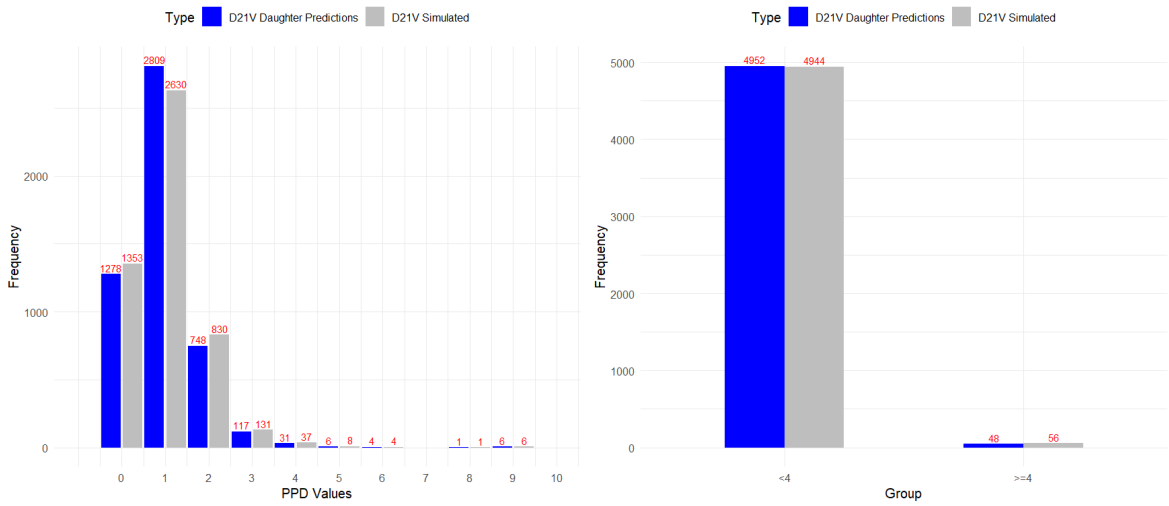
Site V

For site V, as shown in Table IV.8, significant differences were observed for PPD values of 1 and 2, with chi-squared statistics of 5.891 and 4.261, respectively, and corresponding p-values less than 0.05. For other PPD values, no significant differences were observed. Here, 22.2% of comparisons had p-values less than 0.05, 44.4% had p-values between 0.05 and 1, and 33.3% had p-values equal to 1. The aggregate PPD values (PPD ≥ 4) also showed no significant difference.

Table 4.27: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
21V	0	1278	1353	2.138	0.144
	1	2809	2630	5.891	**0.015**
	2	748	830	4.261	**0.039**
	3	117	131	0.790	0.374
	4	31	37	0.529	0.467
	5	6	8	0.286	0.593
	6	4	4	0	1
	8	1	1	0	1
	9	6	6	0	1
PPD ≥ 4		48	56	0.476	0.490

Figure 4.13



Histograms of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm

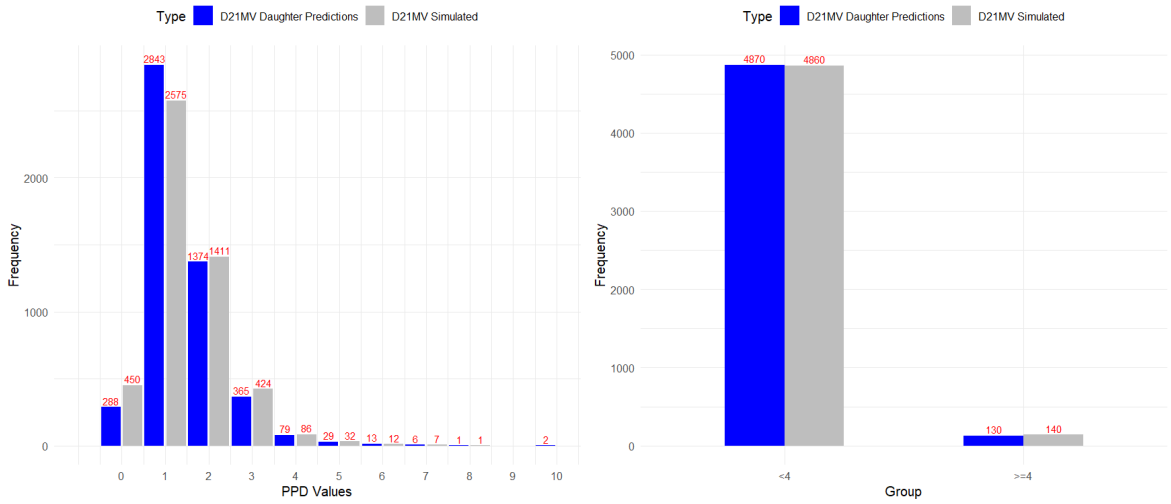
Table 4.28: Chi-squared Test Results for Comparison of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
21MV	0	288	450	35.561	**2.47e-09**
	1	2843	2575	13.257	**2.70e-04**
	2	1374	1411	0.492	0.483
	3	365	424	4.412	**0.036**
	4	79	86	0.297	0.586
	5	29	32	0.148	0.701
	6	13	12	0.040	0.841
	7	6	7	0.077	0.782
	8	1	1	0	1
	10	2	2	0	1
PPD ≥ 4		130	140	0.308	0.579

Site MV

The results for site MV in Table IV.9 reveal significant differences for PPD values of 0, 1, and 3, with respective chi-squared statistics of 35.561, 13.257, and 4.412, all with p-values less than 0.05. The other PPD values and the aggregate PPD values (PPD ≥ 4) showed no significant differences. This site had 30% of comparisons with p-values less than 0.05, 50% with p-values between 0.05 and 1, and 20% with p-values equal to 1.

Figure 4.14

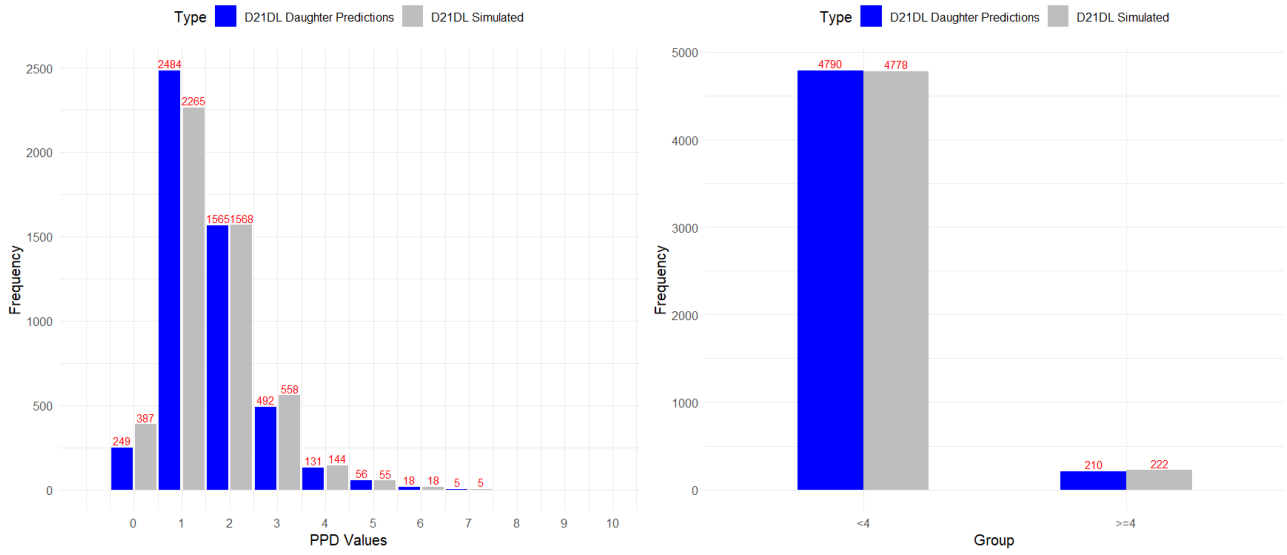


Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Table 4.29: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
21DL	0	249	387	29.943	**4.45e-08**
	1	2484	2265	10.099	**1.48e-03**
	2	1565	1568	0.003	0.957
	3	492	558	4.149	**0.042**
	4	131	144	0.615	0.433
	5	56	55	0.009	0.924
	6	18	18	0	1
	7	5	5	0	1
PPD ≥ 4		210	222	0.293	0.589

Figure 4.15



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Site DL

For site DL, as illustrated in Table IV.10, significant differences were found for PPD values of 0, 1, and 3, with chi-squared statistics of 29.943, 10.099, and 4.149, respectively, and p-values less than 0.05. Other PPD values and the aggregate PPD values (PPD ≥ 4) indicated no significant differences. This site had 37.5% of comparisons with p-values less than 0.05, 50% with p-values between 0.05 and 1, and 12.5% with p-values equal to 1.

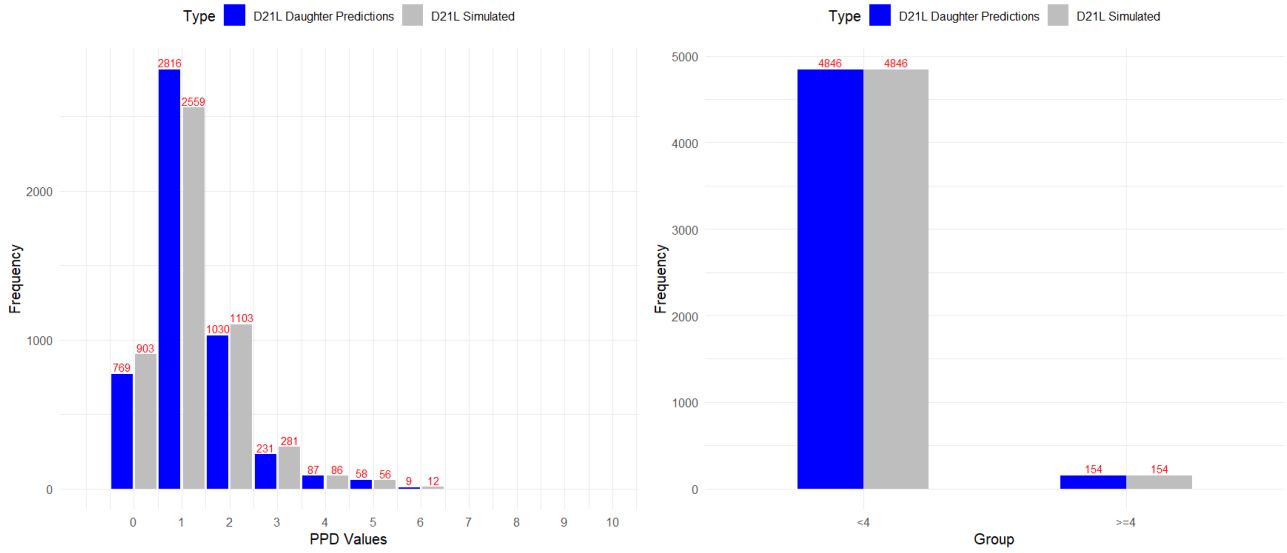
Table 4.30: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
	0	769	903	10.739	**1.05e-03**
	1	2816	2559	12.288	**4.60e-04**
	2	1030	1103	2.498	0.114
21L	3	231	281	4.883	**0.027**
	4	87	86	0.006	0.939
	5	58	56	0.035	0.851
	6	9	12	0.429	0.513
PPD ≥ 4		154	154	0.000	1.000

Site L

The analysis for site L, shown in Table IV.11, demonstrates significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 10.739, 12.288, and 4.883, respectively, and p-values less than 0.05. No significant differences were observed for other PPD values and the aggregate PPD values (PPD ≥ 4) with a chi-squared statistic of 0.00 and a p-value of 1. Here, 42.9% of comparisons had p-values less than 0.05, 57.1% had p-values between 0.05.

Figure 4.16



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

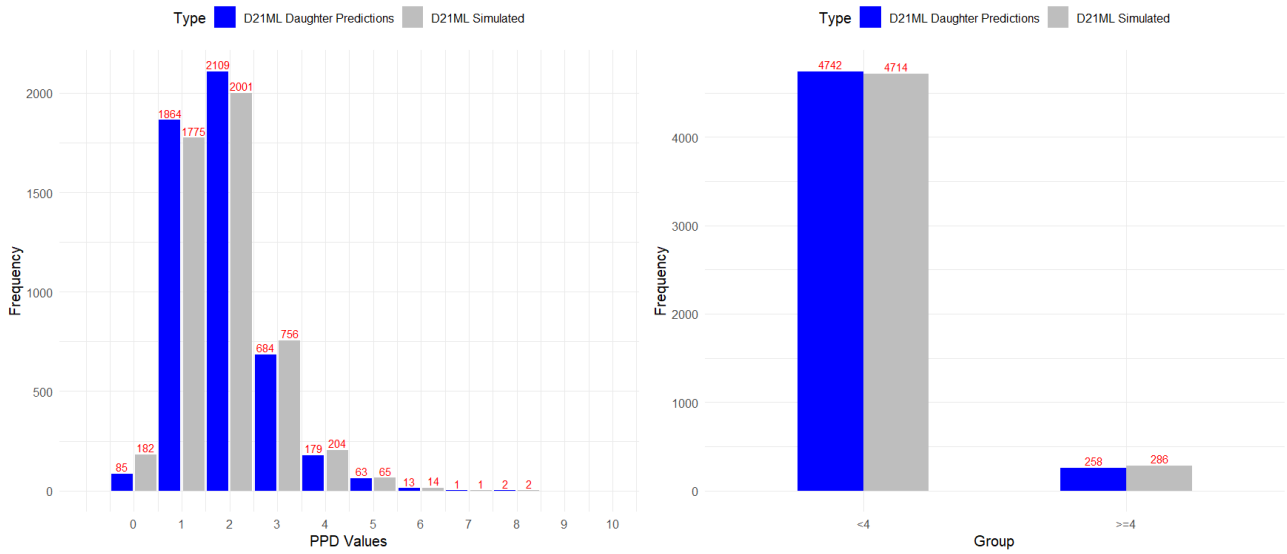
Site ML

For site ML, presented in Table IV.12 a significant difference was noted for PPD value 0 with a chi-squared statistic of 35.240 and a p-value of 2.92e-09. No significant differences were found for other PPD values or the aggregate PPD values (PPD ≥ 4). This site had 11.1% of comparisons with p-values less than 0.05, 66.7% with p-values between 0.05 and 1, and 22.2% with p-values equal to 1.

Table 4.31: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
	0	85	182	35.240	**2.92e-09**
	1	1864	1775	2.177	0.140
	2	2109	2001	2.838	0.092
	3	684	756	3.600	0.058
21ML	4	179	204	1.632	0.201
	5	63	65	0.031	0.860
	6	13	14	0.037	0.847
	7	1	1	0	1
	8	2	2	0	1
PPD ≥ 4		258 286	1.417	0.234	

Figure 4.17



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

Aggregate

Simulated versus predicted with Density estimates

Numerical assessment Table 4.32 summarizes the Kernel Density Estimates (KDE) differences and Kolmogorov-Smirnov (KS) test results for various datasets. The kernel function used was cosine for site 21V and biweight for the other five sites. The bootstrap differences between the KDE of the simulated and predicted data are on the order of a thousandth, from 2.77e-04 to 5.00e-03. Such small differences indicate that the predicted and simulated distributions are very similar in shape and spread. Although these differences are small, they are statistically significant with p-values from 2.29e-13 to 2.96e-04.

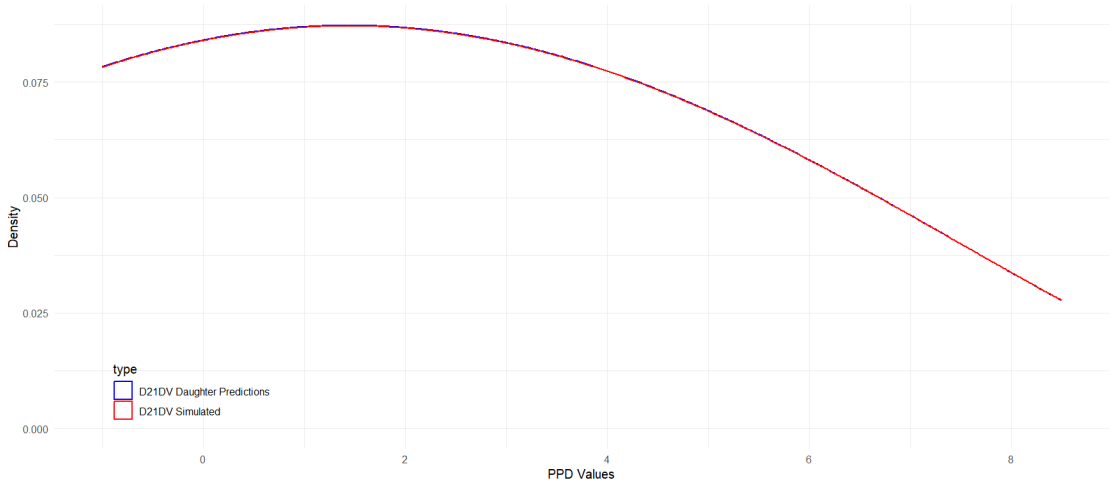
Table 4.32: KDE Differences and Kolmogorov-Smirnov Test

Site	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
21DV	Biweight	2	2	2.77e-04	[3.25e-05, 2.47e-03]	0.076	7.77e-13
21V	Cosine	2	2	3.00e-03	[2.00e-03, 6.70e-03]	0.063	4.24e-09
21MV	Biweight	2	2	5.34e-04	[2.09e-04, 1.71e-03]	0.077	2.29e-13
21DL	Biweight	2	2	4.25e-03	[1.01e-04, 5.13e-03]	0.054	8.36e-07
21L	Biweight	2	2	5.00e-03	[4.68e-03, 5.30e-03]	0.066	9.05e-10
21ML	Biweight	2	2	1.00e-03	[2.65e-04, 3.00e-03]	0.042	2.96e-04

Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

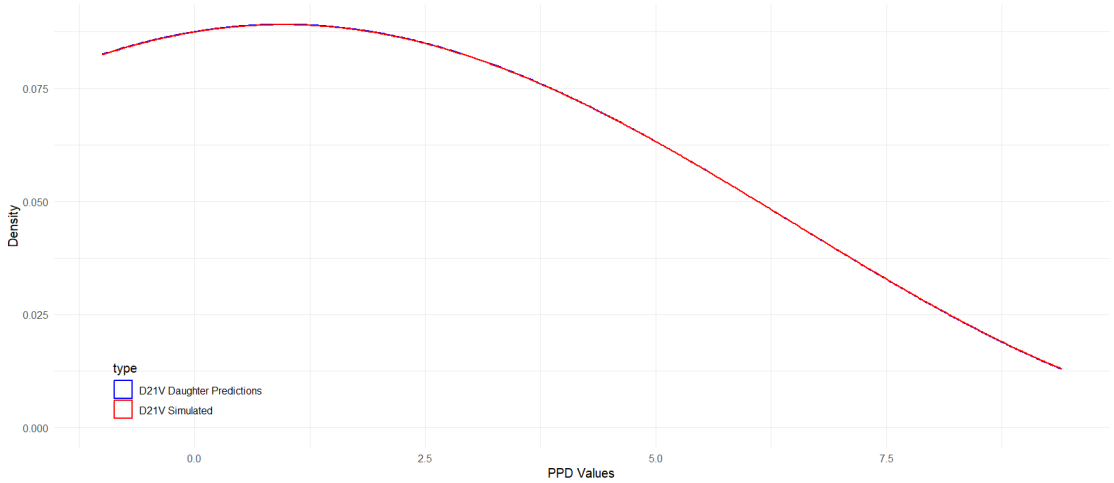
Graphical assessment The provided plot compares the KDE for simulated data (D27ML Simulated) and predicted data (D27ML Daughter Predictions).

Figure 4.18



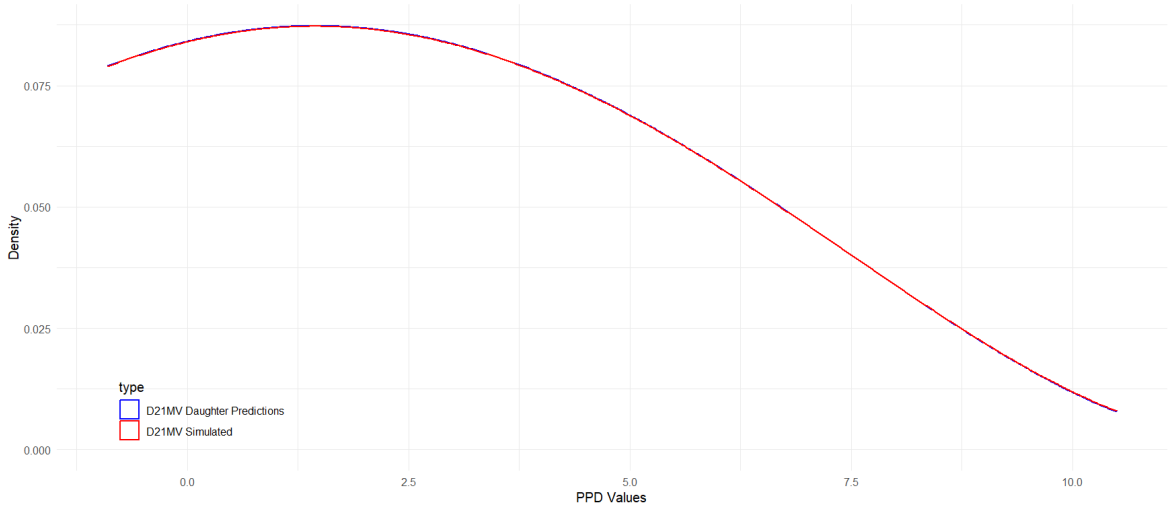
Optimal Kernel Density Plots of 21DV for Simulated and Predicted Data

Figure 4.19



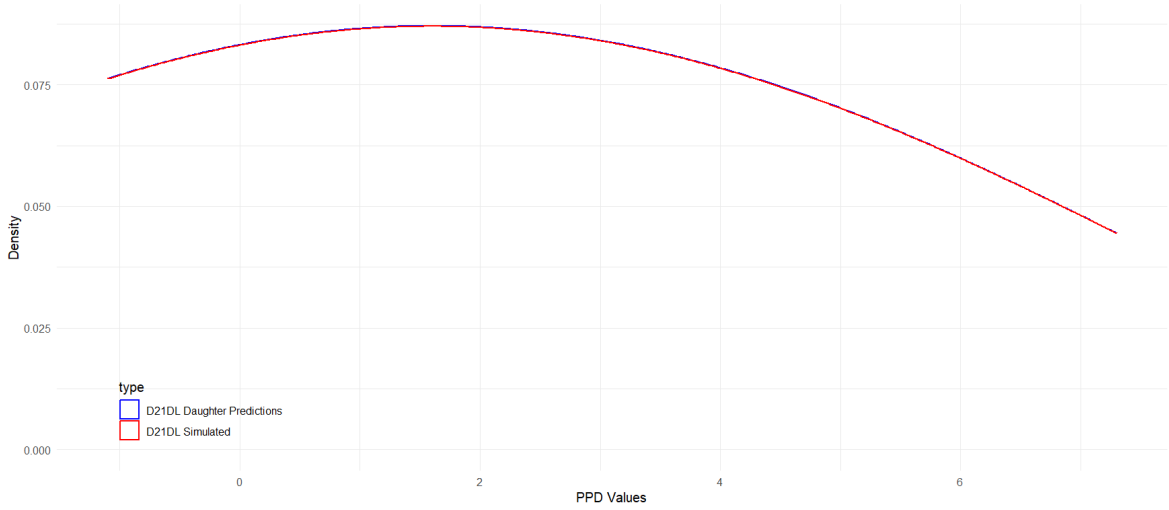
Optimal Kernel Density Plots of 21V for Simulated and Daughter Predicted

Figure 4.20



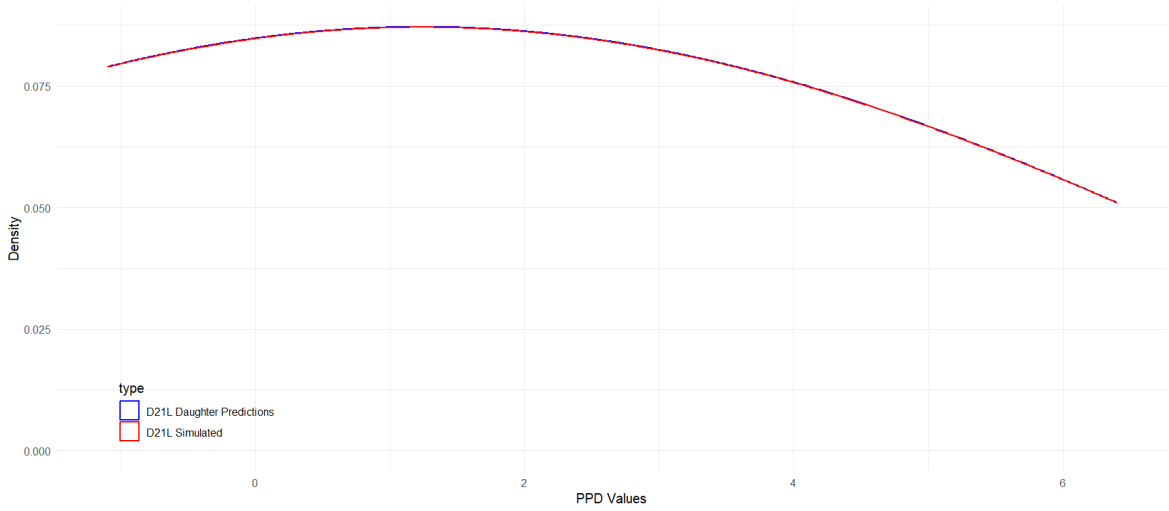
Optimal Kernel Density Plots of 21MV for Simulated and Daughter Predicted

Figure 4.21



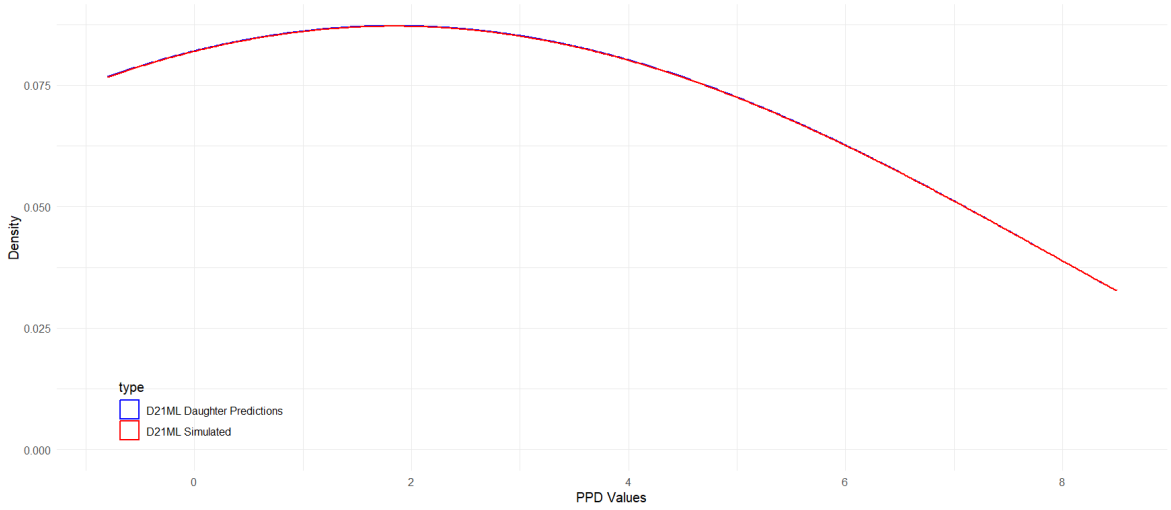
Optimal Kernel Density Plots of 21DL for Simulated and Daughter Predicted

Figure 4.22



Optimal Kernel Density Plots of 21L for Simulated and Daughter Predicted

Figure 4.23



Optimal Kernel Density Plots of 21ML for Simulated and Daughter Predicted

In the KDE plots (4.18, 4.19, 4.20, 4.21, 4.22, 4.23) the curves are almost perfectly overlapped, indicating that the distribution of the simulated data closely matches the distribution of the predicted data. The overall shape of the distribution is a smooth descending curve, suggesting high density at lower PPD (Periodontal Probing Depth) values and gradually decreasing density as the PPD values increase.

The near-perfect overlap of the density curves for the simulated and predicted data suggests that the simulation model accurately represents the distribution of the predicted data. In other words, the predictive model is very good, as the probability distributions of the simulated and predicted values are nearly identical.

This level of agreement indicates that the model can be considered reliable for accurately predicting the distribution of PPD values. The simulation has effectively captured the structure of the predicted data, which is a strong indicator of the model's validity.

4.5.2 Tooth 22 - MoDau Imputation

Mother Models M22

Basic Characteristics of M22 mother models (Table IV.14) revealed differences in model size, iterations, and RMSE values across various sites. The M22DV model had the

largest size at 386.10 Kb and required 366 iterations, whereas the M22L model was the smallest at 179.80 Kb with 168 iterations. Initial training RMSE values ranged from 1.018 for M22L to 1.311 for M22DL, while final training RMSE values varied from 1.68e-02 for M22DL to 4.38e-02 for M22L, indicating effective model training across all sites.

Performance metrics for the M22 mother models, as shown in Table IV.15. The models exhibited excellent performance with RMSE values ranging from 0.017 (M22DL) to 0.044 (M22L). MAE values were consistently low across all sites, indicating minimal absolute errors in predictions. MSE values were also low, with the highest being 1.92e-03 for M22L. The R^2 values were exceptionally high, exceeding 99.75% for all models, indicating a very strong fit to the data.

Feature importance metrics, presented in Table IV.16, highlighted the significance of the directional symmetry measure (γSM) and original NHANES 2011/2012 sites. The gain, cover, and frequency metrics were used to quantify feature importance. For example, the γSM.DV feature in the M22DV model had the highest gain (0.630) and frequency (0.706), indicating its substantial contribution to the model's predictive power. Similarly, the γSM.DL feature in the M22DL model exhibited the highest cover (0.515), emphasizing its critical role in the model's accuracy.

Daughter Model D22

The results of the D22 models predictions for the six site of upper left lateral incisor are presented in Table IV.21 and IV.20, and illustrated in the histograms IV.13, IV.15, IV.17, IV.19, IV.21, IV.23 in appendix IV.

Daughter models predictions by site

For site 22DV, the results show significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 32.169, 9.290, and 4.861, respectively, all with p-values less than 0.05. Other PPD values counts showed no significant differences. Of the comparisons, 30% had p-values less than 0.05, 30% had p-values between 0.05 and 1, and 40% had p-values equal to 1.

For site 22V, the results present significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 29.943, 10.099, and 4.149, respectively, and p-values less than 0.05. No significant differences were observed for other PPD values. Here, 37.5% of comparisons had p-values less than 0.05, 37.5% had p-values between 0.05

and 1, and 22.3% had p-values equal to 1.

For site 22MV significant differences were found for PPD values of 0 and 1, with chi-squared statistics of 37.038 and 13.155, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. This site had 22.2% of comparisons with p-values less than 0.05, 66.7% with p-values between 0.05 and 1, and 11.1% with p-values equal to 1.

In site 22DL, significant differences were identified for PPD values of 0, 1, and 3, with chi-squared statistics of 31.758, 9.732, and 5.761, respectively, all with p-values less than 0.05. Other PPD values showed no significant differences. Among the comparisons, 36.4% had p-values less than 0.05, 45.5% had p-values between 0.05 and 1, and 18.2% had p-values equal to 1, and 1 NA.

For site 22L, significant differences were found for counts of PPD 0, 1, 2, and 3, with chi-squared statistics of 11.663, 18.142, 6.975, and 5.161, respectively, and p-values less than 0.05. Other PPD values showed no significant differences. This site had 27.3% of comparisons with p-values less than 0.05, 40% with p-values between 0.05 and 1, and 10% with p-values equal to 1, and 1 NA.

Finally, in site 22ML, significant differences were observed for PPD values of 0 and 1, with chi-squared statistics of 27.806 and 9.572, respectively, and p-values less than 0.05. No significant differences were noted for other PPD values. This site had 20% of comparisons with p-values less than 0.05, 60% with p-values between 0.05 and 1, and 10% with p-values equal to 1, and 1 NA.

Aggregate No significant differences were found for the proportion of aggregate PPD values less than 4 mm or PPD values greater than or equal to 4 mm between the simulated and predicted data for each site (Table IV.20). The chi-squared statistics range from 0.078189 at site 22L to 0.69156 at site 22ML, with p-values spanning from 0.740 to 0.406.

4.5.3 Tooth 23 - MoDau Imputation

Mother Model M23

Basic Characteristics of M23 mother models (Table IV.23) revealed differences in model size, iterations, and RMSE values across various sites. The M23L model had the largest size at 386.70 Kb and required 374 iterations, whereas the M23V model was the smallest

at 215.50 Kb with 207 iterations. Initial training RMSE values ranged from 0.783 for M23V to 1.386 for M23DL, while final training RMSE values varied from 1.76e-02 for M23L to 4.99e-02 for M23DV, indicating effective model training across all sites.

Performance Metrics for the M23 mother models, as shown in Table IV.24. The models exhibited excellent performance with RMSE values ranging from 0.017 (M23L) to 0.038 (M23DV and M23V). MAE values were consistently low across all sites, indicating minimal absolute errors in predictions. MSE values were also low, with the highest being 2.49e-03 for M23DV. The R^2 values were exceptionally high, exceeding 99.67% for all models, indicating a very strong fit to the data.

Feature importance metrics, presented in Table IV.25, highlighted the significance of the directional symmetry measure (γSM) and original NHANES 2011/2012 sites. The gain, cover, and frequency metrics were used to quantify feature importance. For example, the γSM.DV feature in the M23DV model had the highest gain (0.637) and the original 13DV feature exhibited the highest cover (0.621), indicating their substantial contributions to the model's predictive power. Similarly, the γSM.L feature in the M23L model had the highest cover (0.589) and frequency (0.656), emphasizing its critical role in the model's accuracy.

Daughter Models D23

The results of the D23 models predictions for the six site of upper left lateral incisor are presented in Table IV.30 and IV.29, and illustrated in the histograms IV.25, IV.27, IV.29, IV.31, IV.33, IV.35 in appendix IV

Daughter Models Predictions by site

The results for site D23DV (Table IV.30) show significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 42.483, 7.069, and 4.961, respectively, and p-values less than 0.05. Other PPD values counts did not show significant differences. Specifically, 27.3% of the comparisons had p-values less than 0.05, 54.5% had p-values between 0.05 and 1, and 18.2% had p-values equal to 1.

For site D23V, significant differences were found for PPD values of 2, with a chi-squared statistic of 4.364 and a p-value less than 0.05. Other PPD values showed no significant differences. In this site, 10% of the comparisons had p-values less than 0.05, 40% had p-values between 0.05 and 1, and 50% had p-values equal to 1.

The results for site D23MV indicate significant differences for PPD values of 0, 1,

and 3, with chi-squared statistics of 38.071, 8.428, and 4.429, respectively, and p-values less than 0.05. No significant differences were observed for other PPD values. Here, 30% of the comparisons had p-values less than 0.05, 60% had p-values between 0.05 and 1, and 10% had p-values equal to 1.

For site D23DL, significant differences were identified for PPD values of 0 and 1, with chi-squared statistics of 44.416 and 7.688, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. This site had 10% of the comparisons with p-values less than 0.05, 30% with p-values between 0.05 and 1, and 50% with p-values equal to 1.

In site D23L, significant differences were found for PPD values of 0, 1, and 3, with chi-squared statistics of 15.884, 12.389, and 6.590, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. This site had 27.3% of the comparisons with p-values less than 0.05, 45.5% with p-values between 0.05 and 1, and 18.2% with p-values equal to 1 and 1 NA.

For site D23ML, significant differences were observed for PPD values of 0, 1, and 3, with chi-squared statistics of 41.160, 9.643, and 5.191, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. This site had 27.3% of the comparisons with p-values less than 0.05, 45.5% with p-values between 0.05 and 1, and 27.3% with p-values equal to 1.

Aggregate

No significant differences were found for the proportion of aggregate PPD values less than 4 mm or PPD values greater than or equal to 4 mm between the simulated and predicted data for each site (Table IV.29). The chi-squared statistics range from 0.004 at site 23L to 1.114 at site 23DL, with p-values spanning from 0.947 to 0.291.

4.5.4 Tooth 24 - MoDau Imputation

Mother Models M24

Basic characteristics of M24 mother models (Table IV.32) revealed differences in model size, iterations, and RMSE values across various sites. The M24L model had the largest size at 418.50 Kb and required 396 iterations, whereas the M24V model was the smallest at 163.50 Kb with 158 iterations. Initial training RMSE values ranged from 0.761 for M24V to 1.412 for M24DL, while final training RMSE values varied from 4.78e-03 for

M24L to 3.30e-02 for M24ML, indicating effective model training across all sites.

Performance metrics for the M24 mother models, as shown in Table IV.33. The models exhibited excellent performance with RMSE values ranging from 0.005 (M24L) to 0.033 (M24DV and M24ML). MAE values were consistently low across all sites, indicating minimal absolute errors in predictions. MSE values were also low, with the highest being 1.09e-03 for M24ML. The R^2 values were exceptionally high, exceeding 99.81

Feature importance metrics presented in Table IV.34, highlighted the significance of the directional symmetry measure (γSM) and original NHANES 2011/2012 sites. The gain, cover, and frequency metrics were used to quantify feature importance. For example, the γSM.DV feature in the M24DV model had the highest gain (0.655) and the original 14DV feature exhibited the highest cover (0.538), indicating their substantial contributions to the model's predictive power. Similarly, the γSM.L feature in the M24L model had the highest cover (0.568) and frequency (0.671), emphasizing its critical role in the model's accuracy.

Daughter Models D24

Daughter Models Predictions by site

The results of the D24 models predictions for the six site of upper left lateral incisor are presented in Table IV.39 and IV.38, and illustrated in the histograms IV.37, IV.39, IV.41, IV.43, IV.45, IV.47 in appendix IV

The results for site D24DV show significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 37.253, 7.761, and 7.012, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. Specifically, 33.3% of the comparisons had p-values less than 0.05, 55.6% had p-values between 0.05 and 1, and 11.1% had p-values equal to 1.

For site D24V, no significant differences were found for any PPD values, with p-values greater than 0.05 for all comparisons. This site had 0% of comparisons with p-values less than 0.05, 71.4% with p-values between 0.05 and 1, and 28.6% with p-values equal to 1.

The results for site D24MV indicate significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 65.878, 8.547, and 8.117, respectively, and p-values less than 0.05. No significant differences were observed for other PPD values. Here, 33.3% of the comparisons had p-values less than 0.05, 55.6% had p-values between 0.05

and 1, and 11.1% had p-values equal to 1.

For site D24DL, significant differences were identified for PPD values of 0 and 1, with chi-squared statistics of 55.779 and 3.937, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. This site had 22.2% of the comparisons with p-values less than 0.05, 55.6% with p-values between 0.05 and 1, and 22.2% with p-values equal to 1.

In site D24L, significant differences were found for PPD values of 0, 1, and 3, with chi-squared statistics of 20.873, 14.560, and 4.765, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. This site had 37.5% of the comparisons with p-values less than 0.05, 50% with p-values between 0.05 and 1, and 12.5% with p-values equal to 1.

For site D24ML, significant differences were observed for PPD values of 0 and 1, with chi-squared statistics of 32.816 and 7.493, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. This site had 22.2% of the comparisons with p-values less than 0.05, 55.6% with p-values between 0.05 and 1, and 22.2% with p-values equal to 1.

Aggregate

No significant differences were found for the proportion of aggregate PPD values less than 4 mm or PPD values greater than or equal to 4 mm between the simulated and predicted data for each site (Table IV.20). The chi-squared statistics range from 0.0000 at site 24V to 1.0133 at site 24DL, with p-values spanning from 1 to 0.314.

4.5.5 Tooth 25 - MoDau Imputation

Mother Models M25

Basic Characteristics of M25 mother models (Table IV.41) revealed differences in model size, iterations, and RMSE values across various sites. The M25L model had the largest size at 615.30 Kb and required 591 iterations, whereas the M25V model was the smallest at 133.10 Kb with 123 iterations. Initial training RMSE values ranged from 0.824 for M25V to 1.441 for M25DL, while final training RMSE values varied from 5.70e-03 for M25L to 3.54e-02 for M25V, indicating effective model training across all sites.

Performance Metrics for the M25 mother models, as shown in Table IV.42. The models exhibited excellent performance with RMSE values ranging from 0.006 (M25L)

to 0.035 (M25V). MAE values were consistently low across all sites, indicating minimal absolute errors in predictions. MSE values were also low, with the highest being 1.25e-03 for M25V. The R^2 values were exceptionally high, exceeding 99.79% for all models, indicating a very strong fit to the data.

Feature importance metrics, presented in Table IV.43, highlighted the significance of the directional symmetry measure (γSM) and original NHANES 2011/2012 sites. The gain, cover, and frequency metrics were used to quantify feature importance. For example, the γSM.DV feature in the M25DV model had the highest gain (0.722) and the original 15DV feature exhibited the highest cover (0.522), indicating their substantial contributions to the model's predictive power. Similarly, the γSM.L feature in the M25L model had the highest cover (0.563) and frequency (0.685), emphasizing its critical role in the model's accuracy.

Daughter Models D25

The results of the D25 models predictions for the six site of upper left lateral incisor are presented in Table IV.48 and IV.47, and illustrated in the histograms IV.49, IV.51, IV.53, IV.55, IV.57, IV.59 in appendix IV.

Daughter Models Predictions by site

The results for site D25DV (Table IV.48) revealed significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 35.665, 7.306, and 9.060, respectively, and p-values less than 0.05. Other PPD values showed no significant differences. Specifically, 33.3% of the comparisons had p-values less than 0.05, 33.3% had p-values between 0.05 and 1, and 33.3% had p-values equal to 1.

For site D25V, significant differences were identified for PPD values of 0, 1, and 2, with chi-squared statistics of 5.330, 8.524, and 5.598, respectively, and p-values less than 0.05. No significant differences were observed for other PPD values. This site had 37.5% of comparisons with p-values less than 0.05, 62.5% with p-values between 0.05 and 1, and 12.5% with p-values equal to 1.

The results for site 25MV indicated significant differences for PPD values of 0, 1, and 3, with chi-squared statistics of 38.892, 4.763, and 5.863, respectively, and p-values less than 0.05. No significant differences were found for other PPD values. In this site, 33.3% of comparisons had p-values less than 0.05, 33.3% had p-values between 0.05 and 1, and 33.3% had p-values equal to 1.

For site D25DL, significant differences were detected for PPD values of 0, 1, and 3, with chi-squared statistics of 76.281, 11.061, and 3.893, respectively, and p-values less than 0.05. Other PPD values showed no significant differences. This site had 30% of comparisons with p-values less than 0.05, 50% with p-values between 0.05 and 1, and 20% with p-values equal to 1.

In site D25L, significant differences were found for PPD values of 0, 1, and 3, with chi-squared statistics of 18.731, 13.952, and 6.297, respectively, and p-values less than 0.05. No significant differences were noted for other PPD values. This site had 37.5% of comparisons with p-values less than 0.05, 62.5% with p-values between 0.05 and 1, and 0% with p-values equal to 1.

For site D25ML, significant differences were observed for the PPD value of 0, with a chi-squared statistic of 32.801 and a p-value less than 0.05. No significant differences were noted for other PPD values counts. This site had 11.1% of comparisons with p-values less than 0.05, 77.8% with p-values between 0.05 and 1, and 11.1% with p-values equal to 1.

Aggregate

No significant differences were found for the proportion of aggregate PPD values less than 4 mm or PPD values greater than or equal to 4 mm between the simulated and predicted data for each site (Table IV.20). The chi-squared statistics range from 0.01399 at site 25V to 1.5036 at site 22ML, with p-values spanning from 0.906 to 0.220.

4.5.6 Tooth 26 - MoDau Imputation

Mother Model M26

Basic characteristics of M26 mother models (Table IV.50) revealed differences in model size, iterations, and RMSE values across various sites. The M26MV model had the largest size at 488.30 Kb and required 472 iterations, whereas the M26V model was the smallest at 201.70 Kb with 191 iterations. Initial training RMSE values ranged from 0.967 for M26V to 1.652 for M26DL, while final training RMSE values varied from 1.15e-02 for M26L to 3.17e-02 for M26V, indicating effective model training across all sites.

Performance Metrics for the M26 mother models, as shown in Table IV.51. The models exhibited excellent performance with RMSE values ranging from 0.011 (M26L)

to 0.032 (M26V). MAE values were consistently low across all sites, indicating minimal absolute errors in predictions. MSE values were also low, with the highest being 1.01e-03 for M26V. The R^2 values were exceptionally high, exceeding 99.84% for all models, indicating a very strong fit to the data.

Feature importance metrics, presented in Table IV.52, highlighted the significance of the γSM and original NHANES 2011/2012 sites. The gain, cover, and frequency metrics were used to quantify feature importance. For example, the γSM.MV feature in the M26MV model had the highest gain (0.709) and frequency (0.650), indicating its substantial contribution to the model's predictive power. Similarly, the γSM.L feature in the M26L model had the highest cover (0.514) and frequency (0.635), emphasizing its critical role in the model's accuracy.

Daughter Models D26

Daughter Models Predictions by site

The results of the D26 models predictions for the six site of upper left lateral incisor are presented in Table IV.57 and IV.56, and illustrated in the histograms IV.61, IV.63, IV.65, IV.67, IV.69, IV.71 in appendix IV.

By Site

For site D26DV, significant differences were observed for PPD values of 0, 1, 2, and 3, with chi-squared statistics of 111.420, 20.689, 8.909, and 20.205, respectively, all with p-values less than 0.05. Other PPD values did not show significant differences. Overall, 40% of the comparisons had p-values less than 0.05, 50% had p-values between 0.05 and 1, and 10% had p-values equal to 1.

At site D26V, significant differences were found for PPD values of 0 and 1, with chi-squared statistics of 8.187 and 7.641, respectively, and p-values less than 0.05. Other PPD values showed no significant differences. Here, 22.2% of the comparisons had p-values less than 0.05, 44.4% had p-values between 0.05 and 1, and 33.3% had p-values equal to 1.

For site D26MV, significant differences were identified for PPD values of 0, 1, and 3, with chi-squared statistics of 27.824, 4.459, and 3.879, respectively, all with p-values less than 0.05. Other PPD values did not show significant differences. In this case, 37.5% of the comparisons had p-values less than 0.05, 50% had p-values between 0.05 and 1, and 12.5% had p-values equal to 1.

In site D26DL, significant differences were found for PPD values of 0, 2, and 3, with chi-squared statistics of 50.920, 6.349, and 4.840, respectively, all with p-values less than 0.05. Other PPD values showed no significant differences. This site had 33.3% of comparisons with p-values less than 0.05, 55.6% with p-values between 0.05 and 1, and 11.1% with p-values equal to 1.

At site D26L, significant differences were noted for PPD values of 0, 1, and 3, with chi-squared statistics of 15.651, 5.837, and 4.899, respectively, all with p-values less than 0.05. Other PPD values did not show significant differences. Here, 37.5% of the comparisons had p-values less than 0.05, 50% had p-values between 0.05 and 1, and 12.5% had p-values equal to 1.

Finally, for site D26ML, significant differences were observed for PPD values of 0, 1, and 3, with chi-squared statistics of 42.378, 5.229, and 3.945, respectively, all with p-values less than 0.05. Other PPD values did not show significant differences. This site had 33.3% of comparisons with p-values less than 0.05, 55.6% with p-values between 0.05 and 1, and 11.1% with p-values equal to 1.

Aggregate

No significant differences were found for the proportion of aggregate PPD values less than 4 mm or PPD values greater than or equal to 4 mm between the simulated and predicted data for each site (Table IV.20). The chi-squared statistics range from 0.069094 at site 26V to 1.5863 at site 26DV, with p-values spanning from 0.793 to 0.208.

4.5.7 Tooth 27 - MoDau Imputation

Mother Models M27

The analysis of the M27 Mother Models focuses on three primary aspects: model characteristics, performance metrics, and feature importance.

Model Characteristics

Table IV.59 presents the distinctive characteristics of the M27 models across different sites. Key metrics include the model size, number of iterations (N.Iter.), initial training RMSE (Root Mean Square Error), and final training RMSE.

The Mother Model Size varies significantly among the models, with M27ML being the largest at 709.40 Kb and M27DL and M27L the smallest at approximately 198.30 Kb. The number of iterations follows a similar pattern, where M27ML required the most iterations (672), indicating a potential correlation between model size and computational complexity.

Initial training RMSE values indicate the error rate at the beginning of the training process, with M27MV showing the highest initial RMSE of 1.729 and M27V the lowest at 1.172. Final training RMSE values show a substantial reduction across all models, with the smallest final RMSE observed in M27ML (9.68e-03), suggesting superior training efficacy for this model.

Performance Metrics

The performance of the M27 models, as shown in Table IV.60, is evaluated using RMSE, MAE (Mean Absolute Error), MSE (Mean Squared Error), and R^2 (coefficient of determination).

M27ML outperforms other models across all performance metrics, with the lowest RMSE (0.010), MAE (0.002), and MSE (9.36e-05). Its R^2 value is nearly perfect at 99.99%, indicating an excellent fit. Other models also demonstrate high performance, with R^2 values ranging from 99.72% to 99.85%, but M27ML's metrics suggest it has the highest predictive accuracy and reliability.

Feature Importance

Feature importance metrics for the M27 models are detailed in Table IV.61. The primary features considered are γSM.Site (Directional Symmetry Measure computed from original data) and the Original 17Site features.

For each model, γSM.Site consistently shows higher importance scores across Gain, Cover, and Frequency metrics compared to the Original 17Site features. For instance, in the D27ML model, γSM.ML has a Gain of 0.672, Cover of 0.650, and Frequency of 0.713, whereas Original 17ML features have lower corresponding values (Gain: 0.328, Cover: 0.350, Frequency: 0.287). This indicates that γSM.Site features are more significant contributors to the model's predictive power.

Conclusion

The analysis reveals significant variability in the size and training iterations required for different M27 Mother Models, with M27ML standing out due to its larger size and higher number of iterations. Performance metrics demonstrate that M27ML is the most accurate and reliable model. Feature importance analysis highlights the dominant role of γSM.Site features in driving model performance across all M27 models.

Daughter Models D27

The analysis of the D27 Daughter Models is structured around three main dimensions: distinctive model characteristics, performance metrics, and feature importance.

Model Characteristics

Table IV.62 provides an overview of the distinctive characteristics of the D27 models. The key metrics analyzed include model size, number of iterations (N.Iter.), initial training RMSE, and final training RMSE.

The model sizes vary from 6.6 Mb (D27DV) to 32.4 Mb (D27L), indicating a range in model complexity and storage requirements. The number of iterations shows significant differences, with most models reaching around 30,000 iterations, except D27DV and D27DL, which had 6,287 and 21,060 iterations respectively. Initial training RMSE values suggest that D27MV started with the highest error (1.851), while D27V had the lowest (1.263). Final training RMSE values display a notable reduction for all models, with D27V achieving the lowest final RMSE (0.147), indicating the effectiveness of the training process.

Performance Metrics

Table IV.63 summarizes the performance metrics of the D27 models, evaluated using RMSE, MAE, MSE, and R^2.

The performance metrics reveal that D27V and D27L models have the lowest RMSE (0.147 and 0.148 respectively) and MSE (0.022 for both), signifying high accuracy. The MAE values are also lowest for these models (0.116 for D27V and 0.117 for D27L), reinforcing their strong predictive performance. The R^2 values, which measure the proportion of variance explained by the models, are above 97

Feature Importance

Feature importance metrics, presented in Table IV.64, focus on three types of features: Mother Predictions, New Mother Predictions, and the Original 17Site features.

For each model, Mother Predictions generally have the highest Gain, Cover, and Frequency values. For instance, in the D27DV model, Mother Predictions have a Gain of 0.534, Cover of 0.357, and Frequency of 0.375, indicating their significant contribution to the model. The Original 17Site features, while also important, show variable importance across models. For example, in D27V, the 17V feature has a Gain of 0.274, Cover of 0.595, and Frequency of 0.604, showing substantial importance. The New Mother Predictions tend to have the lowest importance metrics, suggesting their lesser influence on model performance.

Conclusion

The analysis highlights substantial variability in the size and training iterations among the D27 Daughter Models, with D27V and D27L standing out due to their superior performance metrics. Feature importance analysis underscores the dominant role of Mother Predictions across all models, although the Original 17Site features also play a critical role in some cases. Overall, the D27 models exhibit high predictive accuracy and reliability, with certain models demonstrating particularly strong performance metrics.

Daughter Model D27 Predictions

The results of the D27 models predictions for the six sites of the upper left second molar are presented in Tables IV.66 and IV.65, and illustrated in the histograms IV.73, IV.75, IV.77, IV.79, IV.81, IV.83 in Appendix IV.

Chi-squared Test Results for PPD Values $\geq$ 4mm

Table IV.65 presents the chi-squared test results comparing the proportions of PPD (probing pocket depth) values $\geq$ 4mm between the original and predicted data across different D27 models. The test evaluates whether the predicted proportions significantly differ from the simulated data proportions.

For all sites, the chi-squared statistic values are relatively low, and the corresponding p-values are greater than 0.05, indicating no significant differences between the predicted and simulated proportions of PPD values $\geq$ 4mm. This suggests that the D27 models accurately predict the distribution of PPD values in this range.

Comparison of Proportions between Simulated and Predicted by Unique Value of PPD

For site D27DV, significant differences were observed for PPD values of 0, 1, 3, and 7, with chi-squared statistics of 114.786, 13.715, 9.142, and 4.455, respectively, all with p-values less than 0.05. Other PPD values did not show significant differences. Overall, 40% of the comparisons had p-values less than 0.05, 40% had p-values between 0.05 and 1, and 20% had p-values equal to 1.

At site D27V, significant differences were found for PPD values of 0, 1, and 3, with chi-squared statistics of 19.660, 8.895, and 4.135, respectively, and p-values less than 0.05. Other PPD values showed no significant differences. Here, 42.9% of the comparisons had p-values less than 0.05, 57.1% had p-values between 0.05 and 1, and 0% had p-values equal to 1.

For site D27MV, significant differences were identified for PPD values of 0 and 2, with chi-squared statistics of 36.511 and 6.331, respectively, and p-values less than 0.05. Other PPD values did not show significant differences. In this case, 20% of the comparisons had p-values less than 0.05, 60% had p-values between 0.05 and 1, and 20% had p-values equal to 1.

In site D27DL, significant differences were found for PPD values of 0, 2, and 3, with chi-squared statistics of 53.228, 9.858, and 9.145, respectively, all with p-values less than 0.05. Other PPD values showed no significant differences. This site had 30% of comparisons with p-values less than 0.05, 50% with p-values between 0.05 and 1, and 20% with p-values equal to 1.

At site D27L, significant differences were noted for PPD values of 0, 1, and 3, with chi-squared statistics of 33.076, 5.197, and 5.118, respectively, all with p-values less than 0.05. Other PPD values did not show significant differences. Here, 30% of the comparisons had p-values less than 0.05, 30% had p-values between 0.05 and 1, and 40% had p-values equal to 1.

Finally, for site D27ML, significant differences were observed for PPD values of 0, 2, and 3, with chi-squared statistics of 33.151, 7.097, and 4.225, respectively, all with p-values less than 0.05. Other PPD values did not show significant differences. This site had 27.3% of comparisons with p-values less than 0.05, 54.5% with p-values between 0.05 and 1, and 9.1% with p-values equal to 1.

Aggregate

No significant differences were found for the proportion of aggregate PPD values less than 4 mm or PPD values greater than or equal to 4 mm between the simulated and predicted data for each site (Table IV.20). The chi-squared statistics range from 0.2839 at site 27V to 1.5904 at site 27DV, with p-values spanning from 0.594 to 0.207.

Conclusion

The chi-squared test results indicate that, overall, the D27 models provide accurate predictions of PPD values $\geq$ 4mm, as evidenced by the non-significant chi-squared statistics. However, when examining specific PPD values, several significant differences emerge, highlighting areas where model predictions could be improved. The most notable discrepancies are found in lower PPD values (0, 1, and 3mm) across multiple models, suggesting a potential area for model refinement. Clinically, these discrepancies are of low relevance since they occur in PPD values that do not affect the diagnosis. For diagnostic purposes, PPD values greater than 3mm are more critical, and the models perform well in predicting these values. Additionally, 31.58% of the tests yielded p-values equal to 1, indicating exact matches in predictions, while 40.35% of the tests had p-values less than 0.05, signifying statistically significant differences.

Overall Results Reports on Simulated versus Predicted with KDE

In Appendix IV, the tables reporting the KDE differences between simulated versus daughter models predicted showed consistent results across tooth sites in terms of bandwidth and adjustment parameters, both set to 2. The best kernel functions observed were the cosine kernel for the following sites: teeth 22, 23, 24, 25, and 26 at DV sites; teeth 21 and 24 at V sites; tooth 27 at MV site; teeth 25, 26, and 27 and DL sites; tooth 24 at L site; and teeth 26 and 27 at ML sites. The biweight kernel was optimal for the remaining sites.

Plots of optimal kernel density estimates across all tooth sites (Appendix IV) for simulated and daughter predicted data almost perfectly overlapped, indicating that the distribution of the simulated data closely matches the distribution of the predicted data. These observations corroborate the small mean differences between kernels. The overall shape of the distribution is a smooth descending curve, suggesting high density at lower PPD (Periodontal Probing Depth) values and gradually decreasing density as the PPD values increase. This pattern aligns with the frequency column plots by

unique PPD value. The bootstrap differences between the KDE of the simulated and predicted data are on the order of a thousandth. Such small differences indicate that the predicted and simulated distributions are very similar in shape and spread. Although these differences are small, the KS test revealed statistically significant differences with small p-values (ranging from 0 to 5.77e-03) across all 42 sites compared.

Chapter 5

Discussion

5.1 Introduction

The main objective of this work was to develop a method for imputing missing data using total contralateral planned omission, achieving satisfactory results, i.e., a correct prediction of the prevalence of cases with pathology (PPD $>$ 3). One of the most promising methodologies involved incorporating rules that capture the diffuse symmetry occurring between periodontal lesions in a neural network model. However, to achieve our primary objective, we tested various combinations of methods.

The combination that yielded results closest to our main objective was the integration of the symmetry degree (SM) from the symmetry function developed during this research, a parameter indicating the directionality of the difference (γ), and information obtained a priori from fully known data with predictors from an XGBoost model. Some of the remaining potential solutions will be developed in future works.

5.2 Assessing symmetry

The kernel density estimation (KDE) plots were used to compare the distributions of two paired PPD variables, which are assumed to have a high degree of similarity or symmetry. Given the nature of our data, non-parametric methods are more adequate to estimate the probability density function of PPD variables. These circumstances included contralateral sites where symmetry was assessed, and pre- and post-imputation scenarios where similarity was evaluated. The KDE approach varied in the degree of smoothing and the type of kernel used, with adjustments made as necessary. The grade of smoothing in kernel estimation is related to the aim, whether to assess the details

of variable behavior or the trends of the distributions. In this work, we used both approaches.

The prevalence of a specific PPD value could be equal on both sides, corresponding to the overlap of the segment of the KDE. If this occurs across all segments of the KDE, the distributions are similar and possibly equal. However, if the same values of PPD do not coincide simultaneously across both sides, they are not symmetrical. When the prevalence of certain values is high and identical in both contralateral sites, the probability of a high number of contralateral values being equal is high. In the limit, two KDEs may overlap perfectly, but the symmetry grade can be relatively low if the PPD values do not coincide on both sides. In summary, even with equal prevalence of a given PPD value, true symmetry requires the values to coincide simultaneously on both sides. Two probability distributions can be equal without physical symmetry of the disease.

When evaluating the magnitude and significance of the effect size of the side using GAMLSS models, or correlation values using Pearson correlation, the PPD enters the computations as paired observations, i.e., in the same row in different columns, both indexed to the participant. Measured this way, a high correlation (if positive, indicating that both sides' values vary in the same direction) means a high grade of symmetry. If the values vary inversely, we would not be talking about reflexive symmetry (although we could argue about some form of symmetry).

This approach relies on hypothesis testing. Perfect symmetry occurs when the null hypothesis, stating that there is no difference between the two parameters being compared, cannot be rejected. This means that the difference between the two parameters lies within a confidence interval that includes zero at a given confidence level, indicating a high probability of being symmetric. Quasi-symmetry within the hypothesis testing framework, on the other hand, allows for a certain grade of difference within the confidence interval. It acknowledges that while the parameters may not be exactly equal, the difference between them is within an acceptable range, indicating a degree of symmetry that is practically significant.

Overall, the Bhattacharyya coefficients across all tables show high similarity in the PPD distributions at contralateral sites, typically ranging from 0.980 to 1.000. This suggests a strong resemblance in the PPD distribution shapes between opposing teeth. The correlation coefficients, which measure the linear association between PPD values, generally fall within the moderate to high range (0.43 to 0.73), indicating a consistent linear relationship in PPD measurements across contralateral dental sites. The

detailed statistical analysis for each dental site and corresponding tooth pairs supports the conclusion that there is a significant degree of symmetry in PPD values between contralateral sites, as reflected by both the Bhattacharyya coefficients and correlation coefficients. These findings are crucial for understanding the bilateral symmetry in periodontal health, as detailed in Tables 4.1 to I.79.

In the context of reflexive or fuzzy symmetry, the concept of symmetry in PPD values can be nuanced. Reflexive symmetry implies a mirror-like symmetry where values on one side are exactly mirrored on the other. This would mean not only equal prevalence of PPD values but also their simultaneous occurrence on corresponding sites. Fuzzy symmetry, on the other hand, allows for some degree of approximation or tolerance. It acknowledges that while exact mirroring may not occur, there is still a significant degree of similarity or pattern consistency between sides. The symmetry measuring function (SM), by incorporating the principles of fuzzy symmetry, emerges as a versatile tool for assessing the symmetry grade between two distinct values, A and A'. It transcends a mere binary categorization, offering a gradient measure adept at capturing subtle nuances in symmetry levels. This function becomes a powerful tool, as demonstrated in the feature importance metrics, to capture symmetry grades, allowing for the explicit advantage of the fuzzy symmetry pattern.

Probing pocket depth (PPD) values between 0 and 3 mm are clinically classified as healthy and epidemiologically as normal. In a fuzzy logic system, these values are covered by a fuzzy rule designated as "healthy." The possible contralateral numerical differences are not relevant. Thus, when the respective contralateral PPD values also fall within this range, the relationship between the two sides of the mouth in terms of PPD can be classified as symmetrical. More broadly, the symmetry measure (SM) value could be "1" under the following conditions: $A = A'$ or A and A' simultaneously belong to the interval [0, 3], which we could designate as "healthy." In these cases, the pair (A, A') is any pair of numbers within the "healthy" interval.

$$\text{SM} = 1 \quad \text{if and only if} \quad A = A' \quad \text{or} \quad A \in \{0, 1, 2, 3\} \quad \text{and} \quad A' \in \{0, 1, 2, 3\}.$$

This interval includes more than 95% of PPD values, which does not imply that two contralateral PPD values simultaneously belong to the "healthy" interval but suggests a high probability. Considering these conditions, the percentage of $SM = 1$, already above 50% (see graphs on the percentage of SM values per site), would increase considerably. In this study, we primarily consider the numerical perspective of our results, providing greater definition to our analysis.

In the visual assessment the KDE overlap provides a visual and quantitative method to assess this symmetry. If the KDEs of PPD values from both sides overlap significantly across their entire range, it suggests similar distributions. However, perfect overlap does not necessarily equate to reflexive symmetry, especially if the corresponding values do not align precisely across the sides. In such cases, despite high overlap, the physical manifestation of the disease may not be symmetrically distributed.

When using statistical methods like GAMLSS or Pearson correlation, the analysis of paired observations allows for a more granular assessment of symmetry. High positive correlation indicates that as PPD values increase or decrease on one side, they do so similarly on the contralateral side, which aligns with the concept of reflexive symmetry. Conversely, an inverse relationship would suggest a different form of symmetry or possibly an asymmetrical pattern, which could still be clinically relevant but not reflexive.

In conclusion, the analysis of symmetry in PPD values using KDE, correlation, and GAMLSS models provides valuable insights into the distribution patterns of periodontal disease. While high prevalence and similar distributions are indicative of potential symmetry, true reflexive symmetry requires precise alignment of values across sides. Fuzzy symmetry allows for a broader interpretation, accommodating the grading of deviations from perfect symmetry while still recognising underlying symmetrical patterns. This nuanced understanding is crucial for accurate clinical assessments and data imputation applications. This approach is complemented with the, important human perception of symmetry addressed in the next paragraph.

The clinical perception of symmetry by trained dentists is highly correlated with SM values, demonstrating the reliability of the SM function in assessing periodontal probing depth symmetry. This finding suggests that the computational method can serve as a robust tool for evaluating periodontal disease asymmetry, potentially reducing the subjectivity inherent in professional assessments.

5.3 MoDau Imputation

In this study, the imputation process using the MoDau method was meticulously designed from the beginning to ensure the accuracy and validity of the imputed values.

The first step involved imputing missing values, as XGBoost models depend on the variability within the data, and missing values can significantly affect the accuracy of predictions.

In dental datasets, teeth or sites might be missing due to various medical reasons such as dental decay, periodontal disease, trauma, or survey methodology. The mechanisms behind missing data can be complex and not well understood. Therefore, we assume that the missing data mechanisms are both MAR (Missing At Random) and MNAR (Missing Not At Random).

The missing data issue was addressed with chose the hot deck imputation methodology, known for its ability to handle different types of missingness mechanisms (Missing Completely At Random, Missing At Random, and some types of Not Missing At Random) while maintaining the inherent structure and distribution of the data. Preserving the original data structure is particularly important for subsequent analyses using machine learning techniques, specifically XGBoost models. Hot deck is advantageous compared with methods like mean imputation or regression imputation can introduce bias, especially when the missing data is not randomly distributed. Hot deck imputation avoids this by using real observed values from similar records, reducing the risk of introducing systematic bias into the dataset. This leads to more reliable XGBoost model outcomes, as discussed by Rubin, 1987b and Little and Rubin, 2002.

Hot deck imputation replaces missing values by matching similar individuals based on available data and using their observed values for imputation. This approach is supported by Little and Rubin, 2002, who highlighted the versatility of hot deck imputation in handling complex missing data mechanisms.

In the context of imputing missing teeth data due to medical reasons before applying XGBoost models, several considerations support its use. By replacing missing values with observed values from similar individuals, we ensure that the imputed values follow the same distribution as the original data. This preservation of data distribution is particularly important in dental studies, where the distribution of missing teeth could be influenced by various demographic and clinical factors. Preserving this distribution helps maintain the integrity of the dataset, which is crucial for better performance of XGBoost models. Previous studies have emphasized the importance of maintaining data distribution for reliable model outcomes (Rubin, 1987b; Schafer, 1999).

Accurate imputation ensures that XGBoost models have a complete and high-quality dataset to learn from, enhancing their predictive capabilities. The effectiveness of hot deck imputation in improving predictive performance has been supported by multiple studies (Little & Rubin, 2002; Schafer & Graham, 2002).

Methods like mean imputation or regression imputation can introduce bias, especially when the missing data is not randomly distributed. Hot deck imputation avoids

this by using real observed values from similar records, reducing the risk of introducing systematic bias into the dataset. This leads to more reliable XGBoost model outcomes, as discussed by Rubin, 1987b and Little and Rubin, 2002. Furthermore maintains the natural variability of the dataset by using actual observed values for imputation. This contrasts with methods like mean imputation that reduce variability by introducing a constant value for missing data. The maintenance of data variability is critical for the predictive performance of machine learning models, as noted by Hastie et al., 2009.

Furthermore, hot deck imputation can utilise auxiliary information to find similar records for imputation. In a dental study, variables such as age, gender, smoking status, and overall dental health can be used to identify similar patients. This results in more accurate imputation compared to methods that do not use such auxiliary information, thus enhancing the quality of the dataset before applying XGBoost models. The use of auxiliary variables for improving imputation accuracy has been well-documented (Schafer & Graham, 2002).

In our study, by utilising only the PPD values, we were able to maintain even proportions of PPD by unique values and by aggregated values (3 and $leq3$). Additionally, the medians and variances were preserved according to the majority of tests performed. Differences in medians were only observed in sites 12DV, 12DL, and 17DL; differences in variance were noted in sites 12DV, 16DL, and 22V. Proportional differences were seen with PPD = 1 in sites 11DV and 17DL, PPD = 2 in 11DV, PPD = 4 in 16DL, and statistically significant differences in the K-S test were found in sites 11DV and 12DL.

The analysis of the KDE plots showed substantial overlap, suggesting a high similarity between pre- and post-imputation distributions, which aligns with the results of the statistical tests.

In conclusion, hot deck imputation is a rational and effective choice for handling missing teeth data before applying XGBoost models. It preserves the data distribution, handles complex missing data mechanisms, utilises auxiliary information, avoids imputation bias, maintains data variability, and ultimately improves the predictive performance of XGBoost models. These advantages make hot deck imputation a suitable method for preparing dental datasets for machine learning analysis, as demonstrated by the presented results.

Generating synthetic data was not our initial option; we decided to simulate data after encountering difficulties in finding real data of sufficient quality and quantity. Our method for simulating data aims to create a dataset of PPD variables that exhibit

a right-skewed distribution with a heavy right tail, similar to an ex-Gaussian distribution. Generating distributions using the ex-Gaussian probability density function would require making arbitrary decisions about the ex-Gaussian parameter values (μ, σ, and ν), which could potentially bias the distributions toward our expectations. To avoid this, we opted for a method of adding controlled noise to the original distribution, where we just control the amount of noise to be added.

Given the goal of creating simulated data that is different enough from the original dataset but not so different that it ceases to be a plausible distribution for PPD, the selection of a noise level of 0.4 was grounded in the following rationale:

Moderate Variation: A noise level of 0.4 introduces a moderate degree of variation into the simulated data, making it sufficiently different from the original dataset while preserving the overall distribution characteristics. This helps in studying the variability of pocket probing depths in a controlled manner.

Plausibility of Distribution: Maintaining the plausibility of the distribution is crucial for simulating PPD data. A noise level of 0.4 ensures that the simulated data remains realistic and comparable to clinically observed variations in PPD, making the findings relevant and applicable to real-world scenarios.

Controlled Uncertainty: The mean difference in densities and the confidence intervals at a noise level of 0.4 show controlled and predictable variability. This is important in our research, where maintaining the clinical relevance of simulated data is essential for deriving meaningful conclusions.

Empirical Evidence: The 95% confidence interval for the mean difference at a noise level of 0.4 is [0.082, 0.158]. This range is sufficiently precise to indicate that the introduced noise provides a realistic spread of data points without excessively diverging from the original measurements.

Avoiding Over fitting and Under fitting: In the context of PPD data simulation, a noise level of 0.4 helps avoid over fitting (which would result in simulated data being too similar to the original) and under-fitting (which would result in implausible distributions). This balance ensures that the simulated PPD data is both diverse and clinically relevant.

To conclude, using a noise level of 0.4 to simulate PPD data aids in generating realistic predictive models. This level of noise ensures that the simulated data captures the inherent variability in periodontal measurements and provides a realistic testing ground for validating periodontal models. The introduced variability effectively challenges the models, ensuring they are tested against a range of plausible clinical scenarios.

Hyperparameter tuning is crucial for optimising machine learning models, especially for complex algorithms like XGBoost. To fine-tune the test model, we established a comprehensive grid of hyperparameter combinations, allowing us to thoroughly explore their interactions and systematically evaluate the XGBoost models for each combination. The goal was to identify the optimal set of parameters that yield the best-performing model configuration and ensure optimal training.

Each hyperparameter combination was evaluated using 5-fold cross-validation, aiming to minimize the squared error and maximize predictive accuracy. This rigorous evaluation process is essential to achieve the highest possible performance from the model.

The search space for the parameters was as follows: for `eta`, the range was from 0.01 to 0.1, with a step size of 0.05; for `max_depth`, from 1 to 6, with a step size of 1; for `gamma`, from 0 to 0.25, with a step size of 0.05; for `colsample_bytree`, from 0.6 to 1.6, with a step size of 0.2; for `min_child_weight`, from 1 to 4, with a step size of 1; and for `subsample`, from 0.2 to 1, with a step size of 0.1. The number of rounds was set to 500, considering the algorithm's early stopping mechanism that halts the process after 10 rounds without improvement.

The results indicated significant variation in performance across different hyperparameter settings, underscoring the importance of thorough tuning. The sequences for `eta` from 0.01 to 0.3 and `max_depth` from 1 to 6 enabled the exploration of both fine updates and larger steps in learning, providing insights into the model's sensitivity to these particularly influential parameters.

The most performant combination of parameters was: `eta` = 0.1; `max_depth` = 3; `gamma` = 0; `colsample_bytree` = 1.0; `min_child_weight` = 3; `subsample` = 0.7.

The selected learning rate (`eta`) of 0.1 allowed for finer updates, enhancing the model's ability to learn gradually without overshooting the optimal solution. Conversely, higher values of `max_depth` enabled the model to capture more complex interactions within the data. Setting a maximum tree depth of 3 ensured the models built relatively shallow trees, which helps prevent overfitting and reduces model complexity. The `gamma` parameter was set to 0, indicating no minimum loss reduction requirement for further partitioning on a leaf node of the tree. A subsample ratio of 1.0 for `colsample_bytree` means all features were used for building each tree, taking advantage of all data available. The minimum sum of instance weight (hessian) needed in a child was set to 3, helping to control overfitting by preventing the model from

learning overly specific relations to the sample data. Additionally, a subsample ratio of 0.7 indicated that 70% of the training data was randomly sampled prior to growing trees, aiding in overfitting prevention. The best model balanced its complexity with its generalization ability, as evidenced by cross-validation metrics.

In conclusion, the comprehensive hyperparameter tuning strategy employed in this study was instrumental in enhancing the performance of the XGBoost models. This approach ensured that the models were well-calibrated to handle the complexity of the data while maintaining the ability to generalize effectively, thus demonstrating the critical role of hyperparameter tuning in machine learning model optimization. This rigorous hyperparameter tuning strategy significantly improved the performance of the XGBoost models.

When performing grid search and training the models we specify the number of cores to 13 out of 16 to avoid overheating. When using all available cores (16) for grid search and model training can significantly speed up processing, though it may also lead to potential issues, considering that other tasks and applications run simultaneously. These issues include performance degradation, as both grid search/model training and other applications compete for CPU resources, causing both tasks to slow down due to the operating system managing the allocation of CPU time slices, leading to potential performance bottlenecks. Additionally, system responsiveness can be affected, making the system less responsive to user interactions, such as opening new applications, browsing the web, or performing simple system operations, which may become sluggish. Running all cores at full capacity can also increase power consumption and generate more heat, potentially affecting hardware performance despite using an updated computer system. Lastly, although modern operating systems are adept at managing CPU resources, contention issues may still arise, especially if both tasks are highly CPU-intensive.

Mother Models

The grid search for hyperparameter tuning explored a wide range of values, ensuring a thorough evaluation of each combination to identify the optimal set of parameters that minimized root mean squared error (RMSE) and maximized predictive accuracy. This comprehensive search covered sequences for `eta` (0.01 to 0.3), `max_depth` (1 to 6), `gamma`, `colsample_bytree`, `min_child_weight`, `subsample`, and `nrounds`, systematically evaluating the XGBoost models using 5-fold cross-validation. The re-

sults highlighted significant performance variation across different settings, emphasizing the importance of thorough tuning. The best-performing models balanced complexity with generalization capability, with `eta` and `max_depth` sequences providing insights into the model's sensitivity to learning rate and tree depth.

Among the Best Practices in this field the "Leave One Core Free" is a common practice is to leave one core free to ensure the system remains responsive and other applications can run smoothly, we decide on to leave 3 cores out . You can achieve this by subtracting one from the total number of detected cores:

The hyperparameter tuning process is critical for optimizing machine learning models, especially for complex models like XGBoost. In this study, an extensive grid search was conducted to ensure that a wide range of parameter combinations were evaluated, ultimately leading to the identification of optimal settings for the task at hand. The chosen sequences for `eta`, `max_depth`, `gamma`, `colsample_bytree`, `min_child_weight`, `subsample`, and `nrounds` were based on best practices and tailored to the specific characteristics of the dataset.

The results from the grid search underscore the importance of balancing model complexity with the ability to generalize to new data. Parameters like `eta` and `max_depth` were particularly influential, with lower values of `eta` providing finer updates and higher values of `max_depth` allowing for more complex interactions. The cross-validation approach ensured that the selected model configurations were not overfitted to the training data, thereby providing reliable performance on unseen data.

Furthermore, this comprehensive hyperparameter tuning strategy significantly improved the performance of the XGBoost models. The findings highlight the necessity of methodical tuning, which not only enhances model accuracy but also ensures robustness and generalizability. Future work could explore more advanced tuning methods, such as Bayesian optimization, to further refine the hyperparameter search process. These advanced methods could potentially reduce the computational burden while maintaining or even improving the model's performance.

In conclusion, the extensive hyperparameter tuning conducted in this study has demonstrated substantial improvements in the performance of XGBoost models. The meticulous approach adopted herein provides a strong foundation for the application of XGBoost in similar predictive modeling tasks. The results advocate for continued exploration and optimization of hyperparameters to achieve the best possible model performance, ensuring that machine learning models are both effective and generalizable.

The size of the mother models indicates a compact design suitable for deployment in various computational environments without excessive resource demands. These models demonstrate robust and accurate predictive modeling for periodontal probing depth across different tooth sites. The variation in model sizes and the number of iterations highlights the differing complexities and data requirements for accurate predictions across sites. The significant reduction in RMSE values from initial to final training stages underscores the models' capability to learn and adapt effectively to the data, highlighting the consistent training methodology applied across all models.

The performance metrics indicate that all models achieve high accuracy, with RMSE values remaining low and R^2 values consistently high. This reflects the models' ability to account for the variability in periodontal probing depth and make precise predictions. The consistently low MAE and MSE values further support the models' reliability in minimizing prediction errors.

Feature importance analysis reveals the pivotal role of the directional symmetry measure (γSM) across all sites. The high gain, cover, and frequency values associated with γSM features indicate their critical influence on the models' predictive performance. This emphasizes the importance of incorporating symmetry measures in predictive models for periodontal health, as they significantly enhance model accuracy and reliability.

In conclusion, the mother models exhibit strong predictive capabilities, with high accuracy and reliable performance across different teeth sites, providing quality soft labels for the daughter models to learn from. The incorporation of symmetry measures as key features further strengthens these models, making them valuable tools for predicting periodontal probing depth and potentially guiding clinical decision-making in periodontal health management.

Daughter Models

The feature importance metrics provided are crucial for understanding the role of each feature in the model and for improving model interpretability and performance. The results showed that the soft labels provide a nuanced representation of the PPD values, capturing the inherent variability and uncertainty in the data, consistently having the highest gain across all daughter models, implying they are the most important feature in terms of improving accuracy. However, "Simulated data" has the highest cover and frequency, suggesting it is used most frequently and affects the largest portion of the

dataset during the splits, although its impact on accuracy improvement is lower than that of soft labels.

This dichotomy between the gain and the cover/frequency metrics highlights the different roles these features play within the model. While the soft labels are paramount in refining the model's predictive power, the simulated data serves as a foundational element, ensuring that the model is well-grounded and extensively informed by a wide array of data points. This balance between high-gain and high-frequency features is critical for building robust models that can generalize well to new data.

The chi-squared test results across the various sites reveal differing levels of predictive accuracy for the daughter models. The statistically significant differences were found mostly in lower PPD values, indicating potential discrepancies in model predictions for those specific categories. However, the lack of significant differences in higher PPD values and the aggregate PPD ≥ 4 suggests that the model performs better for these categories. Similar patterns are observed for sites V, MV, DL, L, and ML, where significant differences are primarily concentrated in the lower PPD value ranges, while higher values and aggregates show better alignment with the simulated data.

These findings highlight the strengths and limitations of the daughter models in predicting PPD values. In general, the daughter models demonstrate strong performance in higher PPD categories, which are clinically significant as they often indicate more severe periodontal conditions. However, the discrepancies in lower PPD values suggest areas for potential improvement in the model training process, possibly by incorporating more refined training data or enhancing the model's sensitivity to subtle variations in lower PPD values. The detailed statistical analysis emphasizes the need to address site-specific differences to enhance predictive accuracy and improve the clinical applicability of the models.

Overall, the daughter models exhibit promising accuracy and reliability, particularly in predicting higher PPD values. Future work should focus on addressing the observed discrepancies in lower PPD values to further enhance the predictive capabilities of these models, ensuring comprehensive and precise periodontal health assessments.

When assessing similarity with a high level of detail, we perform proportion tests pre- and post-imputation for each value of PPD, focusing on clinical diagnosis by comparing the values aggregated in PPD ≥ 4 pre- and post-MoDau imputation this results were presented numericaly in tables and in bar graphs.

Considering the results of the comparative analysis of values grouped into "healthy" (PPD < 4) and "unhealthy" (PPD ≥ 4) categories using the MoDau imputation, in

both categories, our method can predict similar proportions of PPD ≥ 4 compared to the simulated values, without significant over- or under-estimations in either category. These results are in disagreement with those found by other authors, where biases in "case" classification are systematic Borrell and Papapanou, 2012; Couto et al., 2018; P. I. Eke et al., 2015; Kingman et al., 2008; Kingman and Albandar, 2002; Kingman and Albandar, 2008; Papapanou, 2012; Tran et al., 2013. It is worth noting that the case definitions and the pattern of intentional data omission also varied among the referenced studies, and none of them used total unilateral omission of periodontal data, making direct comparison with our results impossible.

The pre- and post-MoDau imputation results were also compared with optimized KDE. To find the best combination of parameters we perform a grid search for this comparison, including bandwidths from 0.1 to 2.5 (step = 0.1) and adjustments from 0.5 to 2.5 (step = 0.1). The most probable kernel functions considered were: "gaussian," "epanechnikov," "rectangular," "triangular," "biweight," "cosine," and "optcosine." The best combinations identified were the KDE functions "biweight" and "cosine" (with "biweight" being more common), a bandwidth of 2, which provided a high smoothing grade, and an adjustment of 2 to fine-tune the graphical comparison of distributions. Additionally, we performed numerical computations of the difference between kernels, with respective confidence intervals, using bootstrap methods and the Kolmogorov-Smirnov (K-S) test for distribution comparisons.

Chapter 6

Conclusion

This study aimed to explore and validate the use of fuzzy of symmetry concept in half mouth complete imputation method for handling planned missing data in periodontal surveys, specifically focusing on periodontal health probing pocket depth parameter.

1. Validation of SM Function: The Symmetry Measure (SM) function grounded on fuzzy sets theory, developed in this study, was validated for evaluating periodontal probing depth symmetry, showing strong correlation with professional scores and indicating its clinical applicability behind its use in imputation methodology.

2. Performance of Daughter Models: The XGBoost daughter models exhibited high predictive accuracy, especially in higher PPD categories, though some discrepancies in lower PPD values suggest areas for model improvement.

3. Importance of Symmetry Measures: Incorporating symmetry measures, particularly the directional symmetry measure (γSM), significantly improved models accuracy and reliability.

4. Accuracy of Imputation Methods: The hot deck (H-D) imputation method effectively preserved the original statistical properties of PPD data, with no significant differences observed in median and variance values between original and imputed datasets.

5. Generation of Synthetic Data: Generating synthetic data with controlled noise was effective in the absence of new data to test the MoDau imputation model, maintaining plausibility and clinical relevance of the original data.

6. Epidemiological Applicability: The models developed, particularly those incorporating symmetry measures, have strong potential for guiding imputation methods that use partial mouth assessment in contralateral sites in periodontal health surveys.

7. Statistical Robustness: The application of statistical methods like gamlss models, k-folds cross-validation and bootstrap methods ensured robust model evaluations and reliable conclusions.

8. Feature Importance in Predictive Models: Directional symmetry measures were found to be critical for predictive performance, highlighting the need to include these features in future modelling efforts.

Contributions to the Field

This research contributes new methodologies for handling planned missing data in health sciences, validates machine learning-fuzzy symmetry approach, and demonstrates the importance of symmetry measures in periodontal contralateral imputation.

Limitations and Future Research

While the study provided significant insights, limitations include potential biases in simulated data and the need for more diverse datasets. Future research should explore different methodological approaches and broader applications in health sciences.

0pt

Bibliography

Abdella, M., & Marwala, T. (2005). The use of genetic algorithms and neural networks to approximate missing data in database. *IEEE 3rd International Conference on Computational Cybernetics, 2005. ICCC 2005.*, 207–212.

Acock, A. (1997). Working with missing data. *Family Science Review*, *10*(1), 76–102.

Agiwal, V., & Chaudhuri, S. (2024). Methods and implications of addressing missing data in health-care research. *Current Medical Issues*, *22*(1), 60–62.

Ainamo, J., Barmes, D., Beagrie, G., Cutress, T., Martin, J., & Sardo-Infirri, J. (1982). Development of the world health organization (who) community periodontal index of treatment needs (cpitn). *Int Dent J*, *32*(3), 281–91.

Alavi, D. G., BeGole, E. A., & Schneider, B. J. (1988). Facial and dental arch asymmetries in class ii subdivision malocclusion. *American Journal of Orthodontics and Dentofacial Orthopedics*, *93*(1), 38–46.

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, *78*(1), 40–64.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. *Advanced Structural Equation Modeling: Issues and Techniques*, 243–277.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40–79.

Atkins, P. (1986). *Physical chemistry*. Oxford University Press.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*(1), 5–37.

Bengio, Y. (2009). *Learning deep architectures for ai*. Now Publishers Inc.

Bengio, Y., & Gingras, F. (1996). Recurrent neural networks for missing or asynchronous data. *Advances in neural information processing systems*, 395–401.

Benigeri, M., Brodeur, J.-M., Payette, M., Charbonneau, A., & Ismaıl, A. I. (2000). Community periodontal index of treatment needs and prevalence of periodontal conditions. *Journal of clinical periodontology*, *27*(5), 308–312.

Bhattacharyya, A. K. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99–109.

Borrell, L. N., & Papapanou, P. N. (2012). Use of partial-mouth recording protocols and sample weighting to monitor prevalence of periodontitis in the united states. *Journal of Clinical Periodontology*, *39*(12), 1139–1146.

Bratthall, D., Petersen, P. E., Stjernswärd, J. R., & Brown, L. J. (2006). Oral and craniofacial diseases and disorders. *Disease Control Priorities in Developing Countries. 2nd edition.*

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society: Series B (Methodological)*, *22*(2), 302–306. https://doi.org/https://doi.org/10.1111/j.2517-6161.1960.tb00375.x

Burden, R. L., & Faires, J. D. (2010). *Numerical analysis* (9th) [Newton-Raphson Method]. Brooks Cole.

Butt, K., Butt, R., & Sharma, P. (2019). The burden of periodontal disease. *Dental Update*, *46*(10), 907–913.

Campesato, O. (2020). *Artificial intelligence, machine learning, and deep learning*. Stylus Publishing, LLC.

Carroll, S. B. (1995). Homeotic genes and the evolution of arthropods and chordates. *Nature*, *376*(6540), 479–485.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Duxbury Press.

Caton, J. G., Armitage, G., Berglundh, T., Chapple, I. L., Jepsen, S., Sanz, M., & Tonetti, M. S. (2018). A new classification scheme for periodontal and peri-implant diseases and conditions – introduction and key changes from the 1999 classification. *Journal of Clinical Periodontology*, *45*(S20), S1–S8.

Cheeseman, P. C., Self, M., Kelly, J., Taylor, W., Freeman, D., & Stutz, J. C. (1988). Bayesian classification. *AAAI*, *88*, 607–611.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Cheng, C.-H., Huang, C.-H., Ruess, H., Yasuoka, H., et al. (2018). Towards dependability metrics for neural networks. *2018 16th ACM/IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE)*, 1–4.

Choi, J., Dekkers, O. M., & le Cessie, S. (2019). A comparison of different methods to handle missing data in the context of propensity score analysis. *Eur J Epidemiol*, *34*(1), 23–36. https://doi.org/10.1007/s10654-018-0447-z

Choudhury, S. J., & Pal, N. R. (2019). Imputation of missing data with neural networks for classification. *Knowledge-Based Systems*, *182*, 104838.

Cohen, J. (1983). *Applied multiple regression: Correlation analysis for the behavioral sciences* (Report).

Cohen, J., & Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. L. Erlbaum Associates.

Copeland, B. (2020). *Artificial intelligence*. https://www.britannica.com/technology/artificial-intelligence (accessed: 03.03.2021).

Couto, P., Pereira, P. A., Nunes, M., & Mendes, R. A. (2018). Validation of a portuguese version of the oral health impact profile adapted to people with mild intellectual disabilities (ohip-14-mid-pt). *PLoS One*, *13*(6), e0198840. https://doi.org/10.1371/journal.pone.0198840

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977a). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977b). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Dye, B. A., & Thornton-Evans, G. (2007). A brief history of national surveillance efforts for periodontal disease in the united states. *Journal of periodontology*, *78*, 1373–1379.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.

Eke, P. I., Page, R. C., Wei, L., Thornton-Evans, G., & Genco, R. J. (2012). Update of the case definitions for population-based surveillance of periodontitis. *J Periodontol*, *83*(12), 1449–54. https://doi.org/10.1902/jop.2012.110664

Eke, P. I., Dye, B. A., Wei, L., Thornton-Evans, G. O., & Genco, R. J. (2012). Prevalence of periodontitis in adults in the united states: 2009 and 2010. *Journal of Dental Research*, *91*(10), 914–920.

Eke, P. I., Dye, B. A., Wei, L., Thornton-Evans, G. O., & Genco, R. J. (2015). Prevalence of periodontitis in adults in the united states: 2009 and 2010. *Journal of Dental Research*, *94*(11), 1045–1053.

Eke, P., Thornton-Evans, G., Wei, L., Borgnakke, W., & Dye, B. (2010). Accuracy of nhanes periodontal examination protocols. *Journal of dental research*, *89*(11), 1208–1213.

Enders, C. K. (2001). *The association of missing data techniques and multilevel modeling*. SAGE Publications.

Evans, M., Hastings, N., & Peacock, B. (1999). Statistical distributions. *Wiley Series in Probability and Statistics*.

Félix, C. G. A. (2015). *Una introducción a la simetría*. Editorial UNED.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*, 700–725.

Fritz, P. C., Ward, W. E., & Longo, A. B. (2018). The new global classification system for periodontal and peri-implant diseases: An executive summary for the busy dental professional. *Oral Health*.

Gad, I., Hosahalli, D., Manjunatha, B. R., & Ghoneim, O. A. (2020). A robust deep learning model for missing value imputation in big ncdc dataset. *Iran Journal of Computer Science*. https://doi.org/10.1007/s42044-020-00065-z

Garcia, L., Tasker, G., & Grauch, V. (2010). Patterns of missing data in environmental monitoring data. *Environmental Modelling & Software*, *25*(4), 354–363.

García-Laencina, P. J., Sancho-Gómez, J.-L., Figueiras-Vidal, A. R., & Verleysen, M. (2010). Pattern classification with missing data: A review. *Neural Computing and Applications*, *19*(2), 263–282.

Ghorbani, A., & Zou, J. Y. (2018). Embedding for informative missingness: Deep learning with incomplete data. *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 437–445.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*(1), 549–576.

Gupta, A., & Lam, M. S. (1996). Estimating missing values using neural networks. *The Journal of the Operational Research Society*, *47*(2), 229–238. https://doi.org/10.2307/2584344

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis*.

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1953). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *48*(261), 97–141. https://doi.org/10.1080/01621459.1953.10501117

Hardt, J., Herke, M., & Leonhart, R. (2012). Auxiliary variables in multiple imputation in regression with missing x: A warning against including too many in small sample research. *BMC medical research methodology*, *12*, 1–13.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Haykin, S. S. (1999). *Neural networks: A comprehensive foundation*. Prentice Hall.

Hertel, B. R. (1976). Minimizing error variance introduced by missing data routines in survey analysis. *Sociological Methods Research*, *4*(4), 459–474.

Holló, G. (2015). A new paradigm for animal symmetry. *Interface focus*, *5*(6), 20150032.

Huisman, M. (1999). Missing data in behavioral science research: Investigation of a collection of data sets. *M AND T SERIES*, *32*, 23–46.

IBM. (2020). *Artificial intelligence (ai)*. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence (accessed: 03.03.2021).

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Imtiaz, S., & Shah, S. (2008). Treatment of missing values in process data analysis. *The Canadian Journal of Chemical Engineering*, *86*(5), 838–858.

Jamison, D. T. (2006). *Disease control priorities in developing countries* (2nd ed.). Oxford University Press ; Washington, DC : World Bank.

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, *50*(2), 105–115.

Jönsson, P., & Wohlin, C. (2004). An evaluation of k-nearest neighbour imputation using likert data. *Proceedings - International Software Metrics Symposium*, 108–118. https://doi.org/10.1109/METRIC.2004.1357895

Joreskog, K. G., & Sorbom, D. (1993). Lisrel 8 user's guide. *Chicago: Scientific Software International*.

Josse, J., Prost, N., Scornet, E., & Varoquaux, G. (2019). Handling missing values with tree-based methods. *Journal of Machine Learning Research*, *20*(104), 1–58.

Kalton, G. (1986). The treatment of missing survey data. *Survey methodology*, *12*, 1–16.

Kingman, A., Susin, C., & Albandar, J. M. (2008). Effect of partial recording protocols on severity estimates of periodontal disease. *J Clin Periodontol*, *35*(8), 659–67.

Kingman, A., & Albandar, J. M. (1999). Methodological aspects of epidemiological studies of periodontal diseases. *Periodontology 2000*, *29*(1), 11–30.

Kingman, A., & Albandar, J. M. (2002). Methodological aspects of epidemiological studies of periodontal diseases. *Periodontology 2000*, *29*(1), 11–30.

Kingman, A., & Albandar, J. M. (2008). Partial- and full-mouth approaches to assessing the prevalence of periodontitis. *Journal of Periodontology*, *79*(1), 2079–2086.

Kleinke, K., Reinecke, J., Salfrán, D., & Spiess, M. (2020). *Applied multiple imputation.* Springer.

Köhler, A. E. (1991). A fuzzy symmetry concept for forms with imperfect symmetries. *Computers & Mathematics with Applications*, *22*(9), 35–50.

Kumutha, V., & Palaniammal, S. (2013). An enhanced approach on handling missing values using bagging k-nn imputation. *2013 International Conference on Computer Communication and Informatics*, 1–8.

Lakshminarayan, K., Harp, S. A., Goldman, R. P., & Samad, T. (1996). Imputation of missing data using machine learning techniques. *KDD*, 140–145.

Levene, H. (1960). Robust tests for equality of variances. In I. e. a. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of harold hotelling* (pp. 278–292). Stanford University Press.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data.* John Wiley & Sons.

Little, R. J. A., & Rubin, D. B. (2020). Statistical analysis with missing data. https://proxy.library.cornell.edu/sso/skillport?context=146074

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198–1202.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd). John Wiley Sons.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley Sons.

Liu, M., Li, S., Yuan, H., Ong, M. E. H., Ning, Y., Xie, F., Saffari, S. E., Shang, Y., Volovici, V., Chakraborty, B., et al. (2023). Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques. *Artificial intelligence in medicine*, *142*, 102587.

Lohr, S. L. (2019). *Sampling: Design and analysis.* Chapman; Hall/CRC.

Ma, C., & Zhang, C. (2021). Identifiable generative models for missing not at random data imputation. *Advances in Neural Information Processing Systems*, *34*, 27645–27658.

Mascarenhas, A. K., Garetto, L. P., & Johnson, G. A. (2020). *Burt and eklund's dentistry, dental practice, and the community*. Elsevier Health Sciences.

Mascarenhas, A. K., Okunseri, C., Dye, B., et al. (2020). *Burt and eklund's dentistry, dental practice, and the community-e-book*. Elsevier Health Sciences.

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

Maurice, T. J., & Kula, K. (1998). Dental arch asymmetry in the mixed dentition. *The Angle Orthodontist*, *68*(1), 37–44.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Mitchell, T. M. (1997). *Machine learning* (First). McGraw-Hill.

M'Kendrick, A. (1925). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, *44*, 98–130.

Mombelli, A., & Meier, C. (2001). On the symmetry of periodontal disease. *Journal of clinical periodontology*, *28*(8), 741–745.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Muslim, E., Nurtjahyo, B., Darwita, R. R., & Widinugroho, B. (2012). Working posture evaluation of clinical student in faculty of dentistry university of indonesia for the scaling task in sitting position in a virtual environment. *Makara Journal of Health Research*, *16*. https://doi.org/10.7454/msk.v16i1.1300

Myers, T. A. (2011). Goodbye, listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data. *Communication Methods and Measures*, *5*(4), 297–310.

Myrtveit, I., Stensrud, E., & Olsson, U. H. (2001). Analyzing data sets with missing data: An empirical evaluation of imputation methods and likelihood-based methods. *IEEE Transactions on Software Engineering*, *27*(11), 999–1013.

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372–411.

Newman, M. G., Takei, H. H., & Carranza, F. A. (2006). *Carranza's clinical periodontology* (10th ed.). Saunders/Elsevier.

Oliveira, T. (2004). *Estatística aplicada*. Universidade Aberta.

Orme, J. G., & Reis, J. (1991). Multiple regression with missing data. *Journal of Social Service Research*, *15*(1-2), 61–91.

Owens, J. D., Dowsett, S. A., Eckert, G. J., Zero, D. T., & Kowolik, M. J. (2003). Partial-mouth assessment of periodontal disease in an adult population of the united states. *J Periodontol*, *74*(8), 1206–13. https://doi.org/10.1902/jop.2003.74.8.1206

Owens, J. D., & Palmer, J. D. (1999). Epidemiology of periodontal disease: An update. *Journal of Periodontal Research*, *34*(5), 360–367.

Papapanou, P. N. (2012). Epidemiology of periodontal diseases: An update. *Journal of Clinical Periodontology*, *39*(S12), 207–210.

Pihlstrom, B. L., Michalowicz, B. S., & Johnson, N. W. (2005). Periodontal diseases. *Lancet*, *366*(9499), 1809–20.

Platias, C., & Petasis, G. (2020). A comparison of machine learning methods for data imputation. *11th Hellenic Conference on Artificial Intelligence*, 150–159.

Poulos J., V. R. (2018). Missing data imputation for supervised learning. *Applied Artificial Intelligence*, *32*(2), 186–196.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Ramfjord, S. P. (1959). Indices for prevalence and incidence of periodontal disease. *The Journal of Periodontology*, *30*(1), 51–59.

Rao, J. N. K., & Fuller, W. A. (2017). Big data for finite population inference: Applying quasi-random methods. *International Statistical Review*, *85*(2), 271–293. https://doi.org/10.1111/insr.12205

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, *47*(1), 13–26.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion).

Ritzwoller, D. M., Romano, J. P., & Shaikh, A. M. (2024). Randomization inference: Theory and applications. *arXiv:2406.09521 [econ.EM]*. https://arxiv.org/abs/2406.09521

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Roth, P. L., & Switzer III, F. S. (1995). A monte carlo analysis of missing data techniques in a hrm setting. *Journal of Management*, *21*(5), 1003–1023.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*(3), 227–241. https://doi.org/10.1177/1536867X0400400301

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987a). *Multiple imputation for nonresponse in surveys*. John Wiley Sons, Inc.

Rubin, D. B. (1987b). *Multiple imputation for nonresponse in surveys*. John Wiley Sons.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1*. MIT Press.

Sarle, W. S. (1994). Neural networks and statistical models. *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, 1538–1550.

Schafer, J. L. (1997a). *Analysis of incomplete multivariate data*. Chapman Hall/CRC.

Schafer, J. L. (1997b). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. (1999). *Analysis of incomplete multivariate data*. Chapman Hall/CRC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Scheffer, J. (2002). Dealing with missing data.

Silva-Ramírez, E.-L., Pino-Mejías, R., López-Coello, M., & Cubiles-de-la-Vega, M.-D. (2011). Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, *24*(1), 121–129.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman; Hall.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.

Srijan, S., & Rajagopalan, I. R. (2024). Best practices for handling missing data. *Annals of Surgical Oncology*, *31*(1), 12–13.

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Sun, B., Liu, L., Miao, W., Wirth, K., Robins, J., & Tchetgen, E. J. T. (2018). Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, *28*(4), 1965.

Tanner, M. A., & Wong, W. H. (1987). Calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Templ, M., & Filzmoser, P. (2008). Visualization of missing values using the r-package vim. *Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology.*

Tonetti, M. S., Greenwell, H., & Kornman, K. S. (2018). Staging and grading of periodontitis: Framework and proposal of a new classification and case definition. *J Periodontol*, *89 Suppl 1*, S159–s172.

Tran, D. T., Gay, I., Du, X. L., Fu, Y., Bebermeyer, R. D., Neumann, A. S., Streckfus, C., Chan, W., & Walji, M. F. (2013). Assessing periodontitis in populations: A systematic review of the validity of partial-mouth examination protocols. *Journal of Clinical Periodontology*, *40*(12), 1064–1071.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, *17*(6), 520–525.

Tutz, G. (2010). *Regression for categorical data.* Cambridge University Press.

Twala, B., Jones, M., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950–956.

Vahdati, A., Cotterill, S., Marsden, A., & Kontopantelis, E. (2024). Enhancing data integrity in electronic health records: Review of methods for handling missing data. *medRxiv*, 2024–05.

Van Dyk, D. A., & Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, *10*(1), 1–50.

Vapnik, V. (1995). *The nature of statistical learning theory.* Springer.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning*, 1096–1103.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American psychologist*, *54*(8), 594.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62.

Zabrodsky, H., Peleg, S., & Avnir, D. (1992). Continuous symmetry measures. *Journal of the American Chemical Society*, *114*(20), 7843–7851.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, *8*, 338–353.

Zar, J. H. (2005). *Spearman rank correlation.* Encyclopedia of biostatistics.

Zhu, Q., Chen, J., Shi, D., Zhu, L., Bai, X., Duan, X., & Liu, Y. (2019). Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction. *IEEE Transactions on Sustainable Energy*, *11*(1), 509–523.

Appendix I

Appendix: Assessing Symmetry

I.1 Upper Central Incisors - 11, 21

Table I.1: Summary of statistical tests comparing central incisors 11 and 21 PPD medians and variances across six dental sites; distances between distributions.

Site	Stats	Teeth		Test Results		K-S	Bhat.	Corr.
		11	21	Stats	p		Coef.	Coef.
DV	Median	1	1	W = 3589510	0.638	D = 0.007	1.000	0.58
	Variance	0.685	0.761	F = 0.266	0.606	p = 1		
V	Median	1	1	W = 3474941	**3.02e-03**	D = 0.028	0.999	0.64
	Variance	0.538	0.578	F = 0.065	0.799	p = 0.247		
MV	Median	1	1	W = 3354774	**5.33e-09**	D = 0.079	0.997	0.66
	Variance	0.639	0.702	F = 18.218	**2.00e-05**	p = **9.96e-08**		
DL	Median	1	1	W = 3258241	**2.35e-04**	D = 0.050	0.998	0.62
	Variance	0.729	0.807	F = 11.600	**6.65e-04**	p = **3.18e-03**		
L	Median	1	1	W = 3591491	0.267	D = 0.013	0.999	0.65
	Variance	0.731	0.726	F = 0.230	0.647	p = 0.9772		
ML	Median	2	2	W = 3405239	**5.38e-05**	D = 0.055	0.998	0.73
	Variance	0.826	0.779	F = 7.425	**6.45e-03**	p = **5.21e-04**		

Abbreviations: 11 – Upper right central incisor, 21 – Upper left central incisor; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

I.1.1 Distal Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.2: Assessing Side effect on PPD at Upper Central Incisors Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ (identity)	Intercept	0.682	**<2e-16**	0.681	**<2e-16**	0.975	0.807
	21DV	•	•	0.002	0.91		
σ (log)	Intercept	-1.112	**<2e-16**	-1.105	**<2e-16**		
	21DV	•	•	-0.015	0.698		
ν (log)	Intercept	-0.324	**<2e-16**	-0.337	**<2e-16**		
	21DV	•	•	0.025	0.483		
	n	5172		5172			
	D. F.	3		6			
	Res. D. F.	5169		5166			
	G. D.	10948.65		10947.68			
	AIC	10954.65		10959.68			
	SBC	10974.3		10998.98			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.3: Statistical Comparison of Cross Validation Metrics for Models: PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.115	1.115
Paired t-test for RMSE		
t-value		2.164
Degrees of Freedom		4
p-value		0.096
95% CI		**[-2.2e-05, 1.78e-04]**
Mean Difference		7.8e-05
Mean MAE	0.793	0.793
Paired t-test for MAE		
t-value		2.839
Degrees of Freedom		4
p-value		0.047
95% CI		**[3e-06, 2.47e-04]**
Mean Difference		1.25e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.1.2 Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.4: Assessing Side effect on PPD at Upper Central Incisors Vestibular Sites Using GAMLSS Null (PPD$\sim$ 1) and Side (PPD$\sim$ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD $\sim$ 1 Estimate	PPD $\sim$ 1 Pr($>\|t\|$)	PPD $\sim$ Side Estimate	PPD $\sim$ Side Pr($>\|t\|$)	LRT χ^2	LRT Pr($>\chi^2$)
μ	Intercept	0.363	$<$ **2e-16**	0.347	$<$ **2e-16**		
(identity)	21V	•	•	0.033	0.273		
σ	Intercept	-0.693	$<$ **2e-16**	-0.702	$<$ **2e-16**	8.947	**0.03**
(log)	21V	•	•	0.014	0.715		
ν	Intercept	-0.630	$<$ **2e-16**	-0.658	$<$ **2e-16**		
(log)	21V	•	•	0.052	0.365		
	n	5170		5170			
	D. F.	3		6			
	Res. D. F.	5167		5164			
	G. D.	11017.96		11009.01			
	AIC	11023.96		11021.01			
	SBC	11043.61		11060.32			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.5: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD $\sim$ 1 and PPD $\sim$ Side

	PPD $\sim$ 1	PPD $\sim$ Side
Mean RMSE	0.921	0.919
Paired t-test for RMSE		
t-value		4.637
Degrees of Freedom		4
p-value		**9.76e-03**
95% CI		[**4.8e-04**, **1.92e-03**]
Mean Difference		1.2e-03
Mean MAE	0.743	0.742
Paired t-test for MAE		
t-value		4.045
Degrees of Freedom		4
p-value		**0.016**
95% CI		[**3.3e-04**, **1.78e-03**]
Mean Difference		1.06e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.1.3 Mesial Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.6: Assessing Side effect on PPD at Upper Central Incisors Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.671	< **2e-16**	0.663	< **2e-16**		
(identity)	21MV	•	•	0.021	0.270		
σ	Intercept	-1.088	< **2e-16**	-1.105	< **2e-16**	25.989	**9.59e-06**
(log)	21MV	•	•	0.039	0.295		
ν	Intercept	-0.378	< **2e-16**	-0.446	< **2e-16**		
(log)	21MV	•	•	0.125	**5.93e-05**		
	n	5168		5168			
	D. F.	3		6			
	Res. D. F.	5165		5162			
	G. D.	10667.79		10641.8			
	AIC	10673.79		10653.8			
	SBC	10693.44		10693.1			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.7: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.070	1.067
Paired t-test for RMSE		
t-value		6.538
Degrees of Freedom		4
p-value		2.80e-03
95% CI		[**1.2e-03**, **3.0e-03**]
Mean Difference		2.1e-03
Mean MAE	0.763	0.761
Paired t-test for MAE		
t-value		7.388
Degrees of Freedom		4
p-value		1.8e-03
95% CI		[**1.5e-03**, **3.2e-03**]
Mean Difference		2.3e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.1.4 Distal Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.8: Assessing Side effect on PPD at Upper Central Incisors Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.703	< **2e-16**	0.696	< **2e-16**		
(identity)	21DL	•	•	0.018	0.422		
σ	Intercept	-1.048	< **2e-16**	-1.067	< **2e-16**	12.879	**4.91e-03**
(log)	21DL	•	•	0.043	0.36		
ν	Intercept	-0.216	< **2e-16**	-0.261	< **2e-16**		
(log)	21DL	•	•	0.083	**0.024**		
	n	5238		5238			
	D. F.	3		6			
	Res. D. F.	5235		5232			
	G. D.	12025.03		12012.15			
	AIC	12031.03		12024.15			
	SBC	12050.72		12063.53			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.9: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.188	1.187
Paired t-test for RMSE		
t-value		12.242
Degrees of Freedom		4
p-value		**2.6e-04**
95% CI		[**1.2e-03**, **1.9e-03**]
Mean Difference		1.5e-03
Mean MAE	0.871	0.870
Paired t-test for MAE		
t-value		11.323
Degrees of Freedom		4
p-value		**3.5e-04**
95% CI		[**1.3e-03**, **2.1e-03**]
Mean Difference		1.7e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.1.5 Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.10: Assessing Side effect on PPD at Upper Central Incisors Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ (identity)	Intercept	0.479	< **2e-16**	0.469	< **2e-16**	1.381	0.710
	21L	•	•	0.019	0.471		
σ (log)	Intercept	-0.752	< **2e-16**	-0.758	< **2e-16**		
	21L	•	•	0.013	0.729		
ν (log)	Intercept	-0.416	< **2e-16**	-0.420	< **2e-16**		
	21L	•	•	0.008	0.866		
	n	5168		5168			
	D. F.	3		6			
	Res. D. F.	5165		5162			
	G. D.	11798.54		11797.16			
	AIC	11804.54		11809.16			
	SBC	11824.19		11848.46			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.11: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.077	1.077
Paired t-test for RMSE		
t-value		2.776
Degrees of Freedom		4
p-value		**0.05**
95% CI		[**7.5e-07**, **7.9e-04**]
Mean Difference		3.94e-04
Mean MAE	0.828	0.828
Paired t-test for MAE		
t-value		3.782
Degrees of Freedom		4
p-value		0.0194
95% CI		[**9.52e-05**, **6.21e-04**]
Mean Difference		3.58e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.1.6 Mesial Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.12: Assessing Side Effect on PPD at Upper Central Incisors Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD ∼ Side Estimate	PPD ∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.805	**< 2e-16**	0.787	**< 2e-16**		
(identity)	21ML	•	•	0.228	**2.72e-10**		
σ	Intercept	-1.365	**< 2e-16**	-1.290	**< 2e-16**	18.96	**2.79e-04**
(log)	21ML	•	•	0.454	**4.85e-10**		
ν	Intercept	-0.037	0.025	-0.060	**0.011**		
(log)	11ML	•	•	-0.172	**7.86e-04**		
	n	5175		5175			
	D. F.	3		6			
	Res. D. F.	5172		5169			
	G. D.	12381.25		12400.21			
	AIC	12387.25		12412.21			
	SBC	12406.91		12451.52			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.13: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.317	1.260
Paired t-test for RMSE		
t-value		4.125
Degrees of Freedom		4
p-value		**0.015**
95% CI		[**0.019**, **0.095**]
Mean Difference		0.057
Mean MAE	0.986	0.914
Paired t-test for MAE		
t-value		4.173
Degrees of Freedom		4
p-value		**0.014**
95% CI		[**0.024**, **0.120**]
Mean Difference		0.072

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.2 Upper Lateral Incisors - 12, 22

Table I.14: Summary of statistical tests comparing lateral incisors 12 and 22 PPD medians and variances across six dental sites; distances between distributions

Site	Stats	Teeth		Test Results		K-S	Bhat.	Cor.
		12	22	Stat	p		Coef.	Coef.
DV	Median Variance	1 0.800	1 0.676	W = 2979273 F = 0.489	**8.61e-03** 0.485	D = 0.040 p = **0.036**	1.000	0.55
V	Median Variance	1 0.546	1 0.515	W = 2922411 F = 10.961	**1.45e-04** **9.37e-04**	D = 0.059 p = **3.38e-04**	0.998	0.59
MV	Median Variance	1 0.707	1 0.796	W = 3088533 F = 1.856	0.883 0.173	D = 0.010 p = 1	1.000	0.57
DL	Median Variance	1 0.794	1 0.845	W = 3103952 F = 0.010	0.847 0.921	D = 8.44e-03 p = 1	1.000	0.57
L	Median Variance	1 0.671	1 0.713	W = 3051359 F = 0.379	0.329 0.538	D = 0.010 p = 1	0.999	0.62
ML	Median Variance	1 0.812	1 0.760	W = 3065477 F = 0.024	0.515 0.877	D = 0.008 p = 1	1.000	0.62

Abbreviations: 12 – Upper right lateral incisor, 22 – Upper left lateral incisor; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

I.2.1 Distal Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.15: Assessing Side Effect on PPD at Upper Lateral Incisors Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution.

Parameter (link)		PPD ∼ 1) Estimate	PPD ∼ 1) Pr(> |t|)	PPD ∼ Side Estimate	PPD ∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.700	< **2e-16**	0.686	< **2e-16**		
(identity)	22DV	•	•	0.031	0.085		
σ	Intercept	-1.175	< **2e-16**	-1.180	< **2e-16**	3.46	0.326
(log)	22DV	•	•	0.011	0.777		
ν	Intercept	-0.304	< **2e-16**	-0.300	< **2e-16**		
(log)	22DV	•	•	-0.011	0.76		
	n	4985		4985			
	D. F.	3		6			
	Res. D. F.	4982		4979			
	G. D.	10491.53		10488.07			
	AIC	10497.53		10500.07			
	SBC	10517.07		10539.15			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.16: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.133	1.133
Paired t-test for RMSE		
t-value		2.289
Degrees of Freedom		4
p-value		0.084
95% CI		**[−1e−04, 1.3e−03]**
Mean Difference		0.0006
Mean MAE	0.796	0.7957
Paired t-test for MAE		
t-value		1.907
Degrees of Freedom		4
p-value		0.129
95% CI		[−4e−04, 1.9e−03]
Mean Difference		8e−04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.2.2 Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.17: Assessing Side Effect on PPD at Upper Lateral Incisors Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ (identity)	Intercept	0.316	**<2e-16**	0.252	**<2e-16**	21.351	**8.9e-05**
	22V	•	•	0.122	**7.42e-05**		
σ (log)	Intercept	-0.704	**<2e-16**	-0.729	**<2e-16**		
	22V	•	•	0.037	0.358		
ν (log)	Intercept	-0.659	**<2e-16**	-0.601	**<2e-16**		
	22V	•	•	-0.111	0.065		
	n	4979		4979			
	D. F.	3		6			
	Res. D. F.	4976		4973			
	G. D.	10422.79		10401.44			
	AIC	10428.79		10413.44			
	SBC	10448.33		10452.52			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.18: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	0.894	0.895
Paired t-test for RMSE		
t-value		-0.553
Degrees of Freedom		4
p-value		0.610
95% CI		**[-3.34e-03, 2.23e-03]**
Mean Difference		-5.54e-04
Mean MAE	0.723	0.7201
Paired t-test for MAE		
t-value		4.268
Degrees of Freedom		4
p-value		**0.013**
95% CI		**[1.11e-03, 5.22e-03]**
Mean Difference		3.16e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.2.3 Mesial Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.19: Assessing Side Effect on PPD at Upper Lateral Incisors Mesio Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.691	**<2e-16**	0.702	**<2e-16**		
(identity)	22MV	•	•	-0.023	0.220		
σ	Intercept	-1.120	**<2e-16**	-1.160	**<2e-16**	7.70	0.052
(log)	22MV	•	•	0.076	0.061		
ν	Intercept	-0.267	**<2e-16**	-0.293	**<2e-16**		
(log)	22MV	•	•	0.051	0.153		
	n	4982		4982			
	D. F.	3		6			
	Res. D. F.	4979		4976			
	G. D.	10896.99		10889.29			
	AIC	10902.99		10901.29			
	SBC	10922.53		10940.37			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.20: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.156	1.156
Paired t-test for RMSE		
t-value		-0.541
Degrees of Freedom		4
p-value		0.617
95% CI		**[-6.9e-04, 4.6e-04]**
Mean Difference		-1.1e-04
Mean MAE	0.829	0.829
Paired t-test for MAE		
t-value		0.557
Degrees of Freedom		4
p-value		0.607
95% CI		**[-5.0e-04, 7.5e-04]**
Mean Difference		1.3e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.2.4 Distal Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.21: Assessing Side Effect on PPD at Upper Lateral Incisors Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ (identity)	Intercept	0.738	**<2e-16**	0.739	**<2e-16**	0.299	0.960
	22DL	•	•	-0.001	0.964		
σ (log)	Intercept	-1.136	**<2e-16**	-1.148	**<2e-16**		
	22DL	•	•	0.024	0.636		
ν (log)	Intercept	-0.160	**<2e-16**	-0.160	**<2e-16**		
	22DL	•	•	-0.002	0.961		
	n	5009		5009			
	D. F.	3		6			
	Res. D. F.	5006		5003			
	G. D.	11620.21		11619.91			
	AIC	11626.21		11631.91			
	SBC	11645.77		11671.02			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.22: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.241	1.241
Paired t-test for RMSE		
t-value		1.565
Degrees of Freedom		4
p-value		0.193
95% CI		**[-5e-05, 1.7e-04]**
Mean Difference		6.2e-05
Mean MAE	0.901	0.901
Paired t-test for MAE		
t-value		1.836
Degrees of Freedom		4
p-value		0.140
95% CI		**[-5e-05, 2.4e-04]**
Mean Difference		9.4e-05

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.2.5 Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.23: Assessing Side Effect on PPD at Upper Lateral Incisors Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.498	**<2e-16**	0.498	**<2e-16**		
(identity)	22L	•	•	4e-04	0.99		
σ	Intercept	-0.741	**<2e-16**	-0.736	**<2e-16**	1.583	0.663
(log)	22L	•	•	-0.009	0.811		
ν	Intercept	-0.445	**<2e-16**	-0.467	**<2e-16**		
(log)	22L	•	•	0.042	0.378		
	n	4978		4978			
	D. F.	3		6			
	Res. D. F.	4975		4972			
	G. D.	11263.59		11262.01			
	AIC	11269.59		11274.01			
	SBC	11289.13		11313.09			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.24: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.05	1.05
Paired t-test for RMSE		
t-value		0.637
Degrees of Freedom		4
p-value		0.559
95% CI		**[-3.4e-04, 5.5e-04]**
Mean Difference		1.02e-04
Mean MAE	0.813	0.8129
Paired t-test for MAE		
t-value		-0.194
Degrees of Freedom		4
p-value		0.856
95% CI		**[-4.4e-04, 3.9e-04]**
Mean Difference		-2.9e-05

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.2.6 Mesial Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.25: Assessing Side Effect on PPD at Upper Lateral Incisors Mesio Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.695	**<2e-16**	0.689	**<2e-16**		
(identity)	22ML	•	•	0.011	0.614		
σ	Intercept	-1.036	**<2e-16**	-1.039	**<2e-16**	0.316	0.957
(log)	22ML	•	•	0.006	0.903		
ν	Intercept	-0.216	**<2e-16**	-0.215	**<2e-16**		
(log)	22ML	•	•	-0.004	0.915		
	n	5052		5052			
	D. F.	3		6			
	Res. D. F.	5049		5046			
	G. D.	11636.13		11635.81			
	AIC	11642.13		11647.81			
	SBC	11661.71		11686.98			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.26: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.197	1.197
Paired t-test for RMSE		
t-value		2.061
Degrees of Freedom		4
p-value		0.108
95% CI		**[-1.1e-04, 7.1e-04]**
Mean Difference		3.03e-04
Mean MAE	0.874	0.874
Paired t-test for MAE		
t-value		2.281
Degrees of Freedom		4
p-value		0.085
95% CI		**[-9e-05, 9.5e-04]**
Mean Difference		4.29e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.3 Upper Canines - 13, 23

Table I.27: Summary of statistical tests comparing canines 13 and 23 PPD medians and variances across six dental sites; distances between distributions.

Site	Stats	Teeth		Test Results		K-S	Bhat.	Cor.
		13	**23**	**Stat**	**p**		**Coef.**	**Coef.**
DV	Median	1	1	W = 2979273	**8.61e-03**	D = 0.040	1.000	0.45
	Variance	0.800	0.676	F = 0.489	0.485	p = **0.036**		
V	Median	1	1	W = 2922411	**1.45e-04**	D = 0.059	0.998	0.51
	Variance	0.546	0.515	F = 10.961	**9.37e-04**	p = 3.38e-04		
MV	Median	1	1	W = 3088533	0.883	D = 0.010	1.000	0.48
	Variance	0.707	0.796	F = 1.856	0.173	p = 1		
DL	Median	1	1	W = 3103952	0.847	D = 0.008	1.000	0.48
	Variance	0.794	0.845	F = 0.010	0.921	p = 1		
L	Median	1	1	W = 3051359	0.329	D = 0.010	0.999	0.56
	Variance	0.671	0.713	F = 0.379	0.538	p = 1		
ML	Median	1	1	W = 3065477	0.515	D = 0.008	1.000	0.55
	Variance	0.812	0.760	F = 0.024	0.877	p = 1		

Abbreviations: 13 – Upper right canine, 23 – Upper left canine; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

I.3.1 Distal Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.28: Assessing Side Effect on PPD at Upper Canines Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.772	**<2e-16**	0.776	**<2e-16**		
(identity)	23DV	•	•	-0.003	0.835		
σ	Intercept	-1.529	**<2e-16**	-1.577	**<2e-16**	47.78	**2.37e-10**
(log)	23DV	•	•	0.101	**0.007**		
ν	Intercept	-0.318	**<2e-16**	-0.417	**<2e-16**		
(log)	23DV	•	•	0.184	**5.41e-09**		
	n	5257		5257			
	D. F.	3		6			
	Res. D. F.	5254		5251			
	G. D.	10005.82		9958.04			
	AIC	10011.82		9970.04			
	SBC	10031.52		10009.44			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.29: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.093	1.092
Paired t-test for RMSE		
t-value		16.599
Degrees of Freedom		4
p-value		**7.72e-05**
95% CI		[**1.11e-03**, **1.56e-03**]
Mean Difference		1.33e-03
Mean MAE	0.753	0.751
Paired t-test for MAE		
t-value		9.532
Degrees of Freedom		4
p-value		**6.76e-04**
95% CI		[**1.49e-03**, **2.71e-03**]
Mean Difference		2.1e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.3.2 Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.30: Assessing Side Effect on PPD at Upper Canines Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.314	**<2e-16**	0.261	**<2e-16**		
(identity)	23V	•	•	0.105	**1.27e-04**		
σ	Intercept	-0.730	**<2e-16**	-0.740	**<2e-16**	35.66	**8.85e-08**
(log)	23V	•	•	8.9e-03	0.800		
ν	Intercept	-0.689	**<2e-16**	-0.688	**<2e-16**		
(log)	23V	•	•	-5e-04	0.992		
	n	5266		5266			
	D. F.	3		6			
	Res. D. F.	5263		5260			
	G. D.	10744.37		10708.71			
	AIC	10750.37		10720.71			
	SBC	10770.07		10760.12			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.31: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	0.879	0.877
Paired t-test for RMSE		
t-value		2.296
Degrees of Freedom		4
p-value		0.083
95% CI		[-3.6e-04, 3.8e-04]
Mean Difference		1.72e-03
Mean MAE	0.706	0.702
Paired t-test for MAE		
t-value		6.498
Degrees of Freedom		4
p-value		**2.89e-03**
95% CI		**[2.27e-03, 5.66e-03]**
Mean Difference		3.96e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.3.3 Mesial Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.32: Assessing Side Effect on PPD at Upper Canines Mesio-Vestibular Sites Using GAMLSS Null (PPD$\sim$ 1) and Side (PPD$\sim$ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD $\sim$ 1 Estimate	PPD $\sim$ 1 Pr($>\|t\|$)	PPD$\sim$ Side Estimate	PPD$\sim$ Side Pr($>\|t\|$)	LRT χ^2	LRT Pr($>\chi^2$)
μ	Intercept	0.723	**<2e-16**	0.721	**<2e-16**		
(identity)	23MV	•	•	0.003	0.884		
σ	Intercept	-1.180	**<2e-16**	-1.173	**<2e-16**	0.299	0.960
(log)	23MV	•	•	-0.013	0.752		
ν	Intercept	-0.291	**<2e-16**	-0.284	**<2e-16**		
(log)	23MV	•	•	-0.013	0.705		
	n	5262		5262			
	D. F.	3		6			
	Res. D. F.	5259		5256			
	G. D.	11142.71		11142.41			
	AIC	11148.71		11154.41			
	SBC	11168.42		11193.82			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.33: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD $\sim$ 1 and PPD $\sim$ Side

	PPD $\sim$ 1	PPD $\sim$ Side
Mean RMSE	1.123	1.123
Paired t-test for RMSE		
t-value		0.149
Degrees of Freedom		4
p-value		0.889
95% CI		[-9.62e-05, 1.07e-04]
Mean Difference		5.46e-06
Mean MAE	0.805	0.805
Paired t-test for MAE		
t-value		-0.093
Degrees of Freedom		4
p-value		0.930
95% CI		[-1.53e-04, 1.43e-04]
Mean Difference		-4.95$e-06$

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.3.4 Distal Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.34: Assessing Side Effect on PPD at Upper Canines Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.774	**<2e-16**	0.785	**<2e-16**		
(identity)	23DL	•	•	-0.015	0.340		
σ	Intercept	-1.387	**<2e-16**	-1.343	**<2e-16**	32.075	**5.05e-07**
(log)	23DL	•	•	-0.078	0.107		
ν	Intercept	-0.059	**1.82e-04**	0.005	0.813		
(log)	23DL	•	•	-0.138	**1.24e-05**		
	n	5335		5335			
	D. F.	3		6			
	Res. D. F.	5332		5329			
	G. D.	12559.69		12527.61			
	AIC	12565.69		12539.61			
	SBC	12585.44		12579.11			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.35: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.329	1.326
Paired t-test for RMSE		
t-value		6.568
Degrees of Freedom		4
p-value		**2.8e-03**
95% CI		[**1.4e-03**, **3.5e-03**]
Mean Difference		2.4e-03
Mean MAE	0.969	0.966
Paired t-test for MAE		
t-value		5.176
Degrees of Freedom		4
p-value		6.6e-03
95% CI		[**1.3e-03**, **4.4e-**]
Mean Difference		2.9e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.3.5 Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.36: Assessing Side Effect on PPD at Upper Canines Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.575	**<2e-16**	0.567	**<2e-16**		
(identity)	23L	•	•	0.015	0.564		
σ	Intercept	-0.795	**<2e-16**	-0.782	**<2e-16**	1.588	0.662
(log)	23L	•	•	-0.025	0.523		
ν	Intercept	-0.417	**<2e-16**	-0.418	**<2e-16**		
(log)	23L	•	•	0.002	0.962		
	n	5256		5256			
	D. F.	3		6			
	Res. D. F.	5253		5250			
	G. D.	11792.9		11791.32			
	AIC	11798.9		11803.32			
	SBC	11818.61		11842.72			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.37: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.080	1.080
Paired t-test for RMSE		
t-value		-1.064
Degrees of Freedom		4
p-value		0.347
95% CI		[-3.7e-04, 1.7e-04]
Mean Difference		-1.0e-04
Mean MAE	0.805	0.805
Paired t-test for MAE		
t-value		-0.736
Degrees of Freedom		4
p-value		0.503
95% CI		[-5.1e-04, 3.0e-04]
Mean Difference		-1.1e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.3.6 Mesial Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.38: Assessing Side Effect on PPD at Upper Canines Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.746	**<2e-16**	0.751	**<2e-16**		
(identity)	23ML	•	•	-0.003	0.868		
σ	Intercept	-1.244	**<2e-16**	-1.149	**<2e-16**	24.169	**2.3e-05**
(log)	23ML	•	•	-0.179	**1.56e-04**		
ν	Intercept	-0.158	**<2e-16**	-0.127	**<2e-16**		
(log)	23ML	•	•	-0.073	**0.030**		
	n	5315		5315			
	D. F.	3		6			
	Res. D. F.	5312		5309			
	G. D.	12062.4		12038.23			
	AIC	12068.4		12050.23			
	SBC	12088.13		12089.7			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.39: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.245	1.242
Paired t-test for RMSE		
t-value		4.025
Degrees of Freedom		4
p-value		**0.016**
95% CI		[**9e-04**, **5.1e-03**]
Mean Difference		0.003
Mean MAE	0.894	0.890
Paired t-test for MAE		
t-value		4.191
Degrees of Freedom		4
p-value		**0.014**
95% CI		[**1.4e-03**, **6.8e-03**]
Mean Difference		0.004

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.4 Upper First Premolar - 14, 24

Table I.40: Summary of Statistical Tests Comparing Upper First Premolar 14, 24 PPD medians and variances across six dental sites; distances between distributions

Site	Stats	Teeth		Test Results		K-S	Bhat.	Cor.
		14	**24**	**Stat**	**p**		**Coef.**	**Coef.**
DV	Median	1	1	W = 2168638	**6.64e-15**	D = 0.119	0.993	0.58
	Variance	0.552	0.717	F = 58.151	**2.95e-14**	p = **3.55e-14**		
V	Median	1	1	W = 2332176	**4.99e-04**	D = 0.043	1.000	0.64
	Variance	0.394	0.502	F = 5.329	**0.021**	p = **0.034**		
MV	Median	1	1	W = 2421311	0.319	D = 0.013	1.000	0.66
	Variance	0.726	0.729	F = 0.972	0.324	p = 0.991		
DL	Median	2	2	W = 2720906	**3.39e-11**	D = 0.099	0.993	0.62
	Variance	0.788	0.751	F = 6.350	**0.012**	p = **5.47e-10**		
L	Median	1	1	W = 2345284	**1.93e-03**	D = 0.042	0.999	0.65
	Variance	0.605	0.666	F = 4.103	**0.043**	p = **0.041**		
ML	Median	2	1	W = 2771129	**1.34e-15**	D = 0.120	0.991	0.73
	Variance	0.873	0.691	F = 20.377	**6.53e-06**	p = **2.20e-14**		

Abbreviations: 14 – Upper right premolar, 24 – Upper left premolar; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

I.4.1 Distal Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.41: Assessing Side Effect on PPD at Upper First Premolar Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.748	**<2e-16**	0.756	**<2e-16**		
(identity)	24DV	•	•	-0.001	0.940		
σ	Intercept	-1.329	**<2e-16**	-1.420	**<2e-16**	79.809	**<2.2e-16**
(log)	24DV	•	•	0.205	**6.0e-06**		
ν	Intercept	-0.323	**<2e-6**	-0.457	**<2e-16**		
(log)	24DV	•	•	0.236	**1.2e-10**		
	n	4437		4437			
	D. F.	3		6			
	Res. D. F.	4434		4431			
	G. D.	8833.044		8753.235			
	AIC	8839.044		8765.235			
	SBC	8858.237		8803.621			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.42: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.079	1.074
Paired t-test for RMSE		
t-value		24.066
Degrees of Freedom		4
p-value		**1.77e-05**
95% CI		[**4.1e-03**, **5.1e-03**]
Mean Difference		0.005
Mean MAE	0.764	0.758
Paired t-test for MAE		
t-value		19.018
Degrees of Freedom		4
p-value		**4.50e-05**
95% CI		[**5.6e-03**, **7.5e-03**]
Mean Difference		0.007

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.4.2 Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.43: Assessing Side Effect on PPD at Upper First Premolar Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.440	**<2e-16**	0.460	**<2e-16**		
(identity)	24V	•	•	-0.030	0.302		
σ	Intercept	-0.722	**<2e-16**	-0.717	**<2e-16**	27.614	**4.38e-06**
(log)	24V	•	•	-0.007	0.847		
ν	Intercept	-0.836	**<2e-16**	-0.988	**<2e-16**		
(log)	24V	•	•	0.262	**9.26e-05**		
	n	4441		4441			
	D. F.	3		6			
	Res. D. F.	4438		4435			
	G. D.	8593.765		8566.151			
	AIC	8599.765		8578.151			
	SBC	8618.96		8616.542			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.44: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	0.799	0.796
Paired t-test for RMSE		
t-value		2.762
Degrees of Freedom		4
p-value		0.051
95% CI		[**-1e-05**, **4.3e-03**]
Mean Difference		2.14e-03
Mean MAE	0.665	0.663
Paired t-test for MAE		
t-value		3.294
Degrees of Freedom		4
p-value		0.030
95% CI		[3.4e-04, 3.98e-03]
Mean Difference		2.16e-03

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.4.3 Mesial Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.45: Assessing Side Effect on PPD at Upper First Premolar Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.769	**<2e-16**	0.769	**<2e-16**		
(identity)	24MV	•	•	0.000	0.987		
σ	Intercept	-1.339	**<2e-16**	-1.323	**<2e-16**	1.479	0.687
(log)	24MV	•	•	-0.033	0.504		
ν	Intercept	-0.187	**<2e-16**	-0.203	**<2e-16**		
(log)	24MV	•	•	0.032	0.374		
	n	4439		4439			
	D. F.	3		6			
	Res. D. F.	4436		4433			
	G. D.	9673.91		9672.431			
	AIC	9679.91		9684.431			
	SBC	9699.105		9722.82			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.46: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.189	1.189
Paired t-test for RMSE		
t-value		2.247
Degrees of Freedom		4
p-value		0.088
95% CI		[-4.3e-05, 4.08e-04]
Mean Difference		1.82e-04
Mean MAE	0.861	0.861
Paired t-test for MAE		
t-value		2.423
Degrees of Freedom		4
p-value		0.073
95% CI		[-3.5e-05, 5.19e-04]
Mean Difference		2.42e-04

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.4.4 Distal Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.47: Assessing Side Effect on PPD at Upper First Premolar Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.837	**<2e-16**	1.111	**<2e-16**	89.332	**<2.2e-16**
(identity)	24DL	•	•	-0.281	**<2e-16**		
σ	Intercept	-1.867	**<2e-16**	-0.787	**<2e-16**		
(log)	24DL	•	•	-1.095	**<2e-16**		
ν	Intercept	-0.030	0.062	-0.258	**4.36e-11**		
(log)	24DL	•	•	0.154	**0.001**		
	n	4439		4439			
	D. F.	3		6			
	Res. D. F.	4436		4433			
	G. D.	9968.246		10057.58			
	AIC	9974.246		10069.58			
	SBC	9993.44		10107.97			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.48: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.310	1.213
Paired t-test for RMSE		
t-value		54.985
Degrees of Freedom		4
p-value		**6.55e-07**
95% CI		[**0.092**, **0.101**]
Mean Difference		0.096
Mean MAE	0.978	0.885
Paired t-test for MAE		
t-value		85.879
Degrees of Freedom		4
p-value		**1.10e-7**
95% CI		[**0.090**, **0.096**]
Mean Difference		0.093

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.4.5 Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.49: Assessing Side Effect on PPD at Upper First Premolar Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.589	**<2e-16**	0.578	**<2e-16**		
(identity)	24L	•	•	0.024	0.377		
σ	Intercept	-0.830	**<2e-16**	-0.835	**<2e-16**	9.091	**0.028**
(log)	24L	•	•	0.011	0.796		
ν	Intercept	-0.463	**<2e-16**	-0.502	**<2e-16**		
(log)	24L	•	•	0.072	0.130		
	n	4436		4436			
	D. F.	3		6			
	Res. D. F.	4433		4430			
	G. D.	9591.53		9582.44			
	AIC	9597.53		9594.44			
	SBC	9616.723		9632.825			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.50: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.016	1.015
Paired t-test for RMSE		
t-value		4.025
Degrees of Freedom		4
p-value		**0.016**
95% CI		[**3e-04**, **1.9e-03**]
Mean Difference		0.001
Mean MAE	0.770	0.769
Paired t-test for MAE		
t-value		3.103
Degrees of Freedom		4
p-value		**0.036**
95% CI		[**1e-04**, **2.1e-03**]
Mean Difference		0.001

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.4.6 Mesial Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.51: Assessing Side Effect on PPD at Upper First Premolar Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.761	**<2e-16**	0.788	**<2e-16**		
(identity)	24ML	•	•	-0.040	**0.043**		
σ	Intercept	-1.297	**<2e-16**	-1.247	**<2e-16**	55.471	**5.45e-12**
(log)	24ML	•	•	-0.076	0.172		
ν	Intercept	-0.124	**<2e-16**	-0.045	0.085		
(log)	24ML	•	•	-0.184	**3.71e-07**		
	n	4463		4463			
	D. F.	3		6			
	Res. D. F.	4460		4457			
	G. D.	10231.98		10176.51			
	AIC	10237.98		10188.51			
	SBC	10257.19		10226.93			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.52: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.253	1.246
Paired t-test for RMSE		
t-value		5.333
Degrees of Freedom		4
p-value		**0.006**
95% CI		[**3.2e-03**, **1.03e-02**]
Mean Difference		6.8e-03
Mean MAE	0.916	0.909
Paired t-test for MAE		
t-value		4.534
Degrees of Freedom		4
p-value		**0.011**
95% CI		[**2.8e-03**, **0.012**]
Mean Difference		0.007

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.5 Upper Second Premolars - 15, 25

Table I.53: Summary of statistical tests comparing Upper Second Premolars 15 and 25 PPD medians and variances across six dental sites; distances between distributions.

Site	Stats	Teeth		Test Results		K-S	Bhat.	Cor.
		15	**25**	**Stat**	**p**		**Coef.**	**Coef.**
DV	Median	1	1	W = 2066818	< **2.2e-16**	D = 0.165	0.987	0.55
	Variance	0.536	0.783	F = 114.920	< **2.2e-16**	p = < **2.2e-16**		
V	Median	1	1	W = 2293838	**4.87e-07**	D = 0.048	0.997	0.59
	Variance	0.393	0.483	F = 0.946	0.331	p = **1.29e-02**		
MV	Median	1	2	W = 2359115	**2.28e-03**	D = 0.053	0.999	0.57
	Variance	0.627	0.661	F = 4.914	**0.027**	p = **3.45e-03**		
DL	Median	2	2	W = 2825090	< **2.2e-16**	D = 0.127	0.987	
	Variance	0.832	0.727	F = 6.902	**8.64e-03**	p = **4.44e-16**		0.57
L	Median	1	1	W = 2445321	0.391	D = 0.026	1.000	0.62
	Variance	0.653	0.608	F = 1.906	0.167	p = 0.437		
ML	Median	2	2	W = 2883271	< **2.2e-16**	D = 0.135	0.980	0.62
	Variance	0.787	0.663	F = 3.579	0.057	p = < **2.2e-16**		

Abbreviations: 11 – Upper right second premolar, 21 – Upper left second premolar; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

I.5.1 Distal Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.54: Assessing Side Effect on PPD at Upper Second Premolars Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.752	**<2e-16**	0.764	**<2e-16**		
(identity)	25DV	•	•	0.003	0.88		
σ	Intercept	-1.366	**<2e-16**	-1.481	**<2e-16**	147.22	**<2.2e-16**
(log)	25DV	•	•	0.273	**1.81e-08**		
ν	Intercept	-0.271	**<2e-16**	-0.462	**<2e-16**		
(log)	25DV	•	•	0.323	**<2e-16**		
	n	4454		4454			
	D. F.	3		6			
	Res. D. F.	4451		4448			
	G. D.	9115.988		8968.767			
	AIC	9121.988		8980.767			
	SBC	9141.192		9019.176			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.55: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.120	1.112
Paired t-test for RMSE		
t-value		11.928
Degrees of Freedom		4
p-value		**2.83e-04**
95% CI		[**6.8e-03**, **0.011**]
Mean Difference		0.009
Mean MAE	0.798	0.786
Paired t-test for MAE		
t-value		11.678
Degrees of Freedom		4
p-value		**3.08e-04**
95% CI		[**9.3e-03**, **0.015**]
Mean Difference		0.012

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.5.2 Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.56: Assessing Side Effect on PPD at Upper Second Premolars Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.469	**<2e-16**	0.464	**<2e-16**		
(identity)	25V	•	•	0.018	0.524		
σ	Intercept	-0.725	**<2e-16**	-0.722	**<2e-16**	32.167	**4.84e-07**
(log)	25V	•	•	-0.006	0.863		
ν	Intercept	-0.891	**<2e-16**	-1.012	**<2e-16**		
(log)	25V	•	•	0.208	**0.002**		
	n	4455		4455			
	D. F.	3		6			
	Res. D. F.	4452		4449			
	G. D.	8431.154		8398.987			
	AIC	8437.154		8410.987			
	SBC	8456.359		8449.398			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.57: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	0.78	0.777
Paired t-test for RMSE		
t-value		14.455
Degrees of Freedom		4
p-value		**1.33e-04**
95% CI		[**2.39e-03**, **3.53e-03**]
Mean Difference		2.96e-03
Mean MAE	0.647	0.645
Paired t-test for MAE		
t-value		15.762
Degrees of Freedom		4
p-value		**9.47e-05**
95% CI		[**2.17e-03**, **3.09e-03**]
Mean Difference		0.003

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.5.3 Mesial Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.58: Assessing Side Effect on PPD at Upper Second Premolars Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.785	**<2e-16**	0.772	**<2e-16**		
(identity)	25MV	•	•	0.174	**3.5e-04**		
σ	Intercept	-1.244	**<2e-16**	-1.316	**<2e-16**	23.899	**3.0e-05**
(log)	25MV	•	•	0.495	**<2e-16**		
ν	Intercept	-0.225	**<2e-16**	-0.245	**<2e-16**		
(log)	25MV	•	•	-0.159	**0.005**		
	n	4457		4457			
	D. F.	3		6			
	Res. D. F.	4454		4451			
	G. D.	9656.685		9632.786			
	AIC	9662.685		9644.786			
	SBC	9681.892		9683.200			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.59: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.132	1.098
Paired t-test for RMSE		
t-value		3.351
Degrees of Freedom		4
p-value		**0.029**
95% CI		[**0.005.9e-**, **0.063**]
Mean Difference		0.034
Mean MAE	0.834	0.786
Paired t-test for MAE		
t-value		3.235
Degrees of Freedom		4
p-value		**0.032**
95% CI		[**6.9e-03**, **8.98e-02**]
Mean Difference		0.0483

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.5.4 Distal Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.60: Assessing Side Effect on PPD at Upper Second Premolars Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.821	**<2e-16**	1.123	**<2e-16**		
(identity)	25DL	•	•	-0.319	**<2e-16**		
σ	Intercept	-1.657	**<2e-16**	-0.763	**<2e-16**	8.384	**0.039**
(log)	25DL	•	•	-0.850	**<2e-16**		
ν	Intercept	-0.022	0.177	-0.243	**1.02e-10**		
(log)	25DL	•	•	0.125	**0.005**		
	n	4461		4461			
	D. F.	3		6			
	Res. D. F.	4458		4455			
	G. D.	10332.43		10324.04			
	AIC	10338.43		10336.04			
	SBC	10357.64		10374.46			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.61: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.322	1.218
Paired t-test for RMSE		
t-value		42.638
Degrees of Freedom		4
p-value		**1.81e-06**
95% CI		[**0.098**, **0.112**]
Mean Difference		0.105
Mean MAE	0.990	0.892
Paired t-test for MAE		
t-value		86.421
Degrees of Freedom		4
p-value		**1.08e-07**
95% CI		[**0.095**, **0.101**]
Mean Difference		0.098

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.5.5 Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.62: Assessing Side Effect on PPD at Upper Second Premolars Lingual Sites: Null Model vs Side Model Using Ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.647	**<2e-16**	0.637	**<2e-16**		
(identity)	25L	•	•	0.033	0.267		
σ	Intercept	-0.835	**<2e-16**	-0.903	**<2e-16**	10.575	**0.014**
(log)	25L	•	•	0.148	**0.002**		
ν	Intercept	-0.453	**<2e-16**	-0.439	**<2e-16**		
(log)	25L	•	•	-0.048	0.351		
	n	4454		4454			
	D. F.	3		6			
	Res. D. F.	4451		4448			
	G. D.	9650.548		9639.974			
	AIC	9656.548		9651.974			
	SBC	9675.753		9690.383			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.63: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.118	1.170
Paired t-test for RMSE		
t-value		-19.777
Degrees of Freedom		4
p-value		**3.86e-05**
95% CI		[**-0.060**, **-0.045**]
Mean Difference		-0.053
Mean MAE	0.807	0.882
Paired t-test for MAE		
t-value		-36.707
Degrees of Freedom		4
p-value		**3.29e-06**
95% CI		[**-0.081**, **-0.069**]
Mean Difference		-0.075

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.5.6 Mesial Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.64: Assessing Side Effect on PPD at Upper Second Premolars Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter		PPD ∼ 1		PPD∼ Side		LRT	
(link)		Estimate	Pr(> |t|)	Estimate	Pr(> |t|)	χ^2	Pr(> χ^2)
μ	Intercept	1.101	**<2e-16**	1.215	**<2e-16**	180.48	**<2e-16**
(identity)	25ML	•	•	-0.402	**<2e-16**		
σ	Intercept	-0.760	**<2e-16**	-0.712	**<2e-16**		
(log)	25ML	•	•	-0.760	**<2e-16**		
ν	Intercept	-0.338	**<2e-16**	-0.321	**<2e-16**		
(log)	25ML	•	•	0.188	**3.03e-05**		
	n	4456		4456			
	D. F.	3		6			
	Res. D. F.	4453		4450			
	G. D.	10456.57		10276.09			
	AIC	10462.57		10288.09			
	SBC	10481.78		10326.5			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.65: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.118	1.170
Paired t-test for RMSE		
t-value		-19.777
Degrees of Freedom		4
p-value		**3.86e-05**
95% CI		[**-0.060**, **-0.045**]
Mean Difference		-0.053
Mean MAE	0.807	0.882
Paired t-test for MAE		
t-value		-36.707
Degrees of Freedom		4
p-value		**3.29e-06**
95% CI		[**-0.081**, **-0.069**]
Mean Difference		-0.075

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.6 Upper First Molars - 16, 26

Table I.66: Summary of Statistical Tests Comparing Upper First Molars 16 and 26 PPD Medians and Variances Across Six Dental Sites; Distances Between Distributions

Site	Stats	Teeth		Test Results		K-S	Bhat.	Cor.
		16	**26**	**Stat**	**p**		**Coef.**	**Coef.**
DV	Median	1	2	W = 1661474	**<2.2e-16**	D = 0.178	0.984	0.48
	Variance	0.798	1.12	F = 21.548	**3.56e-06**	p = **<2.2e-16**		
V	Median	1	1	W = 1882430	**2.95e-06**	D = 0.061	0.999	0.50
	Variance	0.451	0.601	F = 9.084	**2.60e-03**	p = **1.01e-03**		
MV	Median	1	2	W = 1874751	**4.02e-06**	D = 0.069	0.997	0.45
	Variance	0.642	0.775	F = 4.038	**4,46-02**	p = **1.37-04**		
DL	Median	2	2	W = 2317438	**<2.2e-16**	D = 0.128	0.988	0.47
	Variance	0.990	0.956	F = 4.150	**4.17e-02**	p = **1.17e-14**		
L	Median	1	1	W = 2089879	0.062	D = 0.022	0.999	0.49
	Variance	0.483	0.506	F = 0.01	0.921	p = 0.696		
ML	Median	2	2	W = 2487157	**<2.2e-16**	D = 0.204	0.963	0.43
	Variance	0.826	0.740	F = 16.52	**4.89e-05**	p = **<2.2e-16**		

Abbreviations: 16 – Upper right first molar, 21 – Upper left first molar; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

I.6.1 Distal Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.67: Assessing Side Effect on PPD at Upper First Molars Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.748	**<2e-16**	0.737	**<2e-16**		
(identity)	26DV	•	•	0.215	**4.64e-12**		
σ	Intercept	-1.171	**<2e-16**	-1.336	**<2e-16**	182.27	**<2.2e-16**
(log)	26DV	•	•	0.644	**<2e-16**		
ν	Intercept	-0.037	0.057	-0.190	**2.43e-13**		
(log)	26DV	•	•	0.095	**0.034**		
	n	4042		4042			
	D. F.	3		6			
	Res. D. F.	4039		4036			
	G. D.	10028.27		9846.001			
	AIC	10034.27		9858.001			
	SBC	10053.18		9895.828			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.68: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.382	1.309
Paired t-test for RMSE		
t-value		38.438
Degrees of Freedom		4
p-value		**2.74e-06**
95% CI		[**0.068**, **0.078**]
Mean Difference		0.073
Mean MAE	1.004	0.913
Paired t-test for MAE		
t-value		46.676
Degrees of Freedom		4
p-value		**1.26e-06**
95% CI		[**0.085**, **0.096**]
Mean Difference		0.090

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.6.2 Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.69: Assessing Side Effect on PPD at Upper First Molars Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.582	**<2e-16**	0.578	**<2e-16**		
(identity)	26V	•	•	0.015	0.622		
σ	Intercept	-0.765	**<2e-16**	-0.762	**<2e-16**	28.104	**3.45e-06**
(log)	26V	•	•	-0.004	0.929		
ν	Intercept	-0.656	**<2e-16**	-0.763	**<2e-16**		
(log)	26V	•	•	0.189	**0.002**		
	n	4033		4033			
	D. F.	3		6			
	Res. D. F.	4030		4027			
	G. D.	8208.392		8180.288			
	AIC	8214.392		8192.288			
	SBC	8233.299		8230.102			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.70: Statistical Comparison of Cross Validation Metrics for Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	0.893	0.890
Paired t-test for RMSE		
t-value		28.13
Degrees of Freedom		4
p-value		**9.50e-06**
95% CI		[**0.003**, **0.004**]
Mean Difference		0.003
Mean MAE	0.697	0.694
Paired t-test for MAE		
t-value		6.842
Degrees of Freedom		4
p-value		**0.002**
95% CI		[**0.002**, **0.005**]
Mean Difference		0.003

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.6.3 Mesial Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.71: Assessing Side Effect on PPD at Upper First Molars Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.787	**<2e-16**	0.781	**<2e-16**		
(identity)	26MV	•	•	0.151	**8.34e-05**		
σ	Intercept	-1.275	**<2e-16**	-1.375	**<2e-16**	38.831	**1.88e-08**
(log)	26MV	•	•	0.539	**1.52e-12**		
ν	Intercept	-0.156	**1.03e-15**	-0.221	**<2e-16**		
(log)	26MV	•	•	-0.040	0.483		
	n	4033		4033			
	D. F.	3		6			
	Res. D. F.	4030		4027			
	G. D.	9078.97		9040.139			
	AIC	9084.97		9052.139			
	SBC	9103.877		9089.952			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.72: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.200	1.165
Paired t-test for RMSE		
t-value		3.470
Degrees of Freedom		4
p-value		**0.026**
95% CI		[**0.007**, **0.064**]
Mean Difference		0.036
Mean MAE	0.887	0.840
Paired t-test for MAE		
t-value		3.413
Degrees of Freedom		4
p-value		**0.027**
95% CI		[**0.009**, **0.086**]
Mean Difference		0.047

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.6.4 Distal Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.73: Assessing Side Effect on PPD at Upper First Molars Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD ∼ Side Estimate	PPD ∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	1.140	< **2e-16**	1.273	< **2e-16**		
(identity)	26DL	•	•	-0.309	**7.08e-06**	89.999	< **2.2e-16**
σ	Intercept	-0.709	< **2e-16**	-0.674	< **2e-16**		
(log)	26DL	•	•	-0.220	0.117		
ν	Intercept	-0.152	**3.46e-09**	-0.177	**6.97e-07**		
(log)	26DL	•	•	0.098	0.220		
	n	4036		4036			
	D. F.	3		6			
	Res. D. F.	4033		4030			
	G. D.	10521.95		10431.95			
	AIC	10527.95		10443.95			
	SBC	10546.86		10481.77			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.74: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.312	1.330
Paired t-test for RMSE		
t-value		-1.570
Degrees of Freedom		4
p-value		0.192
95% CI		[-0.049, 0.013]
Mean Difference		-0.018
Mean MAE	0.965	0.976
Paired t-test for MAE		
t-value		-0.688
Degrees of Freedom		4
p-value		0.529
95% CI		[−0.056, 0.034]
Mean Difference		-0.011

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.6.5 Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.75: Assessing Side Effect on PPD at Upper First Molars Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	0.734	**<2e-16**	0.735	**<2e-16**		
(identity)	26L	•	•	0.003	0.907		
σ	Intercept	-0.948	**<2e-16**	-1.056	**<2e-16**	20.209	**1.5e-04**
(log)	26L	•	•	0.199	**4.0e-04**		
ν	Intercept	-0.521	**<2e-16**	-0.485	**<2e-16**		
(log)	26L	•	•	-0.080	0.156		
	n	4031		4031			
	D. F.	3		6			
	Res. D. F.	4028		4025			
	G. D.	8019.846		7999.637			
	AIC	8025.846		8011.637			
	SBC	8044.752		8049.448			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.76: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	0.920	0.919
Paired t-test for RMSE		
t-value		1.393
Degrees of Freedom		4
p-value		0.236
95% CI		[-0.002, 0.005]
Mean Difference		0.002
Mean MAE	0.685	0.682
Paired t-test for MAE		
t-value		1.333
Degrees of Freedom		4
p-value		0.253
95% CI		[-0.002, 0.007]
Mean Difference		0.002

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.6.6 Mesial Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.77: Assessing Side Effect on PPD at Upper First Molars Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	1.131	**<2e-16**	1.322	**<2e-16**		
(identity)	26ML	•	•	-0.526	**<2e-16**		
σ	Intercept	-0.706	**<2e-16**	-0.664	**<2e-16**	289.46	**<2.2e-16**
(log)	26ML	•	•	-0.745	**<2e-16**		
ν	Intercept	-0.278	**<2e-16**	-0.301	**<2e-16**		
(log)	26ML	•	•	0.217	**2.96e-06**		
	n	4032		4032			
	D. F.	3		6			
	Res. D. F.	4029		4026			
	G. D.	9928.875		9639.415			
	AIC	9934.875		9651.415			
	SBC	9953.781		9689.227			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.78: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.177	1.216
Paired t-test for RMSE		
t-value		-21.930
Degrees of Freedom		4
p-value		**2.59e-05**
95% CI		[**-0.043**, **-0.034**]
Mean Difference		-0.039
Mean MAE	0.868	0.926
Paired t-test for MAE		
t-value		-58.649
Degrees of Freedom		4
p-value		**5.06e-07**
95% CI		[**-0.061**, **-0.055**]
Mean Difference		-0.058

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.7 Upper Second Molars - 17, 27

Table I.79: of Statistical Tests Comparing Upper Second Molars 17 and 27 PPD Medians and Variances Across Six Dental Sites; Distances Between Distributions

Site	Stats	Teeth		Test Results		K-S	Bhat.	Cor.
		17	27	Stat	p		Coef.	Coef.
DV	Median	1	2	W = 1850993	**6.36e-10**	D = 0.084	0.997	0.52
	Variance	0.645	0.906	F = 21.365	**3.91e-06**	p = **1.31e-06**		
V	Median	1	1	W = 1925833	**4.10e-05**	D = 0.058	0.997	0.46
	Variance	0.578	0.629	F = 7.675	**5.63e-03**	p = **2.36e-03**		
MV	Median	2	2	W = 1757368	**<2.2e-16**	D = 0.117	0.987	0.47
	Variance	0.759	1.05	F = 1.241	0.265	p = **1.52e-12**		
DL	Median	2	2	W = 2343649	**6.78e-16**	D = 0.115	0.988	0.49
	Variance	1.20	1.05	F = 0.138	0.710	p = **1.21e-11**		
L	Median	1	1	W = 2115555	0.105	D = 0.020	1.000	0.47
	Variance	0.645	0.589	F = 3.020	0.082	p = 0.826		
ML	Median	2	2	W = 2382157	**< 2.2e-16**	D = 0.134	0.982	0.49
	Variance	0.927	0.869	F = 0.780	0.377	p = **3.33e-16**		

Abbreviations: 17 – Upper right second molar, 27 – Upper left second molar; Stats – Statistics, W – Wilcoxon test; F – F-test; K-S – Kolmogorov-Smirnov test; D – Distance measure; Bhat. Coef. – Bhattacharyya Coefficient; Corr. Coef. – Correlation Coefficient; p – p-value

I.7.1 Distal Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.80: Assessing Side Effect on PPD at Upper Second Molars Distal Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.806	< **2e-16**	0.821	< **2e-16**		
(identity)	27DV	•	•	-0.023	0.086		
σ	Intercept	-1.710	< **2e-16**	-1.832	< **2e-16**	71.466	**2.07e-15**
(log)	27DV	•	•	0.237	**1.01e-06**		
ν	Intercept	-0.152	< **2e-16**	-0.279	< **2e-16**		
(log)	27DV	•	•	0.236	**7.17e-12**		
	n	4072		4072			
	D. F.	3		6			
	Res. D. F.	4069		4066			
	G. D.	8516.055		8444.589			
	AIC	8522.055		8456.589			
	SBC	8540.991		8494.460			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.81: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.232	1.231
Paired t-test for RMSE		
t-value		3.949
Degrees of Freedom		4
p-value		**0.017**
95% CI		[**3e-04**, **1.7e-03**]
Mean Difference		0.001
Mean MAE	0.873	0.870
Paired t-test for MAE		
t-value		6.214
Degrees of Freedom		4
p-value		**0.003**
95% CI		[**1.5e-03**, **4.0e-03**]
Mean Difference		0.003

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.7.2 Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.82: Assessing Side Effect on PPD at Upper Second Molars Vestibular Sites Using GAMLSS Null (PPD$\sim$ 1) and Side (PPD$\sim$ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD $\sim$ 1		PPD$\sim$ Side		LRT	
		Estimate	Pr($>\|t\|$)	Estimate	Pr($>\|t\|$)	χ^2	Pr($>\chi^2$)
μ	Intercept	0.681	$<$ **2e-16**	0.661	$<$ **2e-16**		
(identity)	27V	•	•	0.045	0.147		
σ	Intercept	-0.854	$<$ **2e-16**	-0.858	$<$ **2e-16**	16.90	**7.40e-05**
(log)	27V	•	•	0.011	0.837		
ν	Intercept	-0.444	$<$ **2e-16**	-0.488	$<$ **2e-16**		
(log)	27V	•	•	0.080	0.137		
	n	4063		4063			
	D. F.	3		6			
	Res. D. F.	4060		4057			
	G. D.	8767.603		8751.707			
	AIC	8773.603		8763.707			
	SBC	8792.533		8801.565			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.83: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD $\sim$ 1 and PPD $\sim$ Side

	PPD $\sim$ 1	PPD $\sim$ Side
Mean RMSE	1.009	1.007
Paired t-test for RMSE		
t-value		5.655
Degrees of Freedom		4
p-value		**4.8e-03**
95% CI		[**1.3e-03**, **3.7e-03**]
Mean Difference		0.003
Mean MAE	0.755	0.752
Paired t-test for MAE		
t-value		4.560
Degrees of Freedom		4
p-value		**0.010**
95% CI		[**1.1e-03**, **4.4e-03**]
Mean Difference		0.003

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.7.3 Mesial Vestibular Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.84: Assessing Side Effect on PPD at Upper Second Molars Mesial Vestibular Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.804	< **2e-16**	0.794	< **2e-16**		
(identity)	27MV	•	•	0.199	**1.18e-04**		
σ	Intercept	-1.508	< **2e-16**	-1.511	< **2e-16**	31.65	**6.20e-07**
(log)	27MV	•	•	0.630	**1.24e-08**		
ν	Intercept	0.030	0.085	-0.099	**7.44e-05**		
(log)	27MV	•	•	0.073	0.228		
	n	4071		4071			
	D. F.	3		6			
	Res. D. F.	4068		4065			
	G. D.	9970.845		9939.195			
	AIC	9976.845		9951.195			
	SBC	9995.78		9989.065			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.85: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.408	1.350
Paired t-test for RMSE		
t-value		4.295
Degrees of Freedom		4
p-value		**0.013**
95% CI		[**0.021**, **0.097**]
Mean Difference		0.059
Mean MAE	1.047	0.975
Paired t-test for MAE		
t-value		4.235
Degrees of Freedom		4
p-value		**0.013**
95% CI		[**0.025**, **0.120**]
Mean Difference		0.072

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.7.4 Distal Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.86: Assessing Side Effect on PPD at Upper Second Molars Distal Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	1.105	< **2e-16**	1.216	< **2e-16**	103.57	< **2.2e-16**
(identity)	27DL	•	•	-0.400	< **2e-16**		
σ	Intercept	-0.735	< **2e-16**	-0.687	< **2e-16**		
(log)	27DL	•	•	-0.734	< **2e-16**		
ν	Intercept	-0.047	0.057	-0.035	0.306		
(log)	27DL	•	•	0.147	**4.86e-04**		
	n	4070		4070			
	D. F.	3		6			
	Res. D. F.	4067		4064			
	G. D.	11051.04		10947.47			
	AIC	11057.04		10959.47			
	SBC	11075.97		10997.34			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.87: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.432	1.466
Paired t-test for RMSE		
t-value		-2.223
Degrees of Freedom		4
p-value		0.090
95% CI		[-0.078, 0.009]
Mean Difference		-0.035
Mean MAE	1.035	1.080
Paired t-test for MAE		
t-value		-2.105
Degrees of Freedom		4
p-value		0.103
95% CI		[-0.105, 0.014]
Mean Difference		-0.045

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.7.5 Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.88: Assessing Side Effect on PPD at Upper Second Molars Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	χ^2	LRT Pr(> χ^2)
μ	Intercept	0.811	< **2e-16**	0.7987	< **2e-16**		
(identity)	27L	•	•	0.080	0.155		
σ	Intercept	-1.057	< **2e-16**	-1.113	< **2e-16**	5.853	0.119
(log)	27L	•	•	0.228	0.060		
ν	Intercept	-0.262	< **2e-16**	-0.215	**4.05e-10**		
(log)	27L	•	•	-0.171	**0.044**		
	n	4061		4061			
	D. F.	3		6			
	Res. D. F.	4058		4055			
	G. D.	9006.994		9001.14			
	AIC	9012.994		9013.14			
	SBC	9031.921		9050.995			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.89: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.092	1.087
Paired t-test for RMSE		
t-value		0.750
Degrees of Freedom		4
p-value		0.495
95% CI		[-0.013, 0.022]
Mean Difference		0.005
Mean MAE	0.804	0.796
Paired t-test for MAE		
t-value		0.958
Degrees of Freedom		4
p-value		0.392
95% CI		[-0.016, 0.033]
Mean Difference		0.008

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.7.6 Mesial Lingual Sites

GAMLSS Models, Likelihood-Ratio Test and 5-Fold Cross Validation Results

Table I.90: Assessing Side Effect on PPD at Upper Second Molars Mesial Lingual Sites Using GAMLSS Null (PPD∼ 1) and Side (PPD∼ Side) Models and Likelihood Ratio Test with ex-Gaussian Distribution

Parameter (link)		PPD ∼ 1 Estimate	PPD ∼ 1 Pr(> |t|)	PPD∼ Side Estimate	PPD∼ Side Pr(> |t|)	LRT χ^2	LRT Pr(> χ^2)
μ	Intercept	1.315	< **2e-16**	1.434	< **2e-16**		
(identity)	27ML	•	•	-0.239	**8.87e-12**		
σ	Intercept	-0.677	< **2e-16**	-0.718	< **2e-16**	107.71	< **2e-16**
(log)	27ML	•	•	0.020	0.681		
ν	Intercept	-0.214	< **2e-16**	-0.205	**1.25e-09**		
(log)	27ML	•	•	-0.015	0.769		
	n	4064		4064			
	D. F.	3		6			
	Res. D. F.	4061		4058			
	G. D.	10432.96		10325.25			
	AIC	10438.96		10337.25			
	SBC	10457.89		10375.11			

Abbreviations: ex-Gaussian – Exponential Gaussian; LRT – Likelihood Ratio Test; μ – Location; σ – Scale; ν – Skewness; n – Number of observations; D. F. – Degrees of Freedom; Res. D. F. – Residual degrees of freedom; G. D. – Global Deviance; AIC – Akaike Information Criterion; SBC – Schwarz Bayesian Criterion

Table I.91: Statistical Comparison of Cross Validation Metrics for Predictive Models PPD ∼ 1 and PPD ∼ Side

	PPD ∼ 1	PPD ∼ Side
Mean RMSE	1.254	1.248
Paired t-test for RMSE		
t-value		3.182
Degrees of Freedom		4
p-value		**0.034**
95% CI		[**7e-04**, **1.07e-02**]
Mean Difference		0.006
Mean MAE	0.966	0.951
Paired t-test for MAE		
t-value		8.338
Degrees of Freedom		4
p-value		**0.001**
95% CI		[0.010, 0.020]
Mean Difference		0.015

Abbreviations: 95% CI – 95% Confidence Interval; RMSE – Root Mean Square Error; MAE – Mean Absolute Error

I.8 Descriptive SM for Pairs of Contralateral Sites of Upper Teeth

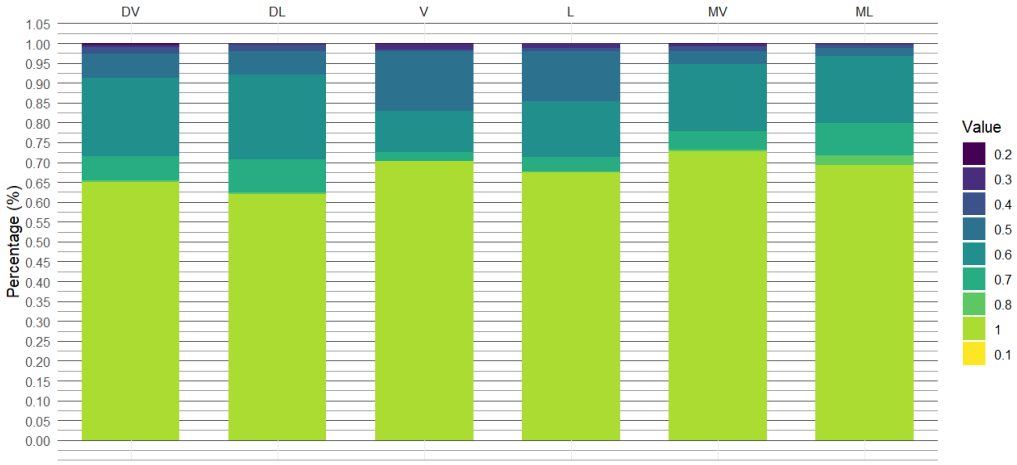
Table I.92: Summary of SM values for upper contralateral teeth pairs across six dental sites.

		Pairs of Contralateral Teeth						
Site	Parameter	(11,21)	(12,22)	(13,23)	(14,24)	(15,25)	(16,26)	(17,27)
DV	Mean	0.85	0.85	0.85	0.82	0.80	0.78	0.83
	Median	1	1	1	1	1	1	1
	Variance	0.042	0.042	0.042	0.048	0.055	0.055	0.046
V	Mean	0.86	0.84	0.84	0.84	0.83	0.83	0.81
	Median	1	1	1	1	1	1	1
	Variance	0.047	0.051	0.051	0.053	0.065	0.059	0.055
MV	Mean	0.89	0.84	0.84	0.82	0.80	0.80	0.78
	Median	1	1	1	1	1	1	1
	Variance	0.036	0.043	0.043	0.048	0.051	0.054	0.051
DL	Mean	0.85	0.84	0.84	0.81	0.78	0.77	0.76
	Median	1	1	1	1	0.7	0.7	0.7
	Variance	0.041	0.043	0.043	0.044	0.052	0.051	0.055
L	Mean	0.86	0.85	0.85	0.85	0.82	0.83	0.81
	Median	1	1	1	1	1	1	1
	Variance	0.046	0.048	0.048	0.048	0.051	0.058	0.046
ML	Mean	0.89	0.85	0.85	0.82	0.80	0.78	0.80
	Median	1	1	1	1	1	1	1
	Variance	0.032	0.041	0.041	0.045	0.050	0.055	0.052

I.9 Upper Central Incisors (11, 21)

Symmetry Measure Bar Plots

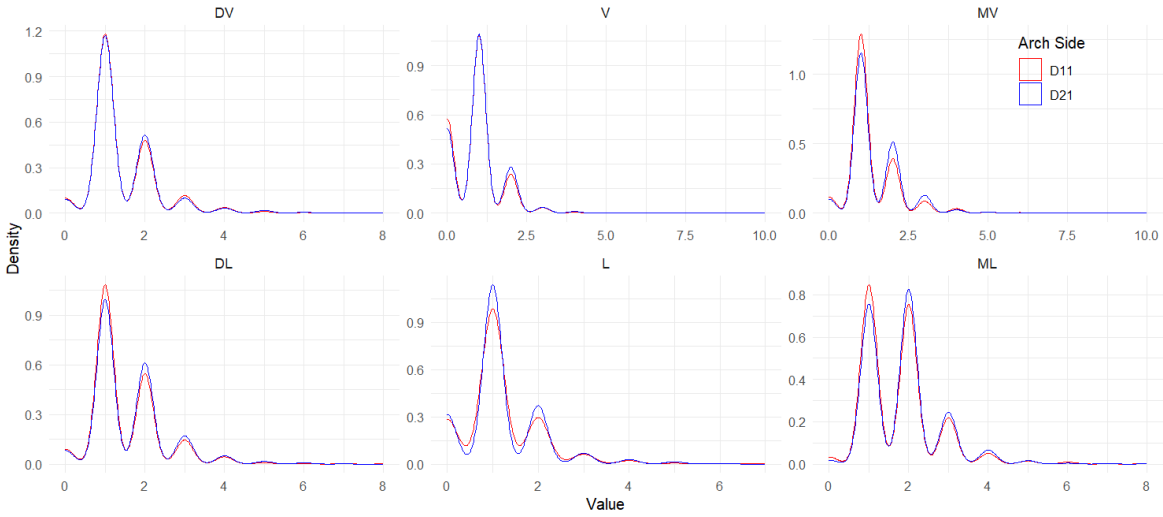
Figure I.1



Upper Central Incisors Percentages of SM Values by Site

Kernel Density Plots Estimates

Figure I.2

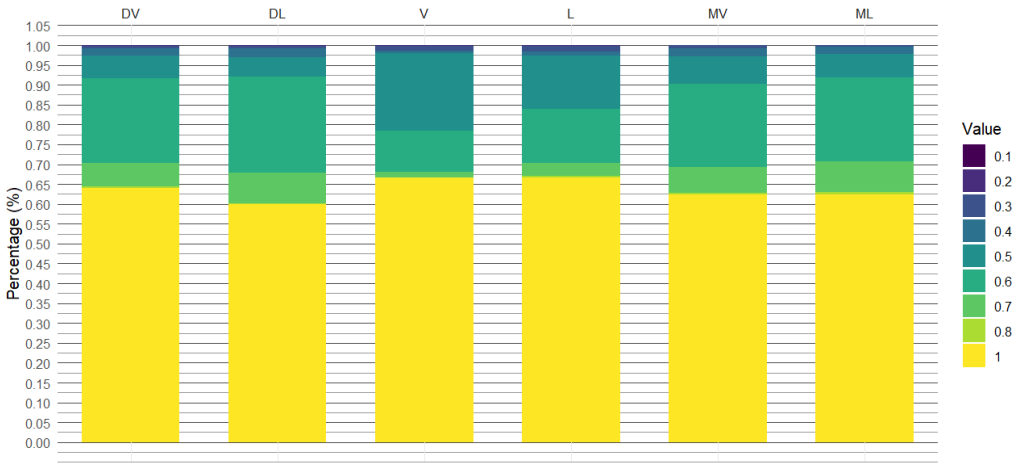


Kernel Density Plots Estimates of Upper Lateral Incisors by Site

I.10 Upper Lateral Incisors (12, 22)

Symmetry Measure Bar Plots

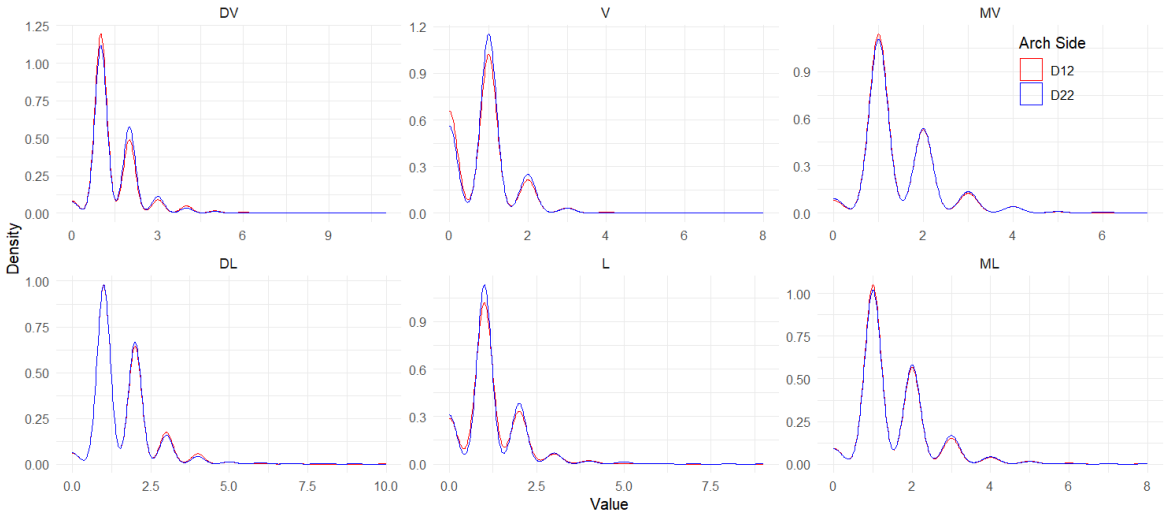
Figure I.3



Upper Lateral Incisors Percentages of SM Values by Site

Kernel Density Plots Estimates

Figure I.4

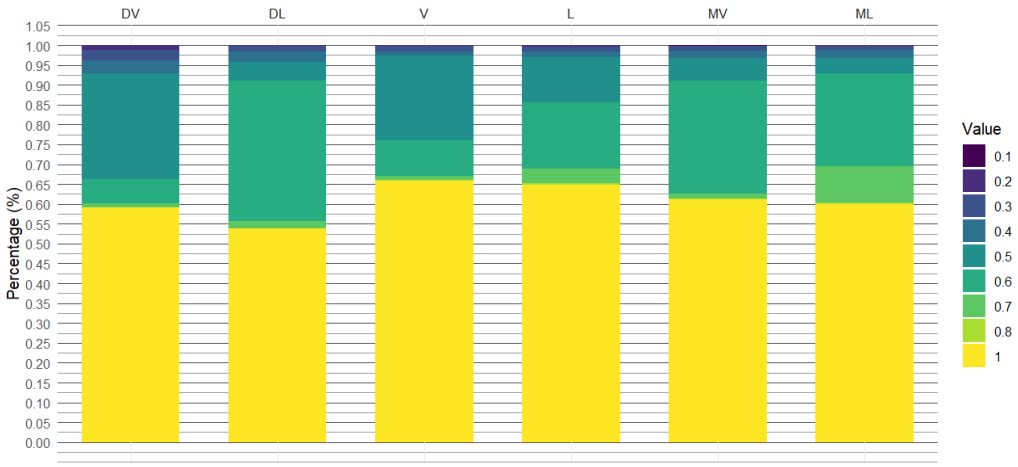


Kernel Density Plots Estimates of Upper Lateral Incisors by Site

I.11 Upper Canines (13, 23)

Symmetry Measure Bar Plots

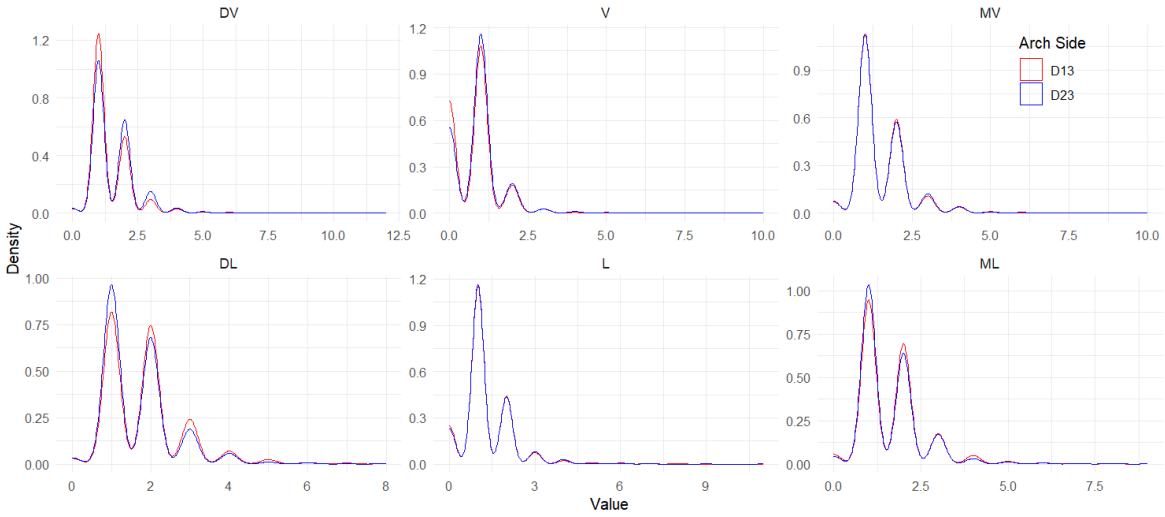
Figure I.5



Upper Canines Percentages of SM Values by Site

Kernel Density Plots Estimates

Figure I.6

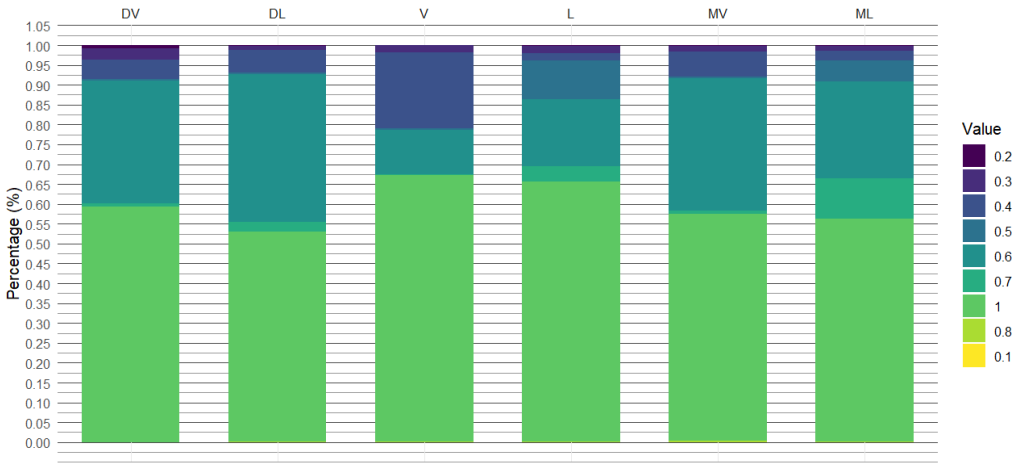


Kernel Density Plots Estimates of Upper Canines by Site

I.12 Upper First Premolars (14, 24)

Symmetry Measure Bar Plots

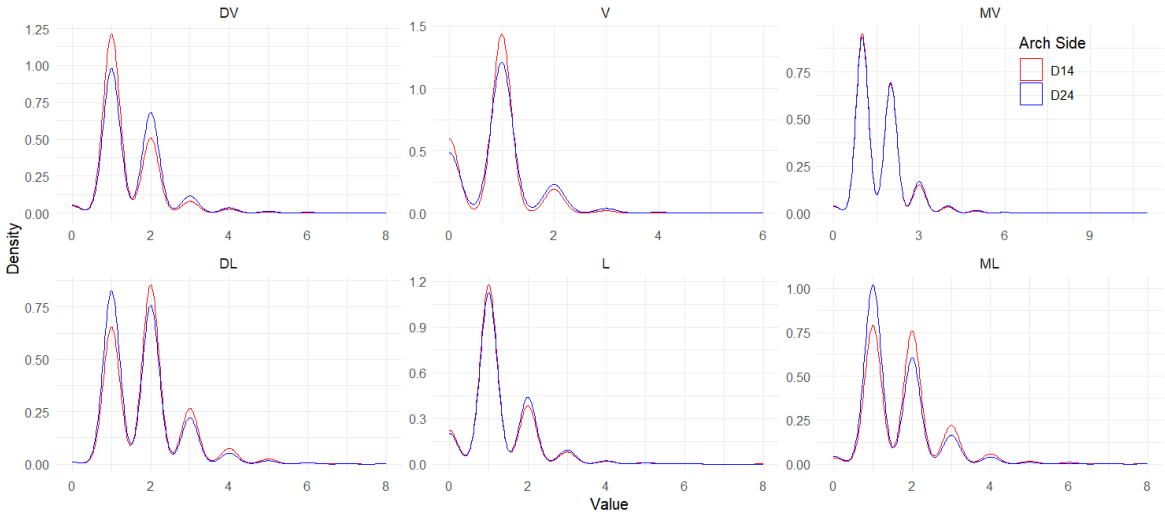
Figure I.7



Upper First Pre Molars Percentages of SM Values by Site

Kernel Density Plots Estimates

Figure I.8

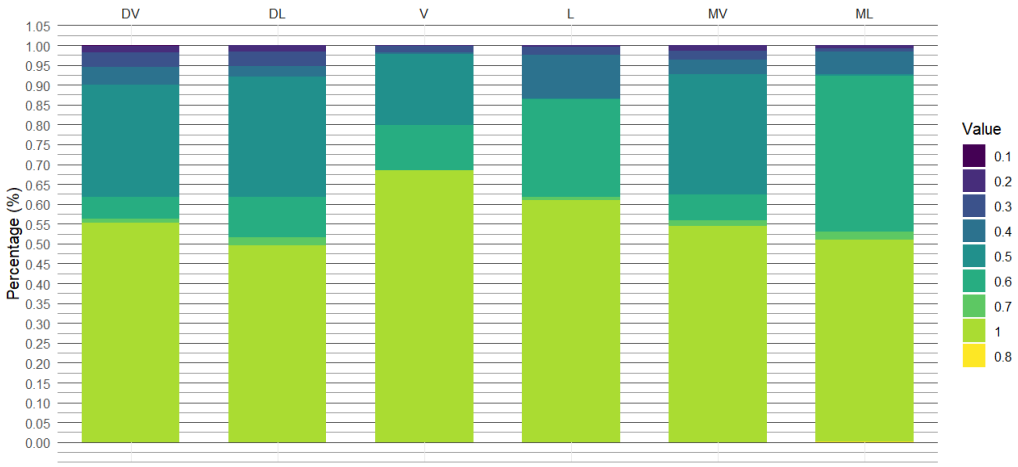


Kernel Density Plots Estimates of Upper First Pre Molars by Site

I.13 Upper Second Premolars (15, 25)

Symmetry Measure Bar Plots

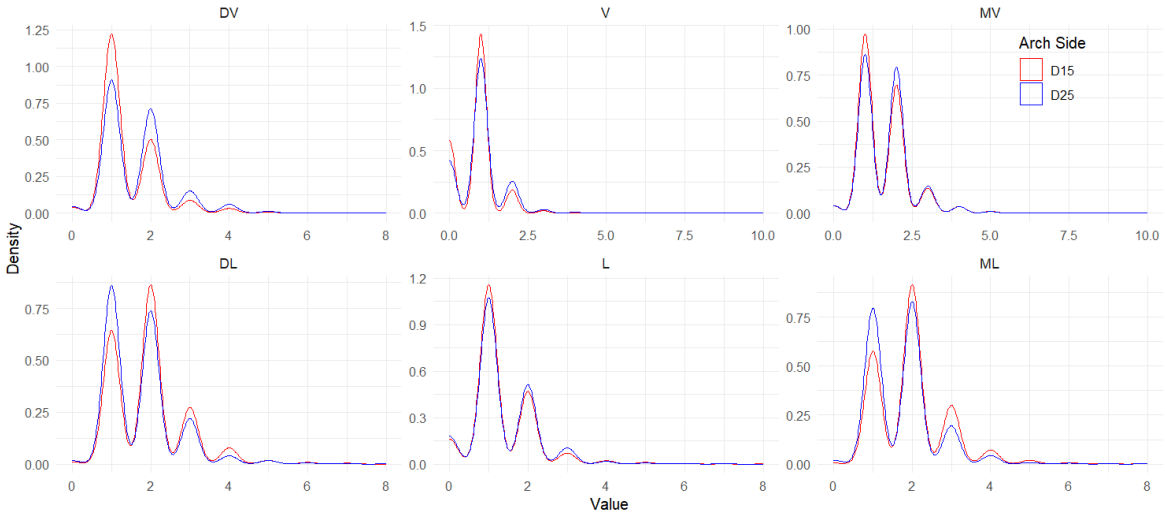
Figure I.9



Upper Second Premolars Percentages of SM Values by Site

Kernel Density Plots Estimates

Figure I.10

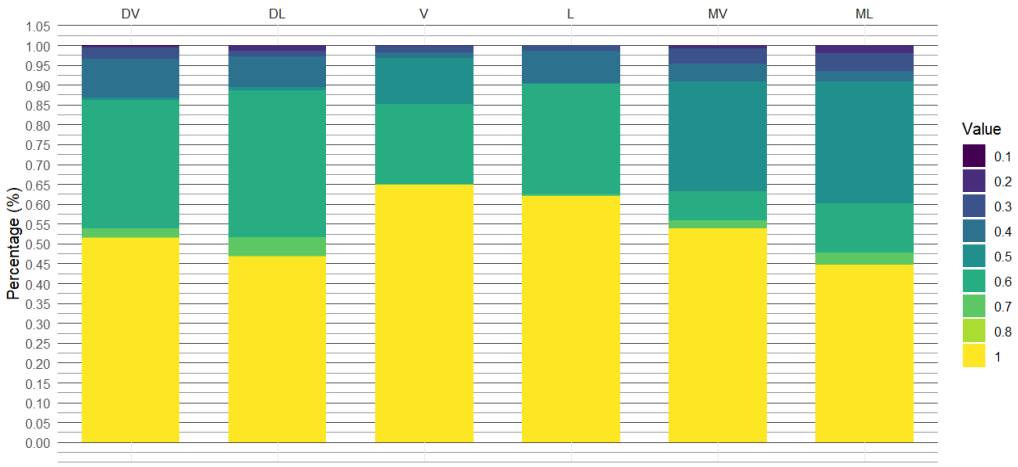


Kernel Density Plots Estimates of Upper Second Pre Molars by Site

I.14 Upper First Molars (16, 26)

Symmetry Measure Bar Plots

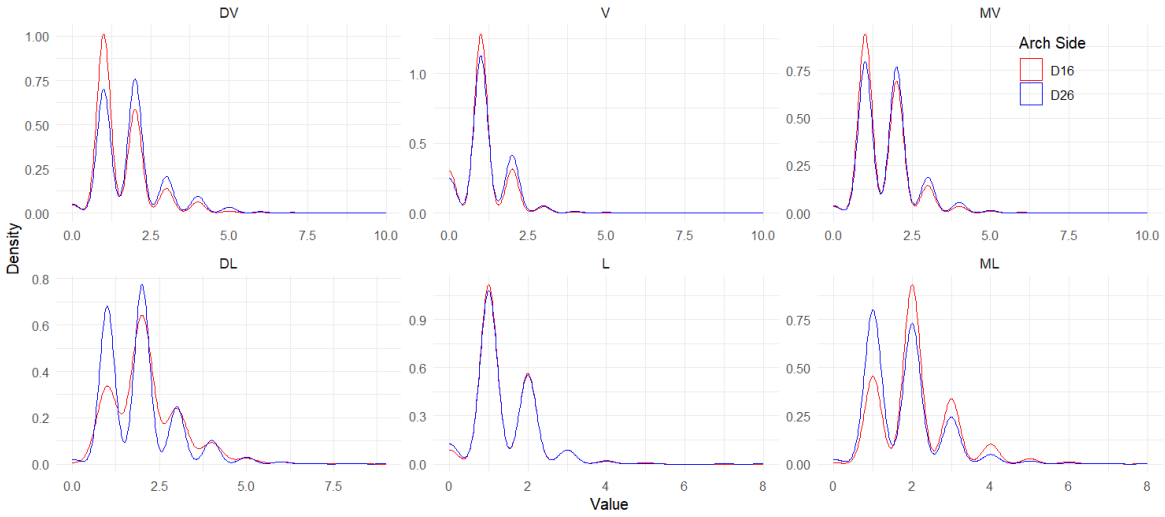
Figure I.11



Upper First Molars Percentages of SM Values by Site

Kernel Density Plots Estimates

Figure I.12

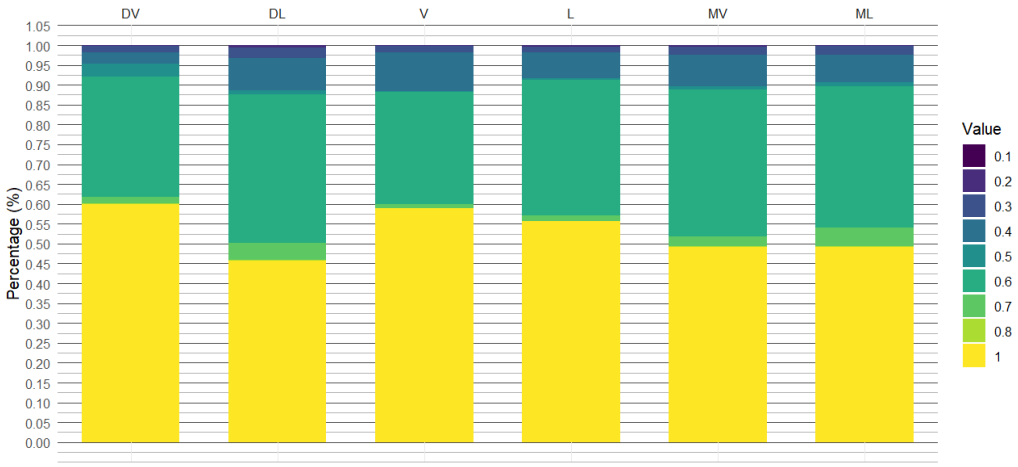


Kernel Density Plots Estimates of Upper First Molars by Site

I.15 Upper Second Molars (17, 27)

Symmetry Measure Bar Plots

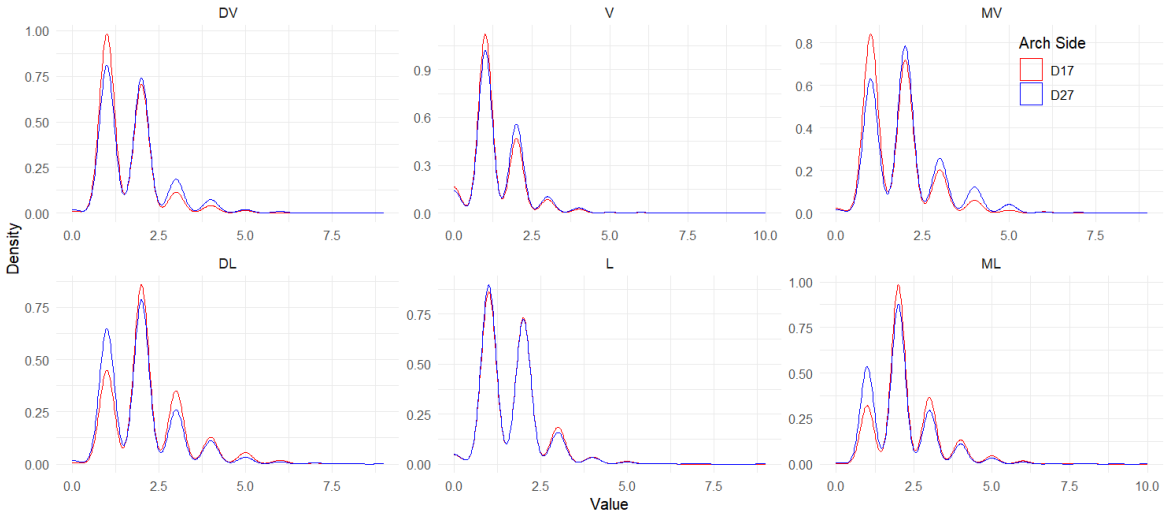
Figure I.13



Upper Second Molars Percentages of SM Values by Site

Kernel Density Plots Estimates

Figure I.14



Kernel Density Plots Estimates of Upper Second Molars by Site

Appendix II

Appendix: Hot Deck Imputation

II.1 Upper Right Central Incisor (11)

Original vs. H-D Imputed PPD by Site

Table II.1: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
11DV	0	6.66e-01	0.414	[-6.38e-03, 1.54e-02]	5.12e-02	4.67e-02
	1	4.86e+00	**0.0276**	**[3.11e-03, 5.25e-02]**	6.11e-01	5.83e-01
	2	5.80e+00	**0.0160**	**[-4.95e-02, -5.17e-03]**	2.49e-01	2.77e-01
	3	4.62e-01	0.497	[-1.64e-02, 7.94e-03]	6.02e-02	6.45e-02
	4	3.91e-01	0.532	[-9.36e-03, 4.81e-03]	1.91e-02	2.14e-02
	5	2.61e-01	0.609	[-2.72e-03, 4.62e-03]	5.77e-03	4.82e-03
	6	1.14e-01	0.736	[-2.02e-03, 2.85e-03]	2.52e-03	2.11e-03
	7	3.25e-02	0.857	[-1.18e-03, 1.42e-03]	7.21e-04	6.02e-04
11V	0	1.51e+00	0.219	[-8.53e-03, 3.71e-02]	2.95e-01	2.81e-01
	1	3.56e-03	0.952	[-2.43e-02, 2.58e-02]	5.60e-01	5.59e-01
	2	2.61e+00	0.106	[-3.09e-02, 2.92e-03]	1.23e-01	1.36e-01
	3	8.86e-02	0.766	[-7.75e-03, 5.70e-03]	1.77e-02	1.87e-02
	4	6.11e-02	0.805	[-3.31e-03, 2.57e-03]	3.24e-03	3.61e-03
	5	1.61e-02	0.899	[-8.61e-04, 9.79e-04]	3.60e-04	3.01e-04
	6	4.82e-02	0.826	[-1.42e-03, 1.77e-03]	1.08e-03	9.04e-04
11MV	0	2.08e-01	0.649	[-9.07e-03, 1.46e-02]	5.97e-02	5.69e-02
	1	1.50e-02	0.903	[-2.23e-02, 2.52e-02]	6.67e-01	6.66e-01
	2	1.37e-01	0.712	[-2.42e-02, 1.65e-02]	2.03e-01	2.07e-01
	3	9.61e-05	0.992	[-1.04e-02, 1.05e-02]	4.49e-02	4.49e-02
	4	2.35e-02	0.878	[-7.12e-03, 6.09e-03]	1.73e-02	1.78e-02
	5	8.48e-02	0.771	[-3.89e-03, 2.88e-03]	4.31e-03	4.82e-03
	6	1.10e-01	0.740	[-2.02e-03, 2.84e-03]	2.52e-03	2.11e-03
	7	3.14e-02	0.859	[-1.18e-03, 1.42e-03]	7.19e-04	6.02e-04
11DL	0	3.07e-01	0.580	[-7.59e-03, 1.36e-02]	4.70e-02	4.40e-02
	1	2.41e+00	0.121	[-5.22e-03, 4.51e-02]	5.63e-01	5.43e-01
	2	2.33e+00	0.127	[-4.10e-02, 5.08e-03]	2.85e-01	3.03e-01
	3	4.31e-01	0.512	[-1.81e-02, 9.00e-03]	7.56e-02	8.01e-02
	4	1.41e-02	0.906	[-7.06e-03, 7.97e-03]	2.27e-02	2.23e-02
	5	1.25e-01	0.724	[-4.25e-03, 2.94e-03]	4.77e-03	5.42e-03
	6	5.85e-02	0.809	[-1.41e-03, 1.81e-03]	1.10e-03	9.04e-04
	7	3.90e-02	0.843	[-1.18e-03, 1.45e-03]	7.34e-04	6.02e-04
11L	0	2.28e-01	0.633	[-1.43e-02, 2.35e-02]	1.71e-01	1.66e-01
	1	7.97e-02	0.778	[-2.12e-02, 2.83e-02]	5.93e-01	5.90e-01
	2	2.66e-01	0.606	[-2.45e-02, 1.43e-02]	1.78e-01	1.83e-01
	3	6.22e-02	0.803	[-1.08e-02, 8.39e-03]	3.73e-02	3.86e-02
	4	5.98e-01	0.439	[-8.37e-03, 3.61e-03]	1.33e-02	1.57e-02
	5	1.55e-01	0.694	[-2.32e-03, 3.48e-03]	3.59e-03	3.01e-03
	6	9.27e-02	0.761	[-1.90e-03, 2.59e-03]	2.15e-03	1.81e-03
	7	1.28e-01	0.721	[-2.39e-03, 1.64e-03]	1.44e-03	1.81e-03
11ML	0	1.53e-02	0.902	[-6.12e-03, 6.94e-03]	1.73e-02	1.69e-02
	1	8.91e-02	0.765	[-2.12e-02, 2.88e-02]	4.38e-01	4.35e-01
	2	4.60e-01	0.498	[-3.32e-02, 1.61e-02]	3.91e-01	3.99e-01
	3	1.28e-01	0.720	[-1.30e-02, 1.88e-02]	1.14e-01	1.11e-01
	4	5.75e-02	0.811	[-7.02e-03, 8.97e-03]	2.63e-02	2.53e-02
	5	8.62e-03	0.926	[-4.71e-03, 4.28e-03]	7.92e-03	8.13e-03
	6	7.35e-02	0.786	[-2.90e-03, 3.82e-03]	4.68e-03	4.22e-03
	7	1.59e-02	0.900	[-8.61e-04, 9.79e-04]	3.60e-04	3.01e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.2: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
11DV	3.38e-02	0.85	[-9.18e-03, 7.60e-03]	2.81e-02	2.89e-02
11V	1.78e-03	0.97	[-3.66e-03, 3.51e-03]	5.04e-03	5.12e-03
11MV	1.19e-02	0.91	[-8.36e-03, 7.48e-03]	2.52e-02	2.56e-02
11DL	8.52e-03	0.93	[-9.03e-03, 8.22e-03]	2.97e-02	3.01e-02
11L	2.42e-01	0.62	[-9.10e-03, 5.44e-03]	2.05e-02	2.23e-02
11ML	7.90e-02	0.78	[-8.39e-03, 1.12e-02]	4.00e-02	3.86e-02

PPD Statistics Before and After H-D Imputation by Site

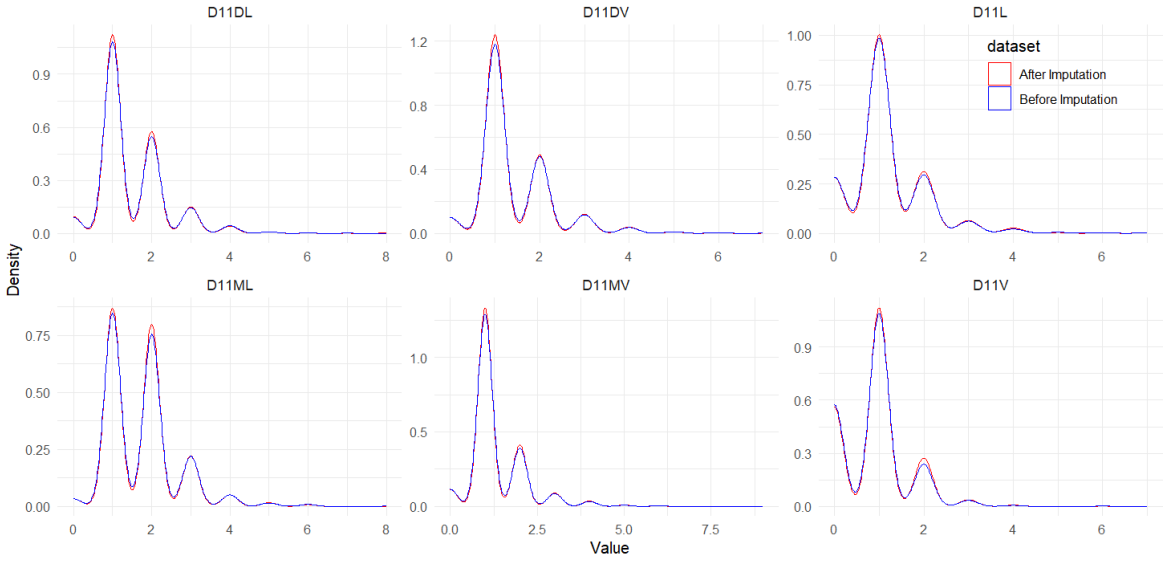
Table II.3: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 11		Test Results		K-S
		BI	AI	Stats	p	
DV	Median Variance	1 0.722	1 0.717	W = 4752302 F = 2.399	**1.29e-02** 0.122	D = 0.032 p = 0.848
V	Median Variance	1 0.549	1 0.555	W = 4712600 F = 0.008	8.78e-02 0.928	D = 0.015 p = 0.882
MV	Median Variance	1 0.675	1 0.666	W = 4644844 F = 0.005	0.640 0.945	D = 0.004 p = 1
DL	Median Variance	1 0.748	1 0.771	W = 4629765 F = 1.693	8.44e-2 0.193	D = 0.023 p = 0.410
L	Median Variance	1 0.757	1 0.773	W = 4673501 F = 0.193	0.422 0.661	D = 0.008 p = 1
ML	Median Variance	2 0.841	2 0.816	W = 4617429 F = 0.481	0.925 0.488	D = 0.004 p = 1

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.1



Kernel Density Plot of Imputed and Original 11 by Site

II.2 Upper Right Lateral Incisor (12)

Original vs. H-D Imputed PPD by Site

Table II.4: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
12DV	0	1.90e+00	0.169	[-2.93e-03, 1.64e-02]	4.14e-02	3.46e-02
	1	2.84e+01	**9.96e-08**	**[4.28e-02, 9.23e-02]**	6.20e-01	5.52e-01
	2	4.35e+01	**4.18e-11**	**[-1.00e-01, -5.46e-02]**	2.53e-01	3.31e-01
	3	1.61e-01	0.689	[-8.35e-03, 1.26e-02]	4.64e-02	4.43e-02
	4	9.69e-01	0.325	[-3.82e-03, 1.14e-02]	2.52e-02	2.14e-02
	5	3.68e-01	0.544	[-3.05e-03, 5.74e-03]	8.28e-03	6.93e-03
	6	3.59e-02	0.850	[-2.68e-03, 3.25e-03]	3.60e-03	3.31e-03
	7	8.11e+00	**0.0044**	**[-7.57e-03, -1.60e-03]**	1.44e-03	6.02e-03
12V	0	7.03e-01	0.402	[-1.36e-02, 3.39e-02]	3.40e-01	3.30e-01
	1	3.53e-02	0.851	[-2.27e-02, 2.76e-02]	5.27e-01	5.25e-01
	2	1.20e+00	0.274	[-2.50e-02, 7.06e-03]	1.11e-01	1.20e-01
	3	3.51e-01	0.554	[-8.77e-03, 4.68e-03]	1.72e-02	1.93e-02
	4	1.52e-01	0.697	[-3.76e-03, 2.50e-03]	3.59e-03	4.22e-03
	5	7.32e-01	0.392	[-3.16e-03, 1.21e-03]	1.44e-03	2.41e-03
	6	1.55e-02	0.901	[-8.61e-04, 9.76e-04]	3.59e-04	3.01e-04
12MV	0	4.89e-01	0.485	[-6.36e-03, 1.34e-02]	4.15e-02	3.80e-02
	1	1.34e-01	0.714	[-2.02e-02, 2.95e-02]	5.89e-01	5.84e-01
	2	3.31e-01	0.565	[-2.92e-02, 1.60e-02]	2.75e-01	2.81e-01
	3	3.30e-02	0.856	[-1.13e-02, 1.36e-02]	6.56e-02	6.45e-02
	4	6.44e-02	0.800	[-8.37e-03, 6.45e-03]	2.16e-02	2.26e-02
	5	6.42e-01	0.423	[-5.40e-03, 2.24e-03]	5.05e-03	6.63e-03
	6	4.21e-03	0.947	[-2.00e-03, 1.87e-03]	1.43e-03	1.53e-03
	7	4.21e-03	0.948	[-2.00e-03, 1.87e-03]	1.44e-03	1.51e-03
12DL	0	1.15e+00	0.283	[-3.86e-03, 1.31e-02]	3.11e-02	2.65e-02
	1	6.13e+00	**0.0133**	**[6.66e-03, 5.71e-02]**	5.05e-01	4.73e-01
	2	4.35e+00	**0.0371**	**[-4.96e-02, -1.58e-03]**	3.35e-01	3.60e-01
	3	4.00e+00	**0.0456**	**[-3.03e-02, -4.00e-04]**	9.01e-02	1.05e-01
	4	6.54e-01	0.419	[-4.89e-03, 1.17e-02]	2.93e-02	2.59e-02
	5	1.40e-01	0.708	[-3.11e-03, 4.57e-03]	6.15e-03	5.42e-03
	6	2.75e-03	0.958	[-2.27e-03, 2.40e-03]	2.17e-03	2.11e-03
	7	5.06e-02	0.822	[-1.42e-03, 1.78e-03]	1.08e-03	9.04e-04
12L	0	9.10e-02	0.763	[-1.58e-02, 2.16e-02]	1.66e-01	1.64e-01
	1	2.42e-03	0.961	[-2.54e-02, 2.42e-02]	5.87e-01	5.88e-01
	2	1.46e-03	0.970	[-1.95e-02, 2.02e-02]	1.93e-01	1.92e-01
	3	1.28e-01	0.720	[-1.13e-02, 7.79e-03]	3.65e-02	3.83e-02
	4	2.18e-02	0.883	[-6.06e-03, 5.21e-03]	1.25e-02	1.30e-02
	5	9.92e-02	0.753	[-3.23e-03, 2.33e-03]	2.86e-03	3.31e-03
	6	5.71e-03	0.940	[-2.00e-03, 1.85e-03]	1.43e-03	1.51e-03
12ML	0	5.40e-02	0.816	[-9.44e-03, 1.20e-02]	4.77e-02	4.64e-02
	1	7.10e-02	0.790	[-2.18e-02, 2.86e-02]	5.45e-01	5.42e-01
	2	4.97e-03	0.944	[-2.39e-02, 2.22e-02]	2.95e-01	2.96e-01
	3	1.60e-02	0.899	[-1.45e-02, 1.28e-02]	7.89e-02	7.98e-02
	4	1.26e-01	0.723	[-8.56e-03, 5.93e-03]	2.04e-02	2.17e-02
	5	2.05e-01	0.650	[-6.08e-03, 3.79e-03]	9.09e-03	1.02e-02
	6	2.27e-01	0.634	[-3.57e-03, 2.16e-03]	2.91e-03	3.61e-03
	7	1.79e-02	0.894	[-8.63e-04, 9.88e-04]	3.64e-04	3.01e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.5: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
12DV	3.82e-02	0.845	[-8.76e-03, 1.07e-02]	3.92e-02	3.82e-02
12V	5.82e-01	0.446	[-5.46e-03, 2.38e-03]	5.39e-03	6.93e-03
12MV	3.58e-01	0.550	[-1.14e-02, 6.04e-03]	2.96e-02	3.22e-02
12DL	8.40e-01	0.359	[-5.10e-03, 1.40e-02]	3.91e-02	3.46e-02
12L	6.97e-02	0.792	[-7.51e-03, 5.72e-03]	1.72e-02	1.81e-02
12ML	3.94e-01	0.530	[-1.22e-02, 6.29e-03]	3.35e-02	3.65e-02

PPD Statistics Before and After H-D Imputation by Site

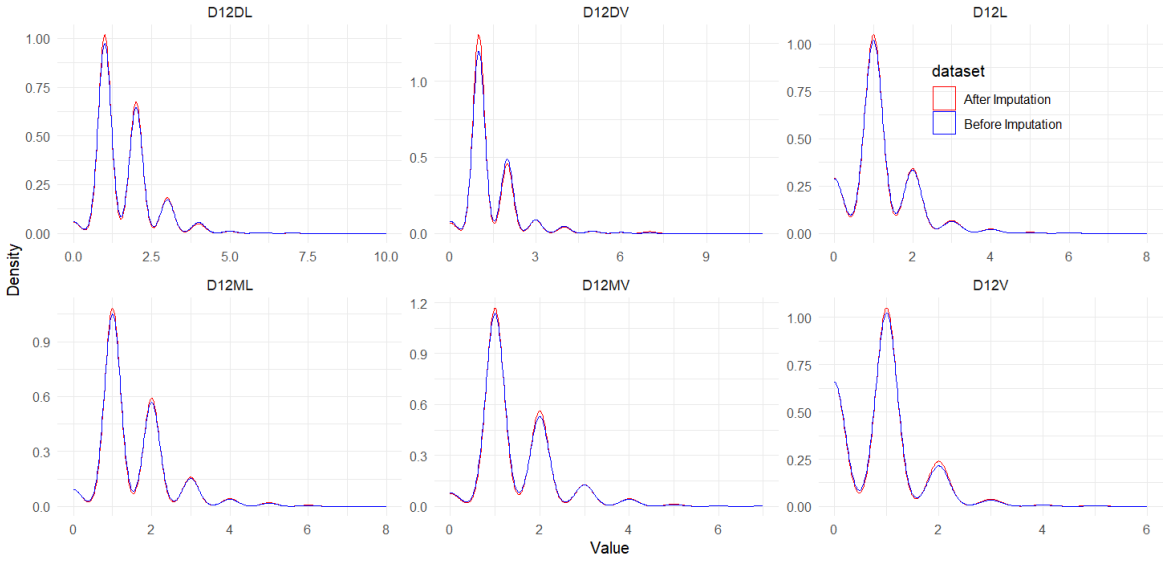
Table II.6: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 12		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4928995	**1.65e-7**	D = 0.074
	Variance	0.836	0.905	F = 10.617	**1.13e-3**	p = **1.124e-7**
V	Median	1	1	W = 4703153	0.196	D = 0.013
	Variance	0.553	0.589	F = 0.313	0.576	p = 0.971
MV	Median	1	1	W = 4651072	0.443	D = 0.008
	Variance	0.726	0.745	F = 0.266	0.606	p = 1
DL	Median	1	1	W = 4758760	**6.40e-3**	D = 0.036
	Variance	0.831	0.799	F = 3.486	0.062	**p = 0.036**
L	Median	1	1	W = 4659273	0.728	D = 0.003
	Variance	0.692	0.703	F = 0.034	0.854	p = 1
ML	Median	1	1	W = 4592928	0.629	D = 0.005
	Variance	0.834	0.861	F = 0.299	0.585	p = 1

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.2



Kernel Density Plot of Imputed and Original 12 by Site

II.3 Upper Right Canine (13)

Original vs. H-D Imputed PPD by Site

Table II.7: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
13DV	0	2.85e-02	0.866	[-6.88e-03, 5.79e-03]	1.63e-02	1.69e-02
	1	2.60e-01	0.611	[-1.77e-02, 3.01e-02]	6.38e-01	6.31e-01
	2	1.77e-01	0.674	[-2.69e-02, 1.74e-02]	2.72e-01	2.77e-01
	3	4.17e-02	0.838	[-1.18e-02, 9.58e-03]	4.83e-02	4.94e-02
	4	1.67e-02	0.897	[-6.46e-03, 5.66e-03]	1.50e-02	1.54e-02
	5	1.25e-04	0.991	[-3.56e-03, 3.52e-03]	5.10e-03	5.12e-03
	6	6.64e-02	0.797	[-2.32e-03, 3.02e-03]	3.06e-03	2.71e-03
	7	2.21e-02	0.882	[-1.43e-03, 1.66e-03]	1.02e-03	9.04e-04
13V	0	7.98e-02	0.778	[-2.04e-02, 2.72e-02]	3.58e-01	3.55e-01
	1	1.01e-01	0.750	[-2.88e-02, 2.07e-02]	5.35e-01	5.39e-01
	2	1.12e-01	0.738	[-1.17e-02, 1.65e-02]	8.92e-02	8.68e-02
	3	2.31e-01	0.631	[-7.34e-03, 4.44e-03]	1.36e-02	1.51e-02
	4	2.29e-01	0.633	[-3.20e-03, 1.94e-03]	2.38e-03	3.01e-03
	5	3.73e-02	0.847	[-1.80e-03, 2.19e-03]	1.70e-03	1.51e-03
	6	1.49e-02	0.903	[-1.18e-03, 1.34e-03]	6.80e-04	6.02e-04
13MV	0	2.10e-02	0.885	[-1.03e-02, 8.88e-03]	3.85e-02	3.92e-02
	1	1.40e-01	0.708	[-1.99e-02, 2.93e-02]	5.74e-01	5.69e-01
	2	3.05e-02	0.861	[-2.49e-02, 2.08e-02]	3.03e-01	3.05e-01
	3	2.00e-01	0.655	[-1.41e-02, 8.88e-03]	5.58e-02	5.84e-02
	4	8.86e-04	0.976	[-7.08e-03, 6.87e-03]	2.01e-02	2.02e-02
	5	2.56e-02	0.873	[-3.21e-03, 3.78e-03]	5.10e-03	4.82e-03
	6	4.47e-02	0.833	[-1.95e-03, 2.41e-03]	2.04e-03	1.81e-03
	7	2.98e-02	0.863	[-1.62e-03, 1.94e-03]	1.36e-03	1.21e-03
13DL	0	9.51e-03	0.922	[-5.87e-03, 6.48e-03]	1.57e-02	1.54e-02
	1	7.05e-02	0.791	[-2.80e-02, 2.13e-02]	4.22e-01	4.25e-01
	2	4.97e-03	0.944	[-2.34e-02, 2.52e-02]	3.84e-01	3.83e-01
	3	2.82e-04	0.987	[-1.66e-02, 1.63e-02]	1.24e-01	1.24e-01
	4	6.80e-02	0.794	[-7.99e-03, 1.04e-02]	3.59e-02	3.46e-02
	5	2.56e-01	0.613	[-4.02e-03, 6.80e-03]	1.25e-02	1.11e-02
	6	2.66e-01	0.606	[-3.73e-03, 2.17e-03]	3.13e-03	3.92e-03
	7	8.44e-02	0.771	[-2.17e-03, 2.93e-03]	2.79e-03	2.41e-03
13L	0	2.25e-02	0.881	[-1.77e-02, 1.52e-02]	1.26e-01	1.27e-01
	1	1.96e-01	0.658	[-1.89e-02, 2.99e-02]	5.90e-01	5.84e-01
	2	6.66e-02	0.796	[-2.35e-02, 1.80e-02]	2.26e-01	2.28e-01
	3	6.12e-03	0.938	[-9.12e-03, 9.88e-03]	3.83e-02	3.80e-02
	4	1.63e-01	0.687	[-6.53e-03, 4.30e-03]	1.15e-02	1.27e-02
	5	3.32e-02	0.856	[-3.92e-03, 3.25e-03]	5.09e-03	5.42e-03
	6	9.82e-02	0.754	[-2.70e-03, 1.96e-03]	2.04e-03	2.41e-03
	7	2.12e-02	0.884	[-1.43e-03, 1.65e-03]	1.02e-03	9.04e-04
13ML	0	1.09e-02	0.917	[-8.75e-03, 7.86e-03]	2.85e-02	2.89e-02
	1	2.21e-02	0.882	[-2.68e-02, 2.30e-02]	4.86e-01	4.88e-01
	2	5.23e-04	0.982	[-2.41e-02, 2.35e-02]	3.56e-01	3.56e-01
	3	1.27e-01	0.722	[-1.16e-02, 1.67e-02]	9.02e-02	8.77e-02
	4	9.95e-04	0.975	[-7.74e-03, 7.99e-03]	2.57e-02	2.56e-02
	5	2.64e-02	0.871	[-4.00e-03, 4.72e-03]	7.89e-03	7.53e-03
	6	1.07e-02	0.918	[-2.86e-03, 3.18e-03]	3.77e-03	3.61e-03
	7	2.75e-01	0.600	[-2.24e-03, 1.28e-03]	1.03e-03	1.51e-03

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.8: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
13DV	4.12-02	0.813	[-7.68-03, 6.12-03]	2.89-02	2.91-02
13V	2.23-03	0.946	[-3.23-03, 3.11-03]	4.93-03	5.10-03
13MV	1.04-02	0.912	[-7.45-03, 7.23-03]	2.45-02	2.53-02
13DL	9.23-03	0.914	[-8.32-03, 7.94-03]	3.01-02	3.01-02
13L	2.57-01	0.621	[-8.79-03, 5.45-03]	2.11-02	2.23-02
13ML	6.87-02	0.785	[-7.98-03, 1.05-02]	3.97-02	3.85-02

PPD Statistics Before and After H-D Imputation by Site

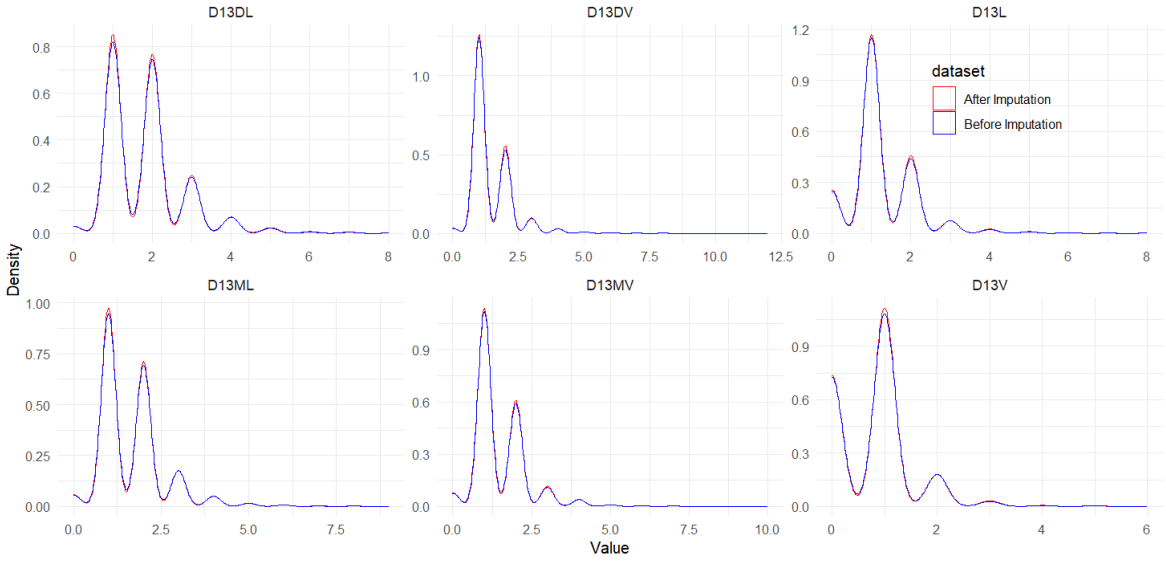
Table II.9: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 13		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4906766	0.684	D = 0.006
	Variance	0.700	0.687	F = 0.062	0.804	p = 1
V	Median	1	1	W = 4892788	0.826	D = 0.003
	Variance	0.517	0.520	F = 0.0232	0.879	p = 1
MV	Median	1	1	W = 4896667	0.777	D = 0.004
	Variance	0.760	0.748	F = 0.040	0.842	p = 1
DL	Median	2	2	W = 4749277	0.780	D = 0.003
	Variance	0.974	0.955	F = 0.030	0.864	p = 1
L	Median	1	1	W = 4908751	0.812	D = 0.004
	Variance	0.763	0.791	F = 0.291	0.590	p = 1
ML	Median	1	1	W = 4822723	0.803	D = 0.003
	Variance	0.883	0.891	F = 0.021	0.886	p = 1

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.3



Kernel Density Plot of Imputed and Original 13 by Site

II.4 Upper Right First Pre Molar (14)

Original vs. H-D Imputed PPD by Site

Table II.10: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
D14DV	0	2.68e-01	0.605	[-1.05e-02, 6.08e-03]	2.52e-02	2.74e-02
	1	1.68e-01	0.682	[-1.97e-02, 3.02e-02]	6.41e-01	6.36e-01
	2	4.71e-01	0.493	[-3.12e-02, 1.50e-02]	2.69e-01	2.77e-01
	3	3.90e-01	0.533	[-7.03e-03, 1.36e-02]	4.24e-02	3.92e-02
	4	6.03e-02	0.806	[-5.28e-03, 6.78e-03]	1.40e-02	1.33e-02
	5	1.88e-01	0.665	[-2.47e-03, 3.84e-03]	4.00e-03	3.31e-03
	6	2.44e-01	0.621	[-1.81e-03, 3.00e-03]	2.40e-03	1.81e-03
D14V	0	4.77e-02	0.827	[-2.55e-02, 2.04e-02]	2.65e-01	2.67e-01
	1	1.31e-03	0.971	[-2.45e-02, 2.54e-02]	6.38e-01	6.38e-01
	2	5.04e-03	0.943	[-1.40e-02, 1.51e-02]	8.61e-02	8.55e-02
	3	1.66e-01	0.684	[-3.76e-03, 5.71e-03]	8.81e-03	7.83e-03
	4	1.63e-01	0.686	[-1.57e-03, 2.36e-03]	1.60e-03	1.21e-03
	5	4.08e-02	0.840	[-8.83e-04, 1.08e-03]	4.00e-04	3.01e-04
D14MV	0	3.69e-02	0.848	[-8.11e-03, 6.66e-03]	2.04e-02	2.11e-02
	1	3.88e-03	0.950	[-2.68e-02, 2.51e-02]	5.03e-01	5.04e-01
	2	2.01e-01	0.654	[-1.93e-02, 3.07e-02]	3.68e-01	3.62e-01
	3	5.48e-01	0.459	[-1.96e-02, 8.84e-03]	7.98e-02	8.52e-02
	4	2.10e-02	0.885	[-7.36e-03, 6.35e-03]	1.76e-02	1.81e-02
	5	3.68e-01	0.544	[-2.66e-03, 5.00e-03]	5.99e-03	4.82e-03
	6	1.24e-04	0.991	[-2.55e-03, 2.52e-03]	2.40e-03	2.41e-03
	7	1.20e-01	0.729	[-1.40e-03, 1.99e-03]	1.20e-03	9.04e-04
D14DL	0	7.44e-04	0.978	[-3.63e-03, 3.53e-03]	4.77e-03	4.82e-03
	1	1.72e+00	0.190	[-8.16e-03, 4.09e-02]	3.46e-01	3.30e-01
	2	1.51e-01	0.698	[-3.09e-02, 2.07e-02]	4.52e-01	4.57e-01
	3	1.51e+00	0.219	[-2.98e-02, 6.75e-03]	1.40e-01	1.52e-01
	4	3.01e-02	0.862	[-9.18e-03, 1.10e-02]	3.98e-02	3.89e-02
	5	3.18e-02	0.859	[-6.39e-03, 5.32e-03]	1.27e-02	1.33e-02
	6	4.37e-02	0.834	[-3.49e-03, 2.82e-03]	3.58e-03	3.92e-03
	7	7.74e-02	0.781	[-1.19e-03, 1.57e-03]	7.95e-04	6.02e-04
D14L	0	6.52e-04	0.980	[-1.69e-02, 1.65e-02]	1.18e-01	1.18e-01
	1	1.87e-02	0.891	[-2.34e-02, 2.69e-02]	6.21e-01	6.19e-01
	2	5.27e-03	0.942	[-2.16e-02, 2.01e-02]	2.03e-01	2.03e-01
	3	1.96e-02	0.889	[-9.58e-03, 1.11e-02]	4.17e-02	4.10e-02
	4	8.37e-02	0.772	[-6.44e-03, 4.78e-03]	1.15e-02	1.24e-02
	5	4.45e-02	0.833	[-3.50e-03, 2.81e-03]	3.57e-03	3.92e-03
	6	1.16e-01	0.734	[-2.83e-03, 1.99e-03]	1.99e-03	2.41e-03
D14ML	0	1.52e+00	0.218	[-1.21e-02, 2.68e-03]	1.88e-02	2.35e-02
	1	7.84e-01	0.376	[-3.72e-02, 1.40e-02]	4.17e-01	4.29e-01
	2	1.09e+00	0.296	[-1.18e-02, 3.89e-02]	4.01e-01	3.87e-01
	3	8.45e-02	0.771	[-1.41e-02, 1.90e-02]	1.17e-01	1.14e-01
	4	3.78e-01	0.539	[-6.05e-03, 1.15e-02]	3.08e-02	2.80e-02
	5	2.08e-02	0.885	[-4.45e-03, 5.15e-03]	8.79e-03	8.43e-03
	6	4.92e-01	0.483	[-5.38e-03, 2.51e-03]	5.19e-03	6.63e-03
	7	3.43e-01	0.558	[-2.59e-03, 1.37e-03]	1.20e-03	1.81e-03

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.11: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
14DV	2.32e-01	0.630	[-5.65e-03, 9.31e-03]	2.20e-02	2.02e-02
14V	2.45e-01	0.621	[-1.81e-03, 2.99e-03]	2.40e-03	1.81e-03
14MV	8.10e-02	0.776	[-7.29e-03, 9.76e-03]	2.83e-02	2.71e-02
14DL	2.49e-03	0.960	[-1.17e-02, 1.23e-02]	5.72e-02	5.69e-02
14L	1.80e-01	0.672	[-8.41e-03, 5.41e-03]	1.75e-02	1.90e-02
14ML	3.71e-03	0.951	[-1.06e-02, 1.13e-02]	4.67e-02	4.64e-02

PPD Statistics Before and After H-D Imputation by Site

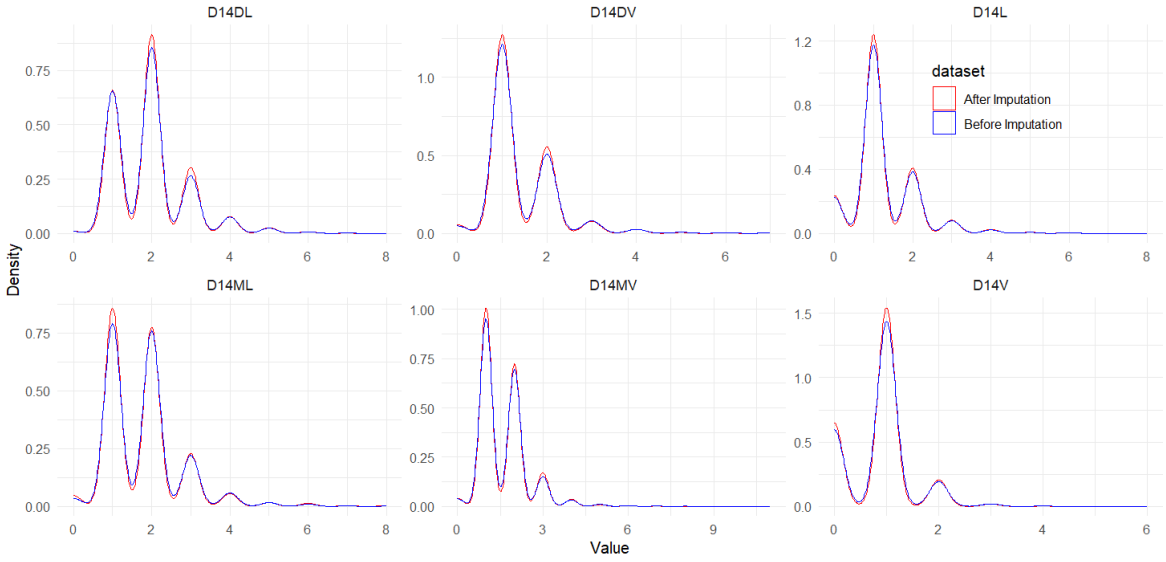
Table II.12: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 14		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4146546	0.973	D = 0.005
	Variance	0.605	0.583	F = 0.023	0.880	p = 1
V	Median	1	1	W = 4130166	0.759	D = 0.003
	Variance	0.394	0.382	F = 0.021	0.884	p = 1
MV	Median	1	1	W = 4158165	0.998	D = 0.004
	Variance	0.769	0.745	F = 0.004	0.950	p = 1
DL	Median	2	2	W = 4258421	0.167	D = 0.016
	Variance	0.865	0.862	F = 0.080	0.777	p = 0.841
L	Median	1	1	W = 4185664	0.917	D = 0.002
	Variance	0.664	0.679	F = 0.066	0.798	p = 1
ML	Median	2	2	W = 4087012	0.238	D = 0.016
	Variance	0.905	0.975	F = 1.795	0.180	p = 0.842

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.4



Kernel Density Plot of Imputed and Original 14 by Site

II.5 Upper Right Second Premolar (15)

Original vs. H-D Imputed PPD by Site

Table II.13: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	2.60e-01	0.611	[-9.76e-03, 5.72e-03]	2.21e-02	2.41e-02
	1	9.95e-01	0.319	[-1.22e-02, 3.75e-02]	6.44e-01	6.31e-01
	2	1.35e+00	0.246	[-3.66e-02, 9.32e-03]	2.64e-01	2.78e-01
	3	4.11e-02	0.839	[-9.75e-03, 1.20e-02]	4.69e-02	4.58e-02
	4	4.46e-01	0.504	[-4.30e-03, 8.68e-03]	1.70e-02	1.48e-02
	5	7.33e-02	0.787	[-3.96e-03, 2.99e-03]	4.34e-03	4.82e-03
	6	7.28e-02	0.787	[-1.19e-03, 1.56e-03]	7.88e-04	6.02e-04
	8	3.64e-02	0.849	[-8.79e-04, 1.07e-03]	3.94e-04	3.01e-04
V	0	8.79e-01	0.348	[-1.18e-02, 3.33e-02]	2.60e-01	2.49e-01
	1	1.90e-01	0.663	[-3.02e-02, 1.92e-02]	6.44e-01	6.49e-01
	2	2.86e-01	0.593	[-1.84e-02, 1.05e-02]	8.37e-02	8.77e-02
	3	1.04e-02	0.919	[-5.19e-03, 4.68e-03]	9.08e-03	9.34e-03
	4	6.06e-01	0.436	[-3.58e-03, 1.50e-03]	1.97e-03	3.01e-03
	5	3.68e-02	0.848	[-8.80e-04, 1.07e-03]	3.95e-04	3.01e-04
	6	2.19e-02	0.882	[-1.61e-03, 1.38e-03]	7.90e-04	9.04e-04
MV	0	4.64e-02	0.829	[-6.59e-03, 8.21e-03]	2.13e-02	2.05e-02
	1	9.07e-02	0.763	[-2.98e-02, 2.19e-02]	5.14e-01	5.18e-01
	2	6.35e-01	0.426	[-1.48e-02, 3.49e-02]	3.67e-01	3.57e-01
	3	9.12e-01	0.340	[-2.02e-02, 6.91e-03]	7.14e-02	7.80e-02
	4	5.46e-02	0.815	[-8.04e-03, 6.32e-03]	1.93e-02	2.02e-02
	5	1.12e-01	0.738	[-2.99e-03, 4.20e-03]	5.13e-03	4.52e-03
	6	5.75e-04	0.981	[-1.81e-03, 1.76e-03]	1.18e-03	1.21e-03
	8	1.22e-01	0.727	[-1.35e-03, 9.29e-04]	3.94e-04	6.02e-04
DL	0	1.19e-02	0.913	[-4.07e-03, 3.64e-03]	5.51e-03	5.72e-03
	1	1.47e-01	0.702	[-1.97e-02, 2.92e-02]	3.39e-01	3.34e-01
	2	2.23e-02	0.881	[-2.38e-02, 2.77e-02]	4.54e-01	4.52e-01
	3	3.75e-01	0.540	[-2.40e-02, 1.26e-02]	1.44e-01	1.50e-01
	4	5.69e-02	0.811	[-1.16e-02, 9.04e-03]	4.09e-02	4.22e-02
	5	2.59e-02	0.872	[-4.54e-03, 5.35e-03]	9.44e-03	9.04e-03
	6	5.98e-02	0.807	[-2.91e-03, 3.73e-03]	4.33e-03	3.92e-03
	7	6.93e-02	0.792	[-2.94e-03, 2.24e-03]	2.36e-03	2.71e-03
L	0	9.85e-02	0.754	[-1.20e-02, 1.66e-02]	8.45e-02	8.22e-02
	1	3.73e-01	0.542	[-1.74e-02, 3.31e-02]	6.10e-01	6.02e-01
	2	1.20e-01	0.729	[-2.63e-02, 1.84e-02]	2.47e-01	2.51e-01
	3	2.64e-01	0.608	[-1.26e-02, 7.36e-03]	3.77e-02	4.04e-02
	4	8.40e-03	0.927	[-5.20e-03, 5.71e-03]	1.14e-02	1.11e-02
	5	1.87e+00	0.172	[-7.52e-03, 1.24e-03]	5.90e-03	9.04e-03
	6	3.32e-01	0.564	[-3.75e-03, 2.02e-03]	2.75e-03	3.61e-03
	7	3.57e-02	0.850	[-8.79e-04, 1.06e-03]	3.93e-04	3.01e-04
ML	0	1.49e-01	0.699	[-3.38e-03, 2.26e-03]	2.75e-03	3.31e-03
	1	7.12e-03	0.933	[-2.27e-02, 2.48e-02]	3.04e-01	3.03e-01
	2	2.54e-01	0.614	[-1.92e-02, 3.24e-02]	4.81e-01	4.75e-01
	3	6.29e-01	0.428	[-2.67e-02, 1.13e-02]	1.58e-01	1.66e-01
	4	6.58e-02	0.798	[-8.49e-03, 1.10e-02]	3.77e-02	3.65e-02
	5	9.20e-02	0.762	[-6.22e-03, 4.54e-03]	1.06e-02	1.15e-02
	6	6.83e-05	0.993	[-3.21e-03, 3.24e-03]	3.93e-03	3.92e-03
	7	2.35e-02	0.878	[-1.61e-03, 1.38e-03]	7.86e-04	9.04e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.14: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
15DV	2.32e-01	0.630	[-5.79e-03, 9.53e-03]	2.33e-02	2.14e-02
15V	4.30e-01	0.512	[-4.16e-03, 2.05e-03]	3.16e-03	4.22e-03
15MV	4.75e-03	0.945	[-8.66e-03, 8.08e-03]	2.68e-02	2.71e-02
15DL	1.66e-02	0.898	[-1.28e-02, 1.12e-02]	5.70e-02	5.78e-02
15L	8.20e-01	0.365	[-1.12e-02, 4.07e-03]	2.08e-02	2.44e-02
15ML	1.05e-02	0.918	[-1.11e-02, 1.23e-02]	5.42e-02	5.36e-02

PPD Statistics Before and After H-D Imputation by Site

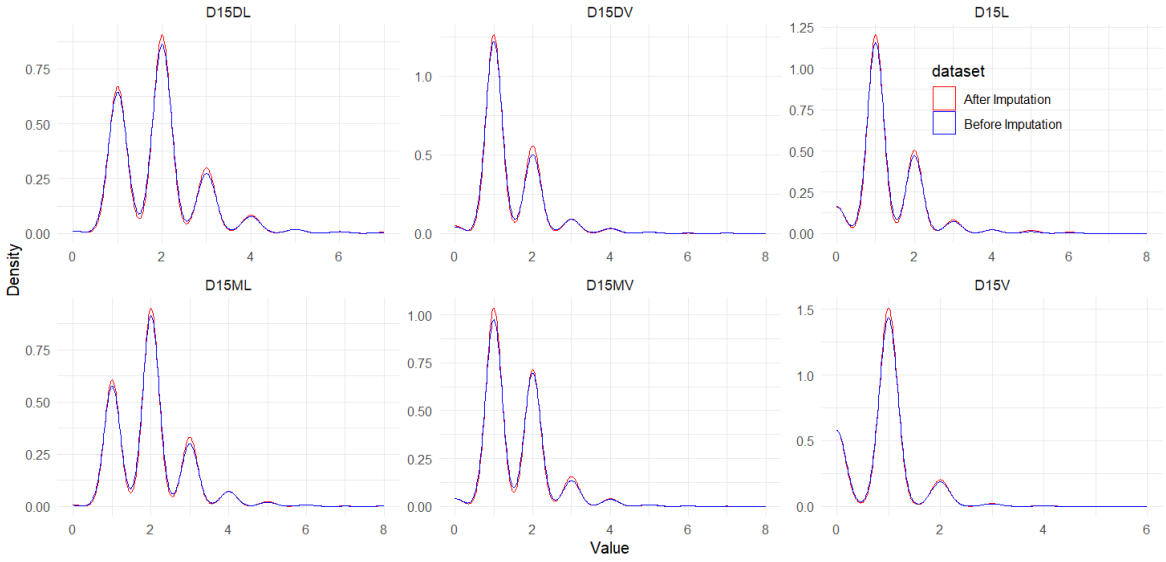
Table II.15: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 15		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4241975	0.575	D = 0.011
	Variance	0.591	0.582	F = 0.164	0.686	p = 0.997
V	Median	1	1	W = 4262404	0.286	D = 0.011
	Variance	0.403	0.413	F = 0.044	0.833	p = 0.996
MV	Median	1	1	W = 4212726	0.959	D = 0.007
	Variance	0.671	0.683	F = 0.020	0.887	p = 1
DL	Median	2	2	W = 4251596	0.594	D = 0.007
	Variance	0.878	0.885	F = 0.023	0.879	p = 1
L	Median	1	1	W = 4274942	0.355	D = 0.010
	Variance	0.683	0.739	F = 1.190	0.276	p = 0.998
ML	Median	2	2	W = 4244277	0.742	D = 0.007
	Variance	0.844	0.848	F = 0.133	0.716	p = 1

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.5



Kernel Density Plot of Imputed and Original 15 by Site

II.6 Upper Right First Molar (16)

Original vs. H-D Imputed PPD by Site

Table II.16: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	7.59e-01	0.384	[-1.24e-02, 4.72e-03]	2.57e-02	2.95e-02
	1	4.30e-01	0.512	[-1.75e-02, 3.51e-02]	5.41e-01	5.32e-01
	2	3.48e-01	0.555	[-1.71e-02, 3.17e-02]	3.14e-01	3.07e-01
	3	4.16e-01	0.519	[-1.87e-02, 9.39e-03]	7.49e-02	7.95e-02
	4	1.12e+00	0.290	[-1.52e-02, 4.47e-03]	3.41e-02	3.95e-02
	5	3.54e-02	0.851	[-3.85e-03, 4.66e-03]	6.73e-03	6.33e-03
	6	2.77e+00	0.096	[-6.11e-03, 3.13e-04]	2.52e-03	5.42e-03
	8	5.64e-02	0.812	[-8.94e-04, 1.13e-03]	4.21e-04	3.01e-04
V	0	3.29e-02	0.856	[-2.08e-02, 1.73e-02]	1.53e-01	1.55e-01
	1	1.45e+00	0.229	[-9.67e-03, 4.06e-02]	6.55e-01	6.39e-01
	2	1.76e+00	0.185	[-3.30e-02, 6.27e-03]	1.61e-01	1.75e-01
	3	1.91e-02	0.890	[-8.59e-03, 7.45e-03]	2.35e-02	2.41e-02
	4	4.03e-01	0.526	[-2.43e-03, 4.68e-03]	5.04e-03	3.92e-03
	5	1.12e-01	0.738	[-1.19e-03, 1.67e-03]	8.40e-04	6.02e-04
	8	2.17e+00	0.140	[-3.05e-03, 2.76e-04]	4.20e-04	1.81e-03
MV	0	3.02e-01	0.583	[-5.01e-03, 8.87e-03]	1.85e-02	1.66e-02
	1	2.23e-01	0.637	[-2.00e-02, 3.27e-02]	5.03e-01	4.96e-01
	2	2.39e-02	0.877	[-2.34e-02, 2.74e-02]	3.72e-01	3.70e-01
	3	1.67e+00	0.197	[-2.40e-02, 4.84e-03]	7.78e-02	8.74e-02
	4	2.53e-04	0.987	[-7.19e-03, 7.30e-03]	1.93e-02	1.93e-02
	5	4.53e-03	0.946	[-4.19e-03, 3.91e-03]	5.89e-03	6.02e-03
	6	1.19e-01	0.730	[-3.24e-03, 2.26e-03]	2.52e-03	3.01e-03
	8	1.25e-02	0.911	[-2.32e-03, 2.07e-03]	1.68e-03	1.81e-03
DL	0	1.93e-01	0.661	[-2.28e-03, 3.56e-03]	3.35e-03	2.71e-03
	1	3.06e-03	0.956	[-2.21e-02, 2.34e-02]	2.49e-01	2.48e-01
	2	3.35e+00	0.067	[-1.74e-03, 5.07e-02]	4.76e-01	4.51e-01
	3	5.58e-01	0.455	[-2.79e-02, 1.25e-02]	1.76e-01	1.84e-01
	4	**3.89e+00**	**0.049**	**[-2.78e-02, -2.87e-04]**	6.79e-02	8.19e-02
	5	4.96e-01	0.482	[-9.93e-03, 4.64e-03]	1.84e-02	2.11e-02
	6	9.98e-01	0.318	[-6.59e-03, 2.06e-03]	5.87e-03	8.13e-03
	7	2.56e-02	0.873	[-1.94e-03, 2.28e-03]	1.68e-03	1.51e-03
L	0	2.01e-02	0.887	[-1.19e-02, 1.03e-02]	4.68e-02	4.76e-02
	1	5.10e-01	0.475	[-1.64e-02, 3.53e-02]	5.91e-01	5.82e-01
	2	2.65e-04	0.987	[-2.43e-02, 2.39e-02]	3.00e-01	3.00e-01
	3	1.84e+00	0.175	[-1.95e-02, 3.43e-03]	4.68e-02	5.48e-02
	4	7.12e-02	0.790	[-6.40e-03, 4.86e-03]	1.13e-02	1.21e-02
	5	4.56e-02	0.831	[-3.12e-03, 3.87e-03]	4.60e-03	4.22e-03
	8	2.87e-02	0.865	[-3.36e-03, 2.82e-03]	3.35e-03	3.61e-03
ML	0	2.87e-02	0.865	[-3.36e-03, 2.82e-03]	3.35e-03	3.61e-03
	1	3.13e-02	0.860	[-2.05e-02, 2.46e-02]	2.43e-01	2.41e-01
	2	5.98e-01	0.439	[-1.59e-02, 3.67e-02]	4.96e-01	4.85e-01
	3	8.60e-01	0.354	[-3.01e-02, 1.07e-02]	1.81e-01	1.90e-01
	4	1.56e-01	0.693	[-1.45e-02, 9.63e-03]	5.48e-02	5.72e-02
	5	1.69e-02	0.897	[-6.77e-03, 5.93e-03]	1.46e-02	1.51e-02
	6	1.42e-02	0.905	[-3.87e-03, 4.37e-03]	6.27e-03	6.02e-03
	8	1.09e-01	0.742	[-1.19e-03, 1.66e-03]	8.37e-04	6.02e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.17: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
16DV	1.745	0.187	[-0.019, 0.004]	0.044	0.052
16V	0.009	0.924	[-0.004, 0.005]	0.007	0.007
16MV	0.023	0.879	[-0.010, 0.008]	0.029	0.030
16DL	4.744	**0.029**	**[-0.034, -0.002]**	0.096	0.115
16L	0.014	0.908	[-0.007, 0.006]	0.016	0.016
16ML	0.115	0.735	[-0.017, 0.012]	0.077	0.080

PPD Statistics Before and After H-D Imputation by Site

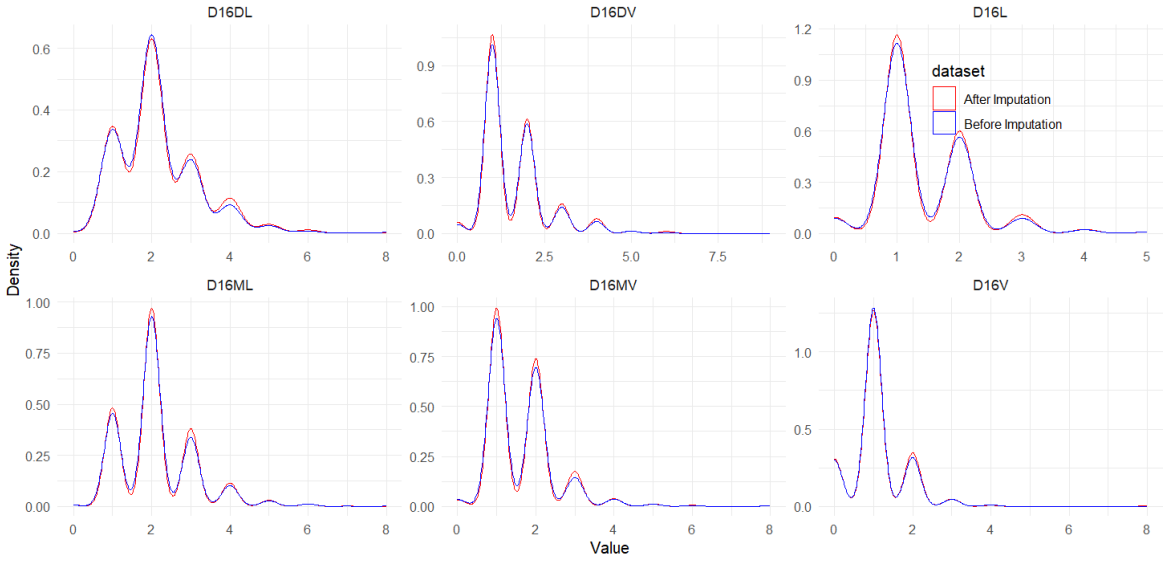
Table II.18: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 16		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 3976425	0.580	D = 0.012
	Variance	0.821	0.906	F = 1.885	0.170	p = 0.985
V	Median	1	1	W = 3989638	0.456	D = 0.014
	Variance	0.508	0.569	F = 1.462	0.227	p = 0.957
MV	Median	1	1	W = 4000295	0.358	D = 0.010
	Variance	0.737	0.763	F = 0.703	0.402	p = 0.999
DL	Median	2	2	W = 4050528	0.119	D = 0.026
	Variance	1.08	1.15	F = 4.334	**0.037**	p = 0.315
L	Median	1	1	W = 4014194	0.458	D = 0.009
	Variance	0.550	0.571	F = 0.934	0.334	p = 0.999
ML	Median	2	2	W = 4008769	0.487	D = 0.012
	Variance	0.926	0.933	F = 0.377	0.539	p = 0.987

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.6



Kernel Density Plot of Imputed and Original 11 by Site

II.7 Upper Right Second Molar (17)

Original vs. H-D Imputed PPD by Site

Table II.19: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	3.24e-01	0.570	[-2.85e-03, 5.12e-03]	6.26e-03	5.12e-03
	1	2.18e-01	0.641	[-3.25e-02, 2.00e-02]	5.24e-01	5.30e-01
	2	1.44e-01	0.704	[-2.05e-02, 3.03e-02]	3.76e-01	3.71e-01
	3	8.76e-03	0.925	[-1.33e-02, 1.21e-02]	6.17e-02	6.24e-02
	4	2.01e-01	0.654	[-5.93e-03, 9.42e-03]	2.25e-02	2.08e-02
	5	6.92e-02	0.793	[-5.24e-03, 3.99e-03]	7.51e-03	8.13e-03
	6	6.45e-02	0.800	[-2.19e-03, 1.68e-03]	1.25e-03	1.51e-03
	8	7.64e-03	0.930	[-1.61e-03, 1.47e-03]	8.34e-04	9.04e-04
V	0	7.21e-01	0.396	[-8.30e-03, 2.09e-02]	8.70e-02	8.07e-02
	1	5.88e-02	0.808	[-2.26e-02, 2.89e-02]	5.99e-01	5.96e-01
	2	6.31e-02	0.802	[-2.57e-02, 1.99e-02]	2.50e-01	2.53e-01
	3	3.01e-01	0.583	[-1.42e-02, 7.96e-03]	4.54e-02	4.85e-02
	4	1.55e+00	0.213	[-1.05e-02, 2.25e-03]	1.33e-02	1.75e-02
	5	1.36e-01	0.713	[-2.22e-03, 3.23e-03]	2.91e-03	2.41e-03
	6	4.45e-03	0.947	[-2.51e-03, 2.69e-03]	2.50e-03	2.41e-03
	7	5.29e-02	0.818	[-8.92e-04, 1.12e-03]	4.16e-04	3.01e-04
MV	0	2.67e+00	0.102	[-1.16e-02, 8.81e-04]	1.21e-02	1.75e-02
	1	1.30e-01	0.718	[-2.13e-02, 3.09e-02]	4.49e-01	4.44e-01
	2	7.93e-02	0.778	[-2.19e-02, 2.92e-02]	3.84e-01	3.81e-01
	3	5.86e-02	0.809	[-1.84e-02, 1.44e-02]	1.09e-01	1.11e-01
	4	2.16e-01	0.642	[-1.18e-02, 7.24e-03]	3.30e-02	3.52e-02
	5	5.77e-05	0.994	[-4.56e-03, 4.52e-03]	7.51e-03	7.53e-03
	6	4.85e-01	0.486	[-1.98e-03, 4.07e-03]	3.76e-03	2.71e-03
	7	1.61e-01	0.688	[-1.40e-03, 2.10e-03]	1.25e-03	9.04e-04
	8	9.08e-02	0.763	[-1.35e-03, 9.84e-04]	4.17e-04	6.02e-04
DL	0	2.72e-01	0.602	[-2.39e-03, 4.08e-03]	4.16e-03	3.31e-03
	1	4.73e+00	**2.96e-02**	**[2.29e-03, 4.64e-02]**	2.39e-01	2.15e-01
	2	2.14e-01	0.644	[-3.23e-02, 2.00e-02]	4.57e-01	4.64e-01
	3	1.35e+00	0.246	[-3.30e-02, 8.39e-03]	1.87e-01	1.99e-01
	4	1.10e+00	0.294	[-2.08e-02, 6.24e-03]	6.86e-02	7.59e-02
	5	4.55e-03	0.946	[-8.56e-03, 9.17e-03]	2.95e-02	2.92e-02
	6	1.00e-01	0.751	[-6.17e-03, 4.45e-03]	9.98e-03	1.08e-02
	7	3.15e-01	0.575	[-1.78e-03, 3.15e-03]	2.50e-03	1.81e-03
	8	1.57e-01	0.692	[-1.40e-03, 2.09e-03]	1.25e-03	9.04e-04
L	0	1.57e-02	0.900	[-7.62e-03, 8.66e-03]	2.49e-02	2.44e-02
	1	4.99e-01	0.480	[-1.67e-02, 3.55e-02]	4.60e-01	4.50e-01
	2	3.72e-02	0.847	[-2.81e-02, 2.31e-02]	3.90e-01	3.93e-01
	3	4.13e-01	0.521	[-2.10e-02, 1.06e-02]	9.84e-02	1.04e-01
	4	3.14e-01	0.575	[-9.36e-03, 5.17e-03]	1.87e-02	2.08e-02
	5	4.24e-02	0.837	[-4.92e-03, 3.98e-03]	7.06e-03	7.53e-03
	6	1.56e-01	0.693	[-1.40e-03, 2.08e-03]	1.25e-03	9.04e-04
ML	0	1.04e-01	0.747	[-1.19e-03, 1.65e-03]	8.31e-04	6.02e-04
	1	1.14e-02	0.915	[-2.09e-02, 1.87e-02]	1.72e-01	1.73e-01
	2	9.09e-01	0.340	[-1.35e-02, 3.90e-02]	5.25e-01	5.12e-01
	3	1.02e+00	0.313	[-3.18e-02, 1.01e-02]	1.95e-01	2.06e-01
	4	9.89e-02	0.753	[-1.57e-02, 1.14e-02]	7.10e-02	7.32e-02
	5	5.07e-03	0.943	[-7.73e-03, 8.32e-03]	2.41e-02	2.38e-02
	6	1.03e-01	0.748	[-4.19e-03, 5.83e-03]	9.55e-03	8.74e-03
	7	6.85e-04	0.979	[-2.43e-03, 2.36e-03]	2.08e-03	2.11e-03
	8	1.93e-03	0.965	[-1.80e-03, 1.88e-03]	1.25e-03	1.21e-03

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.20: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
17DV	2.89e-02	0.865	[-8.42e-03, 1.00e-02]	3.21e-02	3.13e-02
17V	7.97e-01	0.372	[-1.09e-02, 4.02e-03]	1.92e-02	2.26e-02
17MV	3.65e-02	0.849	[-1.21e-02, 9.97e-03]	4.59e-02	4.70e-02
17DL	6.12e-01	0.434	[-2.35e-02, 1.00e-02]	1.12e-01	1.19e-01
17L	2.51e-01	0.616	[-1.09e-02, 6.42e-03]	2.70e-02	2.92e-02
17ML	1.63e-02	0.899	[-1.74e-02, 1.53e-02]	1.08e-01	1.09e-01

PPD Statistics Before and After H-D Imputation by Site

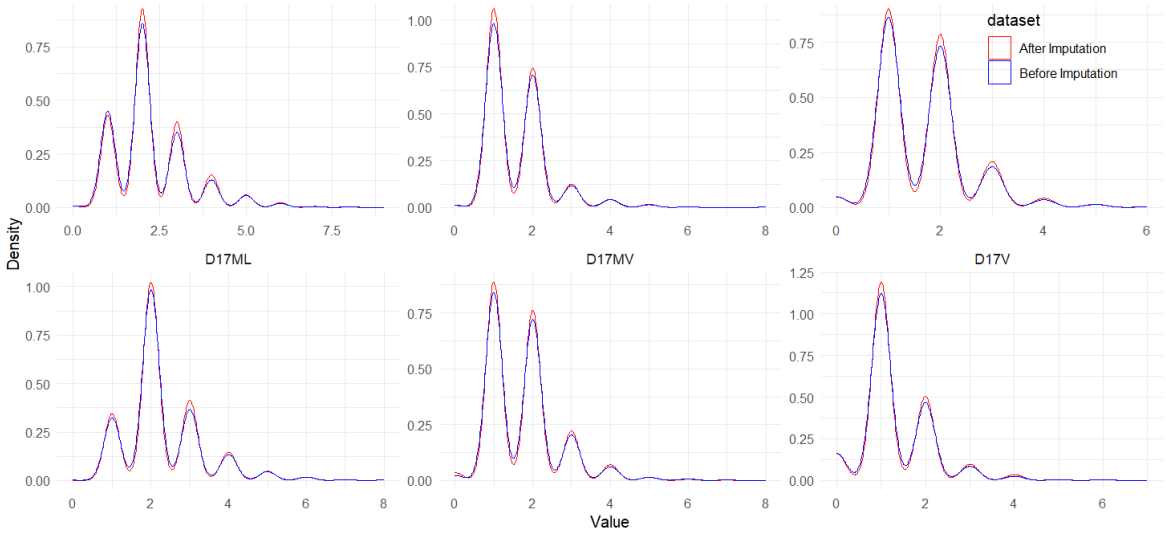
Table II.21: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 17		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 3962640	0.766	D = 0.005
	Variance	0.661	0.666	F = 0.0728	0.787	p = 1
V	Median	1	1	W = 4045030	0.287	D = 0.009
	Variance	0.660	0.676	F = 0.3886	0.533	p = 0.999
MV	Median	2	2	W = 3972103	0.926	D = 0.005
	Variance	0.849	0.855	F = 0.1938	0.660	p = 1
DL	Median	2	2	W = 4118499	**0.030**	D = 0.025
	Variance	1.22	1.18	F = 0.0157	0.900	p = 0.340
L	Median	2	2	W = 4047552	0.376	D = 0.010
	Variance	0.695	0.710	F = 0.004	0.949	p = 0.999
ML	Median	2	2	W = 4029911	0.566	D = 0.012
	Variance	1.04	1.03	F = 0.2659	0.606	p = 0.989

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.7



Kernel Density Plot of Imputed and Original 17 by Site

II.8 Upper Left Central Incisor (21)

Original vs. H-D Imputed PPD by Site

Table II.22: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	3.00e-02	0.863	[-1.17e-02, 9.76e-03]	4.73e-02	4.82e-02
	1	1.32e-01	0.716	[-2.00e-02, 2.91e-02]	6.04e-01	5.99e-01
	2	2.83e-04	0.987	[-2.24e-02, 2.20e-02]	2.66e-01	2.66e-01
	3	4.88e-02	0.825	[-1.24e-02, 9.87e-03]	5.11e-02	5.24e-02
	4	4.21e-01	0.516	[-8.92e-03, 4.46e-03]	1.71e-02	1.93e-02
	5	6.74e-03	0.935	[-4.58e-03, 4.98e-03]	9.24e-03	9.04e-03
	6	1.69e-02	0.897	[-2.61e-03, 2.98e-03]	3.20e-03	3.01e-03
	7	1.37e-02	0.907	[-8.59e-04, 9.67e-04]	3.55e-04	3.01e-04
	8	1.40e-01	0.709	[-2.39e-03, 1.62e-03]	1.42e-03	1.81e-03
V	0	7.80e-01	0.377	[-1.21e-02, 3.20e-02]	2.65e-01	2.55e-01
	1	1.55e-01	0.694	[-2.99e-02, 1.99e-02]	5.64e-01	5.69e-01
	2	9.97e-02	0.752	[-2.06e-02, 1.49e-02]	1.45e-01	1.48e-01
	3	1.51e-01	0.697	[-8.31e-03, 5.55e-03]	1.88e-02	2.02e-02
	4	1.18e-02	0.913	[-3.42e-03, 3.82e-03]	5.32e-03	5.12e-03
	5	2.68e-02	0.870	[-1.18e-03, 1.40e-03]	7.10e-04	6.02e-04
	9	2.07e+00	0.150	[-2.64e-03, 3.40e-04]	3.55e-04	1.51e-03
	10	1.34e-02	0.908	[-8.58e-04, 9.65e-04]	3.55e-04	3.01e-04
MV	0	2.93e-02	0.864	[-1.21e-02, 1.02e-02]	5.17e-02	5.27e-02
	1	3.36e-02	0.855	[-2.23e-02, 2.69e-02]	5.96e-01	5.93e-01
	2	4.82e-04	0.983	[-2.19e-02, 2.24e-02]	2.65e-01	2.65e-01
	3	3.09e-03	0.956	[-1.22e-02, 1.29e-02]	6.66e-02	6.63e-02
	4	4.56e-03	0.946	[-5.85e-03, 5.46e-03]	1.28e-02	1.30e-02
	5	1.35e-01	0.713	[-4.47e-03, 3.05e-03]	5.32e-03	6.02e-03
	6	7.79e-01	0.378	[-3.78e-03, 1.40e-03]	2.13e-03	3.31e-03
	7	2.65e-02	0.871	[-1.18e-03, 1.40e-03]	7.09e-04	6.02e-04
	10	1.32e-02	0.908	[-8.58e-04, 9.65e-04]	3.54e-04	3.01e-04
DL	0	1.33e-01	0.715	[-8.18e-03, 1.19e-02]	4.25e-02	4.07e-02
	1	3.34e-03	0.954	[-2.44e-02, 2.59e-02]	5.14e-01	5.14e-01
	2	5.08e-04	0.982	[-2.37e-02, 2.31e-02]	3.16e-01	3.17e-01
	3	2.61e-02	0.872	[-1.55e-02, 1.31e-02]	8.83e-02	8.95e-02
	4	2.09e-03	0.964	[-7.82e-03, 8.19e-03]	2.61e-02	2.59e-02
	5	4.78e-02	0.827	[-5.12e-03, 4.09e-03]	8.22e-03	8.74e-03
	6	2.72e-01	0.602	[-4.19e-03, 2.42e-03]	3.93e-03	4.82e-03
	7	1.47e-02	0.904	[-8.60e-04, 9.72e-04]	3.57e-04	3.01e-04
L	0	1.10e+00	0.294	[-8.51e-03, 2.81e-02]	1.62e-01	1.52e-01
	1	3.07e-03	0.956	[-2.54e-02, 2.40e-02]	5.86e-01	5.87e-01
	2	3.25e-01	0.569	[-2.56e-02, 1.41e-02]	1.91e-01	1.97e-01
	3	2.51e-01	0.616	[-1.19e-02, 7.07e-03]	3.61e-02	3.86e-02
	4	1.19e-01	0.730	[-7.26e-03, 5.08e-03]	1.49e-02	1.60e-02
	5	1.81e-03	0.966	[-4.42e-03, 4.23e-03]	7.44e-03	7.53e-03
	6	7.89e-02	0.779	[-1.91e-03, 2.55e-03]	2.13e-03	1.81e-03
ML	0	1.12e-01	0.737	[-3.91e-03, 5.51e-03]	9.24e-03	8.43e-03
	1	1.55e-01	0.694	[-1.96e-02, 2.94e-02]	3.91e-01	3.86e-01
	2	7.65e-02	0.782	[-2.13e-02, 2.83e-02]	4.27e-01	4.24e-01
	3	4.77e-01	0.490	[-2.28e-02, 1.09e-02]	1.27e-01	1.33e-01
	4	5.89e-02	0.808	[-1.03e-02, 8.04e-03]	3.41e-02	3.52e-02
	5	8.45e-01	0.358	[-7.21e-03, 2.58e-03]	8.53e-03	1.08e-02
	6	3.81e-04	0.984	[-2.29e-03, 2.33e-03]	2.13e-03	2.11e-03
	7	2.73e-02	0.869	[-1.18e-03, 1.40e-03]	7.10e-04	6.02e-04
	8	1.37e-02	0.907	[-8.59e-04, 9.67e-04]	3.55e-04	3.01e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.23: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
21DV	0.229	0.632	[-1.10e-02, 6.70e-03]	3.13e-02	3.34e-02
21V	0.092	0.762	[-5.09e-03, 3.72e-03]	7.45e-03	8.13e-03
21MV	0.261	0.610	[-9.32e-03, 5.45e-03]	2.13e-02	2.32e-02
21DL	0.054	0.816	[-1.09e-02, 8.59e-03]	3.86e-02	3.98e-02
21L	0.047	0.828	[-8.68e-03, 6.94e-03]	2.44e-02	2.53e-02
21ML	0.359	0.549	[-1.39e-02, 7.39e-03]	4.58e-02	4.91e-02

PPD Statistics Before and After H-D Imputation by Site

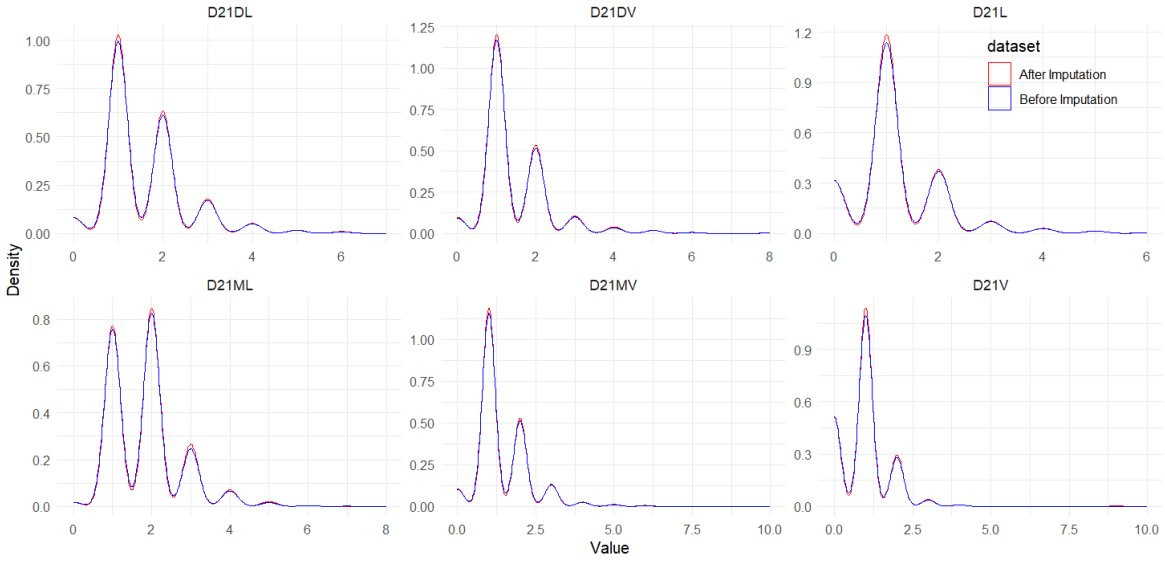
Table II.24: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 21		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4690520	0.770	D = 3.62e-03
	Variance	0.791	0.818	F = 0.258	0.612	p = 1
V	Median	1	1	W = 4736032	0.361	D = 9.92e-03
	Variance	0.611	0.674	F = 0.030	0.863	p = 0.998
MV	Median	1	1	W = 4690267	0.925	D = 1.93e-03
	Variance	0.711	0.739	F = 0.166	0.684	p = 1
DL	Median	1	1	W = 4664852	0.748	D = 2.61e-03
	Variance	0.850	0.867	F = 0.082	0.775	p = 1
L	Median	1	1	W = 4758864	0.248	D = 9.77e-03
	Variance	0.769	0.770	F = 0.024	0.877	p = 0.999
ML	Median	2	2	W = 4723042	0.435	D = 9.23e-03
	Variance	0.803	0.828	F = 0.214	0.644	p = 0.999

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.8



Kernel Density Plot of Imputed and Original 21 by Site

II.9 Upper Left Lateral Incisor (22)

Original vs. H-D Imputed PPD by Site

Table II.25: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	3.86e-02	0.844	[-8.62e-03, 1.05e-02]	3.80e-02	3.71e-02
	1	3.82e-01	0.537	[-1.70e-02, 3.27e-02]	5.80e-01	5.72e-01
	2	1.63e-02	0.899	[-2.45e-02, 2.15e-02]	2.98e-01	2.99e-01
	3	5.72e-01	0.449	[-1.66e-02, 7.32e-03]	5.77e-02	6.24e-02
	4	1.69e-02	0.897	[-7.16e-03, 6.27e-03]	1.79e-02	1.84e-02
	5	1.28e-03	0.972	[-3.84e-03, 3.98e-03]	6.10e-03	6.02e-03
	6	1.65e+00	0.199	[-3.91e-03, 7.56e-04]	1.43e-03	3.01e-03
	7	1.52e-02	0.902	[-8.60e-04, 9.75e-04]	3.59e-04	3.01e-04
	8	8.24e-01	0.364	[-2.44e-03, 8.62e-04]	7.17e-04	1.51e-03
V	0	7.64e-01	0.382	[-1.25e-02, 3.25e-02]	2.81e-01	2.71e-01
	1	3.41e-02	0.854	[-2.72e-02, 2.25e-02]	5.76e-01	5.78e-01
	2	5.77e-01	0.448	[-2.33e-02, 1.03e-02]	1.25e-01	1.31e-01
	3	2.76e-01	0.599	[-8.14e-03, 4.69e-03]	1.57e-02	1.75e-02
	4	7.44e-02	0.785	[-1.77e-03, 2.33e-03]	1.79e-03	1.51e-03
	5	2.97e-02	0.863	[-1.18e-03, 1.41e-03]	7.16e-04	6.02e-04
	8	1.49e-02	0.903	[-8.60e-04, 9.73e-04]	3.58e-04	3.01e-04
MV	0	4.35e-02	0.835	[-9.51e-03, 1.18e-02]	4.75e-02	4.64e-02
	1	6.34e-01	0.426	[-1.48e-02, 3.51e-02]	5.71e-01	5.61e-01
	2	3.11e-01	0.577	[-2.90e-02, 1.62e-02]	2.78e-01	2.84e-01
	3	2.06e-02	0.886	[-1.39e-02, 1.20e-02]	7.07e-02	7.17e-02
	4	2.92e-01	0.589	[-9.50e-03, 5.38e-03]	2.14e-02	2.35e-02
	5	2.39e-02	0.877	[-3.45e-03, 4.04e-03]	5.72e-03	5.42e-03
	6	3.40e-01	0.560	[-4.67e-03, 2.52e-03]	4.65e-03	5.72e-03
DL	0	4.63e-04	0.983	[-8.90e-03, 8.70e-03]	3.15e-02	3.16e-02
	1	2.82e-01	0.596	[-1.84e-02, 3.20e-02]	5.06e-01	4.99e-01
	2	1.98e-01	0.656	[-2.94e-02, 1.85e-02]	3.46e-01	3.51e-01
	3	1.39e-01	0.709	[-1.65e-02, 1.12e-02]	8.17e-02	8.43e-02
	4	1.32e-02	0.909	[-7.83e-03, 6.96e-03]	2.19e-02	2.23e-02
	5	2.73e-01	0.601	[-2.85e-03, 4.91e-03]	6.45e-03	5.42e-03
	6	1.82e-01	0.670	[-2.49e-03, 3.86e-03]	4.30e-03	3.61e-03
	7	3.02e-02	0.862	[-1.18e-03, 1.41e-03]	7.17e-04	6.02e-04
	8	5.47e-03	0.941	[-2.00e-03, 1.85e-03]	1.43e-03	1.51e-03
	9	1.51e-02	0.902	[-8.60e-04, 9.74e-04]	3.58e-04	3.01e-04
L	0	2.40e-01	0.625	[-1.38e-02, 2.29e-02]	1.60e-01	1.56e-01
	1	9.80e-02	0.754	[-2.87e-02, 2.08e-02]	5.86e-01	5.90e-01
	2	1.58e-03	0.968	[-2.04e-02, 1.96e-02]	1.98e-01	1.98e-01
	3	1.72e-02	0.896	[-1.01e-02, 8.82e-03]	3.64e-02	3.71e-02
	4	2.01e-02	0.887	[-4.60e-03, 5.32e-03]	1.00e-02	9.64e-03
	5	1.67e-03	0.967	[-4.33e-03, 4.16e-03]	7.14e-03	7.23e-03
	6	4.16e-04	0.984	[-2.15e-03, 2.11e-03]	1.79e-03	1.81e-03
ML	0	3.75e-01	0.540	[-7.23e-03, 1.38e-02]	4.63e-02	4.31e-02
	1	6.82e-01	0.409	[-1.47e-02, 3.60e-02]	5.31e-01	5.21e-01
	2	4.01e-01	0.526	[-3.09e-02, 1.58e-02]	3.03e-01	3.10e-01
	3	9.14e-01	0.339	[-2.17e-02, 7.42e-03]	8.72e-02	9.43e-02
	4	7.70e-02	0.782	[-6.38e-03, 8.48e-03]	2.24e-02	2.14e-02
	5	3.17e-02	0.859	[-4.90e-03, 4.08e-03]	7.72e-03	8.13e-03
	6	8.01e-02	0.777	[-1.60e-03, 2.13e-03]	1.47e-03	1.21e-03

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.26: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
22DV	0.401	0.526	[-1.10e-02, 5.59e-03]	2.65e-02	2.92e-02
22V	0.149	0.700	[-2.33e-03, 3.46e-03]	3.58e-03	3.01e-03
22MV	0.677	0.411	[-1.31e-02, 5.32e-03]	3.29e-02	3.67e-02
22DL	0.087	0.768	[-7.81e-03, 1.06e-02]	3.51e-02	3.37e-02
22L	0.014	0.908	[-6.59e-03, 7.42e-03]	2.00e-02	1.96e-02
22ML	0.026	0.871	[-8.19e-03, 9.66e-03]	3.24e-02	3.16e-02

PPD Statistics Before and After H-D Imputation by Site

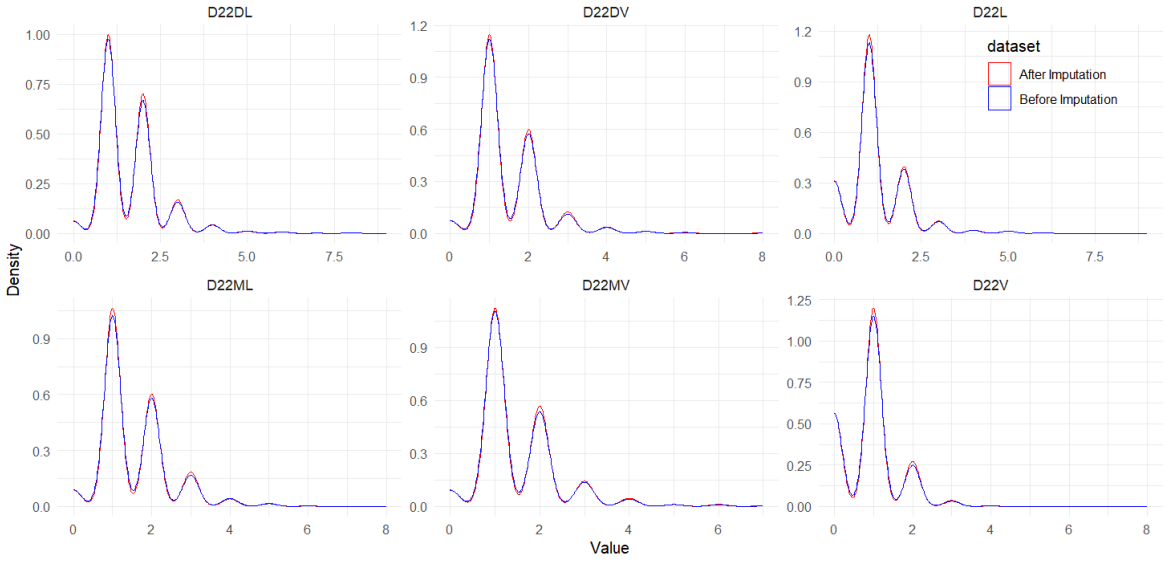
Table II.27: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 22		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4681564	0.393	D = 0.009
	Variance	0.691	0.764	F = 1.327	0.249	**p = 0.001**
V	Median	1	1	W = 4705147	0.283	D = 0.010
	Variance	0.519	0.516	F = 0.025	0.875	p = 0.998
MV	Median	1	1	W = 4703843	0.349	D = 0.011
	Variance	0.809	0.870	F = 1.165	0.281	p = 0.991
DL	Median	1	1	W = 4660903	0.657	D = 0.007
	Variance	0.861	0.839	F = 0.031	0.860	p = 1
L	Median	1	1	W = 4667941	0.765	D = 0.005
	Variance	0.772	0.759	F = 0.061	0.805	p = 1
ML	Median	1	1	W = 4585401	0.241	D = 0.014
	Variance	0.799	0.802	F = 0.595	0.441	p = 0.934

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.9



Kernel Density Plot of Imputed and Original 22 by Site

II.10 Upper Left Canine (23)

Original vs. H-D Imputed PPD by Site

Table II.28: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	4.92e-02	0.825	[-7.23e-03, 5.75e-03]	1.70e-02	1.78e-02
	1	6.93e-03	0.934	[-2.37e-02, 2.58e-02]	5.44e-01	5.43e-01
	2	6.89e-03	0.934	[-2.24e-02, 2.44e-02]	3.33e-01	3.32e-01
	3	9.86e-05	0.992	[-1.33e-02, 1.32e-02]	7.73e-02	7.74e-02
	4	5.21e-02	0.820	[-7.66e-03, 6.06e-03]	1.91e-02	1.99e-02
	5	3.84e-02	0.845	[-3.34e-03, 4.08e-03]	5.79e-03	5.42e-03
	6	2.89e-01	0.591	[-3.08e-03, 1.74e-03]	2.04e-03	2.71e-03
	7	7.61e-03	0.931	[-8.52e-04, 9.31e-04]	3.41e-04	3.01e-04
	8	1.52e-02	0.902	[-1.18e-03, 1.34e-03]	6.81e-04	6.02e-04
	9	2.23e-01	0.637	[-1.33e-03, 8.07e-04]	3.41e-04	6.02e-04
V	0	3.61e-01	0.548	[-1.55e-02, 2.92e-02]	2.84e-01	2.77e-01
	1	1.51e-02	0.902	[-2.59e-02, 2.29e-02]	5.94e-01	5.96e-01
	2	3.07e-01	0.580	[-1.92e-02, 1.07e-02]	9.82e-02	1.02e-01
	3	4.98e-03	0.944	[-5.68e-03, 6.11e-03]	1.44e-02	1.42e-02
	4	1.91e-01	0.662	[-4.65e-03, 2.95e-03]	5.47e-03	6.32e-03
	5	2.18e-01	0.641	[-3.19e-03, 1.96e-03]	2.40e-03	3.01e-03
	6	2.44e-02	0.876	[-1.42e-03, 1.67e-03]	1.03e-03	9.04e-04
MV	0	9.81e-02	0.754	[-7.71e-03, 1.06e-02]	3.61e-02	3.46e-02
	1	4.61e-03	0.946	[-2.37e-02, 2.54e-02]	5.78e-01	5.77e-01
	2	1.15e-01	0.735	[-2.66e-02, 1.88e-02]	2.94e-01	2.98e-01
	3	3.44e-03	0.953	[-1.17e-02, 1.24e-02]	6.30e-02	6.27e-02
	4	1.99e-01	0.656	[-5.51e-03, 8.74e-03]	2.18e-02	2.02e-02
	5	4.01e-03	0.950	[-2.78e-03, 2.97e-03]	3.41e-03	3.31e-03
	6	4.48e-02	0.832	[-2.94e-03, 2.36e-03]	2.73e-03	3.01e-03
	7	7.57e-03	0.931	[-8.52e-04, 9.30e-04]	3.41e-04	3.01e-04
	8	9.65e-02	0.756	[-1.61e-03, 1.17e-03]	6.81e-04	9.04e-04
	9	2.20e-01	0.641	[-1.62e-03, 1.32e-03]	1.03e-03	1.21e-03
DL	0	6.52e-02	0.799	[-7.41e-03, 5.70e-03]	1.72e-02	1.81e-02
	1	3.24e-01	0.569	[-1.77e-02, 3.21e-02]	4.95e-01	4.88e-01
	2	2.69e-02	0.870	[-2.58e-02, 2.18e-02]	3.50e-01	3.52e-01
	3	1.88e-01	0.664	[-1.81e-02, 1.15e-02]	9.64e-02	9.97e-02
	4	1.37e-02	0.907	[-8.98e-03, 7.97e-03]	2.96e-02	3.01e-02
	5	2.21e-03	0.963	[-3.59e-03, 3.76e-03]	5.51e-03	5.42e-03
	6	1.17e-01	0.732	[-2.67e-03, 3.79e-03]	4.48e-03	3.92e-03
	7	7.54e-01	0.385	[-1.78e-03, 6.66e-04]	3.44e-04	9.04e-04
	8	3.52e-01	0.553	[-2.94e-03, 1.56e-03]	1.72e-03	2.41e-03
	9	3.06e-02	0.869	[-1.55e-03, 1.65e-03]	9.02e-04	9.74e-04
L	0	1.04e-01	0.747	[-1.33e-02, 1.85e-02]	1.17e-01	1.15e-01
	1	1.01e-01	0.750	[-2.04e-02, 2.83e-02]	5.96e-01	5.92e-01
	2	2.83e-01	0.595	[-2.64e-02, 1.51e-02]	2.23e-01	2.29e-01
	3	2.32e-02	0.879	[-1.07e-02, 9.15e-03]	4.14e-02	4.22e-02
	4	4.59e-03	0.946	[-5.85e-03, 6.27e-03]	1.53e-02	1.51e-02
	5	8.25e-06	0.998	[-2.58e-03, 2.58e-03]	2.72e-03	2.71e-03
	6	2.18e-02	0.883	[-3.15e-03, 2.71e-03]	3.39e-03	3.61e-03
	7	1.42e-02	0.905	[-1.18e-03, 1.33e-03]	6.79e-04	6.02e-04
	8	7.11e-03	0.933	[-8.51e-04, 9.27e-04]	3.39e-04	3.01e-04
	9	2.26e-01	0.635	[-1.33e-03, 8.04e-04]	3.39e-04	6.02e-04
ML	0	6.01e-02	0.806	[-8.46e-03, 6.58e-03]	2.29e-02	2.38e-02
	1	1.02e-01	0.750	[-2.08e-02, 2.90e-02]	5.31e-01	5.27e-01
	2	3.52e-04	0.985	[-2.37e-02, 2.32e-02]	3.29e-01	3.30e-01
	3	1.11e-01	0.740	[-1.66e-02, 1.18e-02]	8.80e-02	9.04e-02
	4	8.15e-03	0.928	[-6.63e-03, 6.04e-03]	1.63e-02	1.66e-02
	5	2.12e-03	0.963	[-4.03e-03, 3.85e-03]	6.23e-03	6.33e-03
	6	5.51e-03	0.941	[-2.65e-03, 2.86e-03]	3.12e-03	3.01e-03
	7	3.78e-02	0.846	[-1.83e-03, 1.50e-03]	1.04e-03	1.21e-03
	8	5.03e-03	0.943	[-2.17e-03, 2.02e-03]	1.73e-03	1.81e-03
	9	9.74e-03	0.921	[-8.54e-04, 9.44e-04]	3.46e-04	3.01e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.29: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
23DV	8.51e-02	0.771	[-9.55e-03, 7.08e-03]	2.83e-02	2.95e-02
23V	2.69e-01	0.604	[-6.21e-03, 3.60e-03]	9.24e-03	1.05e-02
23MV	8.72e-02	0.768	[-7.00e-03, 9.48e-03]	2.90e-02	2.77e-02
23DL	4.67e-02	0.829	[-1.11e-02, 8.91e-03]	4.17e-02	4.28e-02
23L	1.72e-03	0.967	[-7.56e-03, 7.25e-03]	2.27e-02	2.29e-02
23ML	1.25e-02	0.911	[-8.84e-03, 7.89e-03]	2.87e-02	2.92e-02

PPD Statistics Before and After H-D Imputation by Site

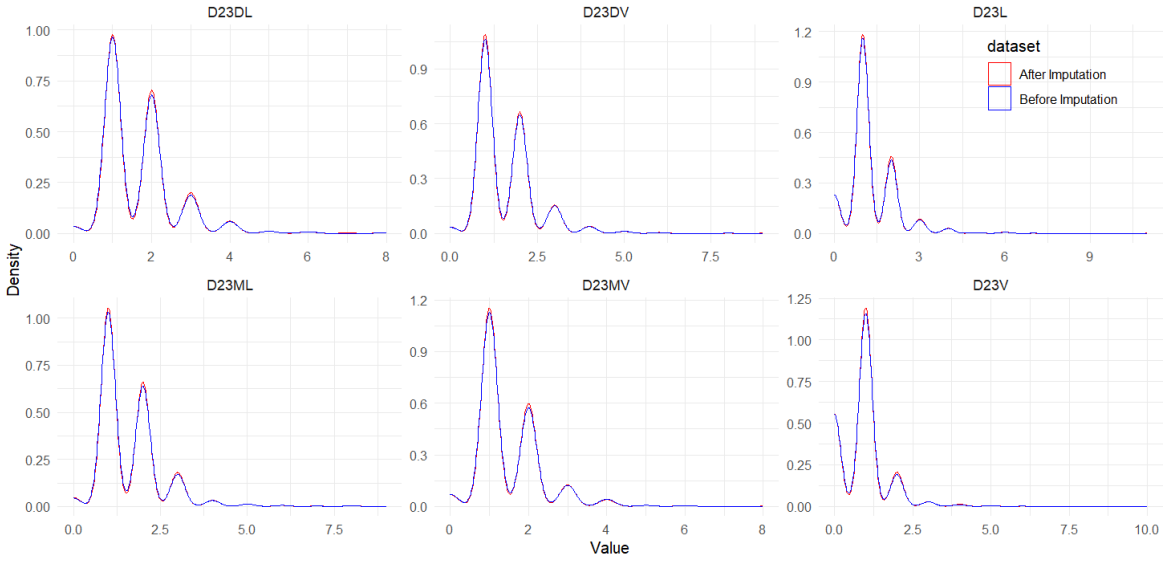
Table II.30: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 23		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4874719	0.967	D = 0.001
	Variance	0.717	0.742	F = 0.067	0.796	p = 1.000
V	Median	1	1	W = 4899951	0.443	D = 0.007
	Variance	0.566	0.577	F = 0.004	0.950	p = 1.000
MV	Median	1	1	W = 4884743	0.862	D = 0.002
	Variance	0.713	0.713	F = 0.001	0.971	p = 1.000
DL	Median	1	1	W = 4854745	0.598	D = 0.006
	Variance	0.864	0.903	F = 0.471	0.493	p = 1.000
L	Median	1	1	W = 4928709	0.561	D = 0.007
	Variance	0.778	0.802	F = 0.122	0.727	p = 1.000
ML	Median	1	1	W = 4810464	0.797	D = 0.003
	Variance	0.821	0.833	F = 0.125	0.724	p = 1.000

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.10



Kernel Density Plot of Imputed and Original 23 by Site

II.11 Upper Left First Premolar(24)

Original vs. H-D Imputed PPD by Site

Table II.31: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	1.49e-01	0.70	[-1.06e-02, 7.11e-03]	2.93e-02	3.10e-02
	1	1.42e-02	0.91	[-2.44e-02, 2.75e-02]	5.18e-01	5.17e-01
	2	2.31e-02	0.88	[-2.30e-02, 2.68e-02]	3.60e-01	3.58e-01
	3	1.11e-01	0.74	[-1.03e-02, 1.45e-02]	6.17e-02	5.96e-02
	4	7.19e-02	0.79	[-6.16e-03, 8.11e-03]	1.97e-02	1.87e-02
	5	3.26e+00	0.07	[-9.48e-03, 2.20e-04]	6.82e-03	1.15e-02
	6	6.87e-02	0.79	[-3.42e-03, 2.61e-03]	3.21e-03	3.61e-03
	7	4.12e-02	0.84	[-8.83e-04, 1.08e-03]	4.01e-04	3.01e-04
V	0	7.71e-01	0.38	[-1.23e-02, 3.22e-02]	2.45e-01	2.35e-01
	1	3.23e-02	0.86	[-2.76e-02, 2.30e-02]	6.13e-01	6.16e-01
	2	8.21e-01	0.36	[-2.47e-02, 9.04e-03]	1.17e-01	1.24e-01
	3	3.56e-02	0.85	[-7.82e-03, 6.44e-03]	1.89e-02	1.96e-02
	4	3.83e-02	0.84	[-2.76e-03, 3.37e-03]	3.62e-03	3.31e-03
	5	1.68e-01	0.68	[-1.57e-03, 2.37e-03]	1.61e-03	1.21e-03
	6	8.39e-02	0.77	[-1.19e-03, 1.59e-03]	8.04e-04	6.02e-04
MV	0	1.47e-01	0.70	[-5.60e-03, 8.30e-03]	1.88e-02	1.75e-02
	1	1.88e-01	0.66	[-2.02e-02, 3.17e-02]	4.95e-01	4.89e-01
	2	7.79e-03	0.93	[-2.61e-02, 2.39e-02]	3.65e-01	3.66e-01
	3	4.76e-02	0.83	[-1.65e-02, 1.32e-02]	8.93e-02	9.10e-02
	4	9.50e-01	0.33	[-1.16e-02, 3.82e-03]	2.08e-02	2.47e-02
	5	2.38e-01	0.63	[-6.12e-03, 3.66e-03]	8.41e-03	9.64e-03
	6	2.46e-01	0.62	[-1.81e-03, 3.00e-03]	2.40e-03	1.81e-03
	7	4.09e-02	0.84	[-8.83e-04, 1.08e-03]	4.01e-04	3.01e-04
DL	0	2.51e-01	0.62	[-2.57e-03, 4.30e-03]	4.78e-03	3.92e-03
	1	1.04e-01	0.75	[-2.15e-02, 2.99e-02]	4.37e-01	4.33e-01
	2	2.48e-04	0.99	[-2.56e-02, 2.52e-02]	4.00e-01	4.01e-01
	3	3.46e-01	0.56	[-2.19e-02, 1.18e-02]	1.18e-01	1.23e-01
	4	1.67e-03	0.97	[-8.35e-03, 8.70e-03]	2.79e-02	2.77e-02
	5	7.39e-02	0.79	[-5.47e-03, 4.13e-03]	8.37e-03	9.04e-03
	6	1.12e-01	0.74	[-2.35e-03, 3.30e-03]	3.19e-03	2.71e-03
	7	3.94e-02	0.84	[-8.82e-04, 1.08e-03]	3.98e-04	3.01e-04
L	0	2.38e-02	0.88	[-1.47e-02, 1.72e-02]	1.06e-01	1.05e-01
	1	5.37e-03	0.94	[-2.45e-02, 2.64e-02]	5.94e-01	5.93e-01
	2	1.47e-01	0.70	[-1.76e-02, 2.61e-02]	2.32e-01	2.28e-01
	3	6.70e-01	0.41	[-1.63e-02, 6.64e-03]	4.94e-02	5.42e-02
	4	1.37e-01	0.71	[-6.13e-03, 4.17e-03]	9.56e-03	1.05e-02
	5	1.61e-01	0.69	[-2.99e-03, 4.51e-03]	5.58e-03	4.82e-03
	6	8.12e-01	0.37	[-4.45e-03, 1.59e-03]	2.79e-03	4.22e-03
	7	2.45e-02	0.88	[-1.61e-03, 1.37e-03]	7.83e-04	9.04e-04
ML	0	1.50e-02	0.90	[-8.37e-03, 7.38e-03]	2.33e-02	2.38e-02
	1	3.59e-01	0.55	[-3.38e-02, 1.80e-02]	5.40e-01	5.48e-01
	2	1.24e-01	0.72	[-1.99e-02, 2.86e-02]	3.21e-01	3.16e-01
	3	4.86e-02	0.83	[-1.30e-02, 1.63e-02]	8.72e-02	8.55e-02
	4	7.33e-02	0.79	[-6.34e-03, 8.37e-03]	2.09e-02	1.99e-02
	5	5.46e-01	0.46	[-2.23e-03, 4.85e-03]	5.22e-03	3.92e-03
	6	3.32e-02	0.86	[-2.34e-03, 1.94e-03]	1.61e-03	1.81e-03
	7	8.36e-02	0.77	[-1.19e-03, 1.59e-03]	8.04e-04	6.02e-04
	8	4.18e-02	0.84	[-8.83e-04, 1.08e-03]	4.02e-04	3.01e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.32: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
D24DV	6.73-01	0.412	[-1.30-02, 5.29-03]	3.05-02	3.43-02
D24V	2.14-01	0.644	[-2.98-03, 4.80-03]	6.03-03	5.12-03
D24MV	7.85-01	0.376	[-1.38-02, 5.14-03]	3.24-02	3.68-02
D24DL	1.19-03	0.973	[-1.00-02, 1.04-02]	4.02-02	4.01-02
D24L	2.10-01	0.647	[-8.66-03, 5.36-03]	1.79-02	1.96-02
D24ML	3.11-01	0.577	[-6.14-03, 1.10-02]	2.89-02	2.65-02

PPD Statistics Before and After H-D Imputation by Site

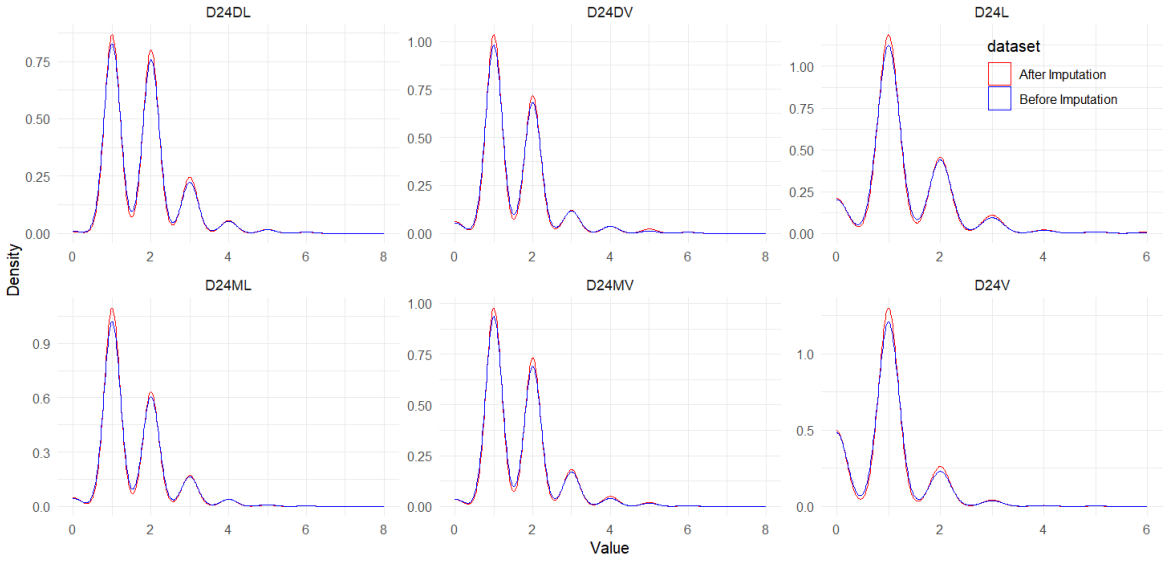
Table II.33: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 24		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4140321	0.996	D = 0.005
	Variance	0.721	0.770	F = 0.300	0.584	p = 1.000
V	Median	1	1	W = 4188050	0.291	D = 0.010
	Variance	0.511	0.499	F = 0.077	0.781	p = 0.999
MV	Median	1	1	W = 4186368	0.475	D = 0.007
	Variance	0.750	0.764	F = 0.473	0.492	p = 1.000
DL	Median	2	2	W = 4196329	0.613	D = 0.005
	Variance	0.783	0.776	F = 0.018	0.894	p = 1.000
L	Median	1	1	W = 4185938	0.730	D = 0.006
	Variance	0.698	0.736	F = 0.325	0.569	p = 1.000
ML	Median	1	1	W = 4092611	0.491	D = 0.008
	Variance	0.727	0.699	F = 0.577	0.448	p = 1.000

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.11



Kernel Density Plot of Imputed and Original 24 by Site

II.12 Upper Left Second PremMolar (25)

Original vs. H-D Imputed PPD by Site

Table II.34: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	3.99e-01	0.53	[-1.08e-02, 5.51e-03]	2.45e-02	2.71e-02
	1	2.47e-01	0.62	[-3.24e-02, 1.93e-02]	4.79e-01	4.86e-01
	2	3.19e-01	0.57	[-1.78e-02, 3.22e-02]	3.74e-01	3.67e-01
	3	4.60e-01	0.50	[-9.09e-03, 1.86e-02]	8.01e-02	7.53e-02
	4	3.21e-01	0.57	[-1.19e-02, 6.56e-03]	3.20e-02	3.46e-02
	5	1.38e-03	0.97	[-4.13e-03, 4.29e-03]	6.71e-03	6.63e-03
	6	4.33e-02	0.84	[-2.19e-03, 2.71e-03]	2.37e-03	2.11e-03
	8	2.43e-01	0.62	[-2.02e-03, 1.19e-03]	7.89e-04	1.21e-03
V	0	2.32	0.13	[-4.72e-03, 3.72e-02]	2.14e-01	1.98e-01
	1	2.21e-01	0.64	[-1.90e-02, 3.09e-02]	6.33e-01	6.27e-01
	2	3.27	0.07	[-3.44e-02, 1.27e-03]	1.31e-01	1.48e-01
	3	1.45	0.23	[-1.09e-02, 2.52e-03]	1.54e-02	1.96e-02
	4	9.24e-01	0.34	[-5.32e-03, 1.76e-03]	3.94e-03	5.72e-03
	5	1.10e-01	0.74	[-1.40e-03, 1.96e-03]	1.18e-03	9.04e-04
	6	5.75e-04	0.98	[-1.81e-03, 1.76e-03]	1.18e-03	1.21e-03
MV	0	8.38e-01	0.36	[-3.95e-03, 1.07e-02]	2.21e-02	1.87e-02
	1	1.38e-01	0.71	[-2.09e-02, 3.06e-02]	4.56e-01	4.51e-01
	2	1.78e-01	0.67	[-2.00e-02, 3.10e-02]	4.18e-01	4.13e-01
	3	1.11	0.29	[-2.17e-02, 6.47e-03]	7.76e-02	8.52e-02
	4	5.52e-01	0.46	[-1.00e-02, 4.48e-03]	1.89e-02	2.17e-02
	5	4.45e-01	0.50	[-5.04e-03, 2.45e-03]	4.73e-03	6.02e-03
	6	2.19	0.14	[-4.60e-03, 5.21e-04]	1.58e-03	3.61e-03
	7	1.22e-01	0.73	[-1.35e-03, 9.29e-04]	3.94e-04	6.02e-04
DL	0	7.33e-03	0.93	[-4.55e-03, 4.96e-03]	8.64e-03	8.43e-03
	1	1.50e-01	0.70	[-2.06e-02, 3.08e-02]	4.53e-01	4.48e-01
	2	1.52e-03	0.97	[-2.47e-02, 2.57e-02]	3.89e-01	3.88e-01
	3	1.25e-01	0.72	[-1.96e-02, 1.36e-02]	1.15e-01	1.18e-01
	4	4.78e-01	0.49	[-4.71e-03, 9.78e-03]	2.12e-02	1.87e-02
	5	6.34e-01	0.43	[-7.24e-03, 3.01e-03]	9.03e-03	1.11e-02
	6	9.28e-02	0.76	[-2.37e-03, 3.23e-03]	3.14e-03	2.71e-03
	7	1.65	0.20	[-3.23e-03, 5.79e-04]	7.86e-04	2.11e-03
	8	4.55	**0.03**	**[-4.25e-03, -3.89e-04]**	3.93e-04	2.71e-03
L	0	2.99e-02	0.86	[-1.37e-02, 1.64e-02]	9.44e-02	9.31e-02
	1	4.95e-01	0.48	[-1.64e-02, 3.48e-02]	5.65e-01	5.56e-01
	2	1.19e-01	0.73	[-2.70e-02, 1.89e-02]	2.71e-01	2.75e-01
	3	2.27e-01	0.63	[-1.48e-02, 9.01e-03]	5.52e-02	5.81e-02
	4	8.34e-01	0.36	[-7.59e-03, 2.71e-03]	9.01e-03	1.14e-02
	5	3.50e-01	0.55	[-4.23e-03, 2.25e-03]	3.53e-03	4.52e-03
	6	1.07e-03	0.97	[-1.81e-03, 1.75e-03]	1.18e-03	1.21e-03
ML	0	8.54e-01	0.36	[-8.22e-03, 2.90e-03]	1.06e-02	1.33e-02
	1	9.58e-02	0.76	[-2.14e-02, 2.95e-02]	4.19e-01	4.15e-01
	2	4.65e-02	0.83	[-2.28e-02, 2.84e-02]	4.36e-01	4.33e-01
	3	8.70e-05	0.99	[-1.56e-02, 1.58e-02]	1.03e-01	1.03e-01
	4	1.52	0.22	[-1.33e-02, 2.95e-03]	2.32e-02	2.83e-02
	5	3.30e-03	0.95	[-3.27e-03, 3.47e-03]	4.32e-03	4.22e-03
	6	2.11e-01	0.65	[-1.83e-03, 2.92e-03]	2.35e-03	1.81e-03
	7	3.51e-02	0.85	[-8.78e-04, 1.06e-03]	3.92e-04	3.01e-04
	8	7.03e-02	0.79	[-1.19e-03, 1.55e-03]	7.85e-04	6.02e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.35: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
25DV	2.65e-01	0.607	[-1.33e-02, 7.72e-03]	4.18e-02	4.46e-02	
25V	3.94e-01	0.530	[-5.84e-03, 2.98e-03]	6.70e-03	8.13e-03	
25MV	1.87e+00	0.172	[-1.48e-02, 2.55e-03]	2.64e-02	3.25e-02	
25DL	3.21e-01	0.571	[-1.24e-02, 6.80e-03]	3.46e-02	3.74e-02	
25L	1.14e+00	0.286	[-1.01e-02, 2.90e-03]	1.45e-02	1.81e-02	
25ML	8.06e-01	0.369	[-1.35e-02, 4.95e-03]	3.10e-02	3.52e-02	

PPD Statistics Before and After H-D Imputation by Site

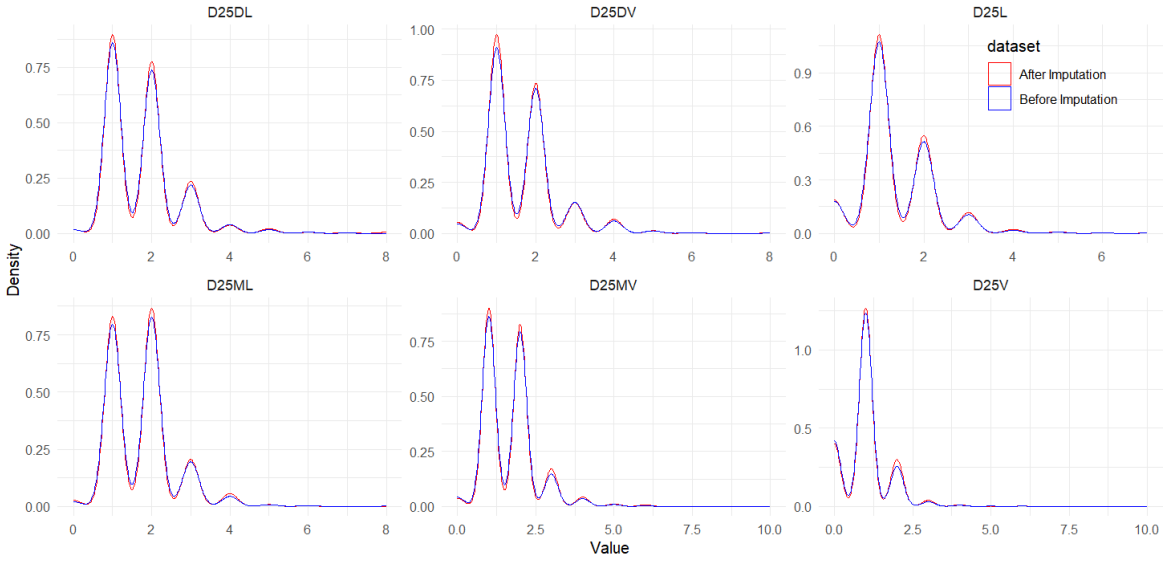
Table II.36: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 25		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	1	1	W = 4166893	0.482	D = 0.009
	Variance	0.790	0.816	F = 0.044	0.834	p = 1
V	Median	1	1	W = 4342589	**0.016**	D = 0.022
	Variance	0.522	0.544	F = 0.587	0.444	p = 0.477
MV	Median	2	2	W = 4281952	0.240	D = 0.014
	Variance	0.687	0.753	F = 0.615	0.433	p = 0.949
DL	Median	2	2	W = 4260340	0.566	D = 0.006
	Variance	0.782	0.912	F = 0.987	0.321	p = 1.000
L	Median	1	1	W = 4288947	0.377	D = 0.011
	Variance	0.667	0.708	F = 1.179	0.278	p = 0.997
ML	Median	2	2	W = 4243066	0.843	D = 0.004
	Variance	0.703	0.714	F = 0.210	0.647	p = 1.000

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.12



Kernel Density Plot of Imputed and Original 25 by Site

II.13 Upper Left First Molar (26)

Original vs. H-D Imputed PPD by Site

Table II.37: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	5.89e-02	0.81	[-9.86e-03, 7.68e-03]	2.78e-02	2.89e-02
	1	1.23	0.27	[-1.11e-02, 4.00e-02]	3.75e-01	3.61e-01
	2	1.16e-01	0.73	[-2.15e-02, 3.05e-02]	4.08e-01	4.03e-01
	3	4.65e-01	0.50	[-2.27e-02, 1.09e-02]	1.11e-01	1.17e-01
	4	3.04	0.08	[-2.31e-02, 1.18e-03]	5.14e-02	6.24e-02
	5	8.62e-03	0.93	[-6.75e-03, 7.42e-03]	1.84e-02	1.81e-02
	6	4.72e-02	0.83	[-4.10e-03, 3.28e-03]	4.71e-03	5.12e-03
	7	9.94e-01	0.32	[-4.25e-03, 1.30e-03]	2.14e-03	3.61e-03
	8	1.25e-01	0.72	[-1.20e-03, 1.70e-03]	8.56e-04	6.02e-04
	10	1.25e-01	0.72	[-1.20e-03, 1.70e-03]	8.56e-04	6.02e-04
V	0	6.12e-01	0.43	[-1.07e-02, 2.48e-02]	1.33e-01	1.26e-01
	1	3.24e-02	0.86	[-2.82e-02, 2.35e-02]	6.04e-01	6.06e-01
	2	1.21e-01	0.73	[-2.59e-02, 1.81e-02]	2.23e-01	2.27e-01
	3	5.12e-01	0.47	[-1.22e-02, 5.63e-03]	2.80e-02	3.13e-02
	4	1.02	0.31	[-2.09e-03, 6.30e-03]	7.22e-03	5.12e-03
	5	6.50e-02	0.80	[-2.61e-03, 3.39e-03]	3.40e-03	3.01e-03
	6	5.97e-02	0.81	[-8.97e-04, 1.14e-03]	4.25e-04	3.01e-04
	7	5.25e-02	0.82	[-2.19e-03, 1.72e-03]	1.27e-03	1.51e-03
	10	5.97e-02	0.81	[-8.97e-04, 1.14e-03]	4.25e-04	3.01e-04
MV	0	2.40e-01	0.62	[-9.52e-03, 5.68e-03]	2.04e-02	2.23e-02
	1	1.00e-01	0.75	[-3.03e-02, 2.19e-02]	4.25e-01	4.30e-01
	2	1.87	0.17	[-7.87e-03, 4.40e-02]	4.12e-01	3.94e-01
	3	1.28	0.26	[-2.56e-02, 6.81e-03]	1.01e-01	1.11e-01
	4	1.47e-01	0.70	[-1.09e-02, 7.34e-03]	3.01e-02	3.19e-02
	5	3.83e-01	0.54	[-6.48e-03, 3.33e-03]	8.07e-03	9.64e-03
	6	1.78e-01	0.67	[-1.40e-03, 2.14e-03]	1.27e-03	9.04e-04
	7	1.78e-01	0.67	[-1.40e-03, 2.14e-03]	1.27e-03	9.04e-04
	10	5.94e-02	0.81	[-8.97e-04, 1.14e-03]	4.24e-04	3.01e-04
DL	0	7.66e-02	0.78	[-6.36e-03, 4.78e-03]	1.10e-02	1.18e-02
	1	1.49	0.22	[-9.53e-03, 4.10e-02]	3.65e-01	3.49e-01
	2	3.26e-03	0.95	[-2.67e-02, 2.52e-02]	4.13e-01	4.14e-01
	3	2.84e-01	0.59	[-2.29e-02, 1.31e-02]	1.33e-01	1.38e-01
	4	9.26e-01	0.34	[-1.83e-02, 6.19e-03]	5.48e-02	6.08e-02
	5	4.51e-01	0.50	[-9.17e-03, 4.45e-03]	1.60e-02	1.84e-02
	6	2.99e-01	0.58	[-5.20e-03, 2.91e-03]	5.48e-03	6.63e-03
	7	8.61e-02	0.77	[-1.36e-03, 9.93e-04]	4.21e-04	6.02e-04
	8	1.71e-01	0.68	[-1.40e-03, 2.12e-03]	1.26e-03	9.04e-04
	9	5.69e-02	0.81	[-8.95e-04, 1.14e-03]	4.21e-04	3.01e-04
L	0	1.08e-01	0.74	[-1.09e-02, 1.53e-02]	6.70e-02	6.48e-02
	1	2.32e-02	0.88	[-2.40e-02, 2.81e-02]	5.76e-01	5.74e-01
	2	1.23e-02	0.91	[-2.27e-02, 2.54e-02]	2.97e-01	2.96e-01
	3	5.42e-02	0.82	[-1.26e-02, 9.89e-03]	4.72e-02	4.85e-02
	4	1.67	0.20	[-9.13e-03, 1.76e-03]	9.26e-03	1.30e-02
	5	2.88e-01	0.59	[-3.60e-03, 2.02e-03]	2.53e-03	3.31e-03
	7	5.67e-02	0.81	[-8.95e-04, 1.13e-03]	4.21e-04	3.01e-04
	8	5.67e-02	0.81	[-8.95e-04, 1.13e-03]	4.21e-04	3.01e-04
ML	0	5.26e-01	0.47	[-8.51e-03, 3.87e-03]	1.30e-02	1.54e-02
	1	5.21e-02	0.82	[-2.30e-02, 2.91e-02]	4.27e-01	4.24e-01
	2	6.37e-02	0.80	[-2.24e-02, 2.90e-02]	3.89e-01	3.86e-01
	3	1.07e-03	0.97	[-1.74e-02, 1.80e-02]	1.30e-01	1.30e-01
	4	6.52e-01	0.42	[-1.23e-02, 5.07e-03]	2.65e-02	3.01e-02
	5	1.80e-01	0.67	[-6.16e-03, 3.95e-03]	8.84e-03	9.94e-03
	6	1.51e-01	0.70	[-2.21e-03, 3.28e-03]	2.95e-03	2.41e-03
	7	5.88e-02	0.81	[-2.19e-03, 1.70e-03]	1.26e-03	1.51e-03
	8	5.64e-02	0.81	[-8.94e-04, 1.13e-03]	4.21e-04	3.01e-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.38: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
26DV	2.540	0.111	[-2.67e-02, 2.60e-03]	7.83e-02	9.04e-02
26V	0.774	0.379	[-3.18e-03, 8.18e-03]	1.27e-02	1.02e-02
26MV	0.212	0.646	[-1.31e-02, 8.11e-03]	4.12e-02	4.37e-02
26DL	1.550	0.213	[-2.37e-02, 5.20e-03]	7.84e-02	8.77e-02
26L	1.671	0.196	[-1.05e-02, 2.04e-03]	1.26e-02	1.69e-02
26ML	0.633	0.426	[-1.49e-02, 6.22e-03]	4.00e-02	4.43e-02

PPD Statistics Before and After H-D Imputation by Site

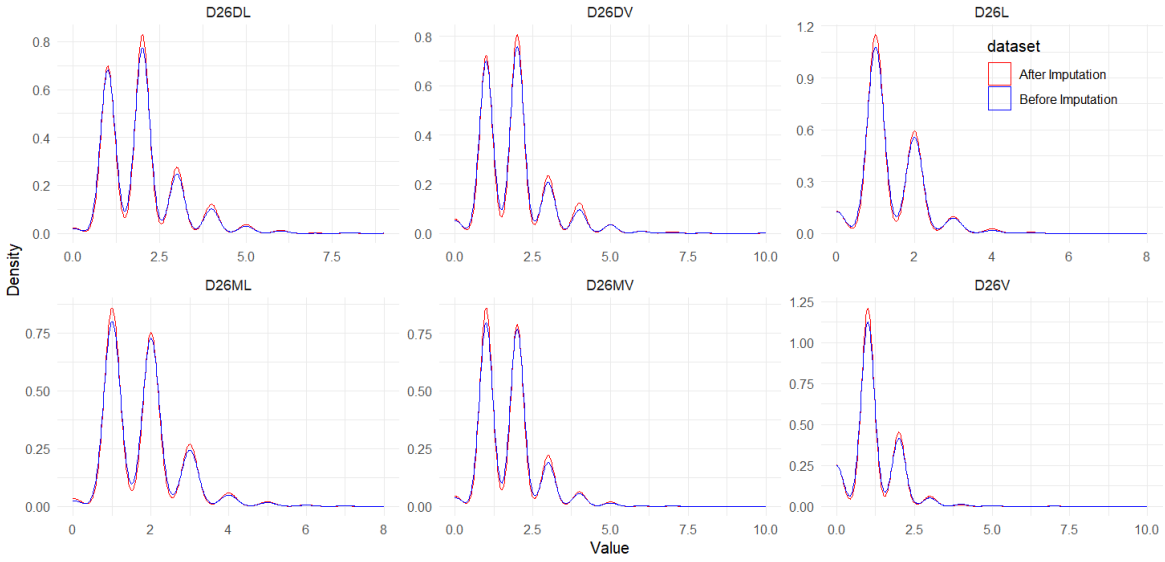
Table II.39: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 26		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	2	2	W = 3960621	0.145	D = 0.01789
	Variance	1.170	1.230	F = 0.847	0.358	p = 0.773
V	Median	1	1	W = 3943518	0.500	D = 7.06e-03
	Variance	0.640	0.617	F = 0.069	0.793	p = 1.000
MV	Median	2	2	W = 3912780	0.974	D = 0.011903
	Variance	0.811	0.833	F = 1.471	0.225	p = 0.990
DL	Median	2	2	W = 4022068	0.149	D = 0.014932
	Variance	1.070	1.110	F = 0.372	0.542	p = 0.917
L	Median	1	1	W = 3971073	0.598	D = 5.57e-03
	Variance	0.577	0.602	F = 0.404	0.525	p = 1.000
ML	Median	2	2	W = 3955062	0.871	D = 4.31e-03
	Variance	0.840	0.867	F = 0.339	0.561	p = 1.000

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value;

Kernel Density Plot of Imputed and Original

Figure II.13



Kernel Density Plot of Imputed and Original 26 by Site

II.14 Upper Left Second Molar (27)

Original vs. H-D Imputed PPD by Site

Table II.40: Comparison of Original and Imputed Periodontal Pocket Depth (PPD) Values Across Different Sites and Unique Value Categories (k)

Site	k	Statistic (D)	P-value	Confidence Interval	Original	Imputed
DV	0	3.64-01	0.55	[-7.33-03, 3.84-03]	1.06-02	1.24-02
	1	4.28-01	0.51	[-1.74-02, 3.49-02]	4.34-01	4.25-01
	2	4.75-03	0.95	[-2.49-02, 2.67-02]	3.97-01	3.96-01
	3	2.67-01	0.61	[-2.02-02, 1.17-02]	9.97-02	1.04-01
	4	1.38-01	0.71	[-1.24-02, 8.45-03]	3.99-02	4.19-02
	5	1.61-01	0.69	[-6.67-03, 4.39-03]	1.06-02	1.18-02
	6	3.63-01	0.55	[-5.18-03, 2.71-03]	5.09-03	6.33-03
	7	1.78-01	0.67	[-1.40-03, 2.14-03]	1.27-03	9.04-04
	8	2.37-01	0.63	[-1.55-03, 2.53-03]	1.70-03	1.21-03
	9	8.28-02	0.77	[-1.36-03, 1.00-03]	4.24-04	6.02-04
V	0	1.44-01	0.70	[-1.12-02, 1.65-02]	7.56-02	7.29-02
	1	5.86-01	0.44	[-1.60-02, 3.65-02]	5.46-01	5.36-01
	2	3.60-01	0.55	[-3.16-02, 1.68-02]	2.98-01	3.05-01
	3	9.43-01	0.33	[-1.85-02, 6.16-03]	5.53-02	6.15-02
	4	1.33-01	0.72	[-5.65-03, 8.22-03]	1.82-02	1.69-02
	5	4.47-01	0.50	[-4.41-03, 2.12-03]	3.38-03	4.52-03
	6	2.94-02	0.86	[-2.57-03, 3.06-03]	2.96-03	2.71-03
	9	5.75-02	0.81	[-8.95-04, 1.14-03]	4.22-04	3.01-04
	10	5.75-02	0.81	[-8.95-04, 1.14-03]	4.22-04	3.01-04
MV	0	2.17-01	0.64	[-6.45-03, 3.95-03]	9.29-03	1.05-02
	1	1.65-01	0.69	[-1.98-02, 3.01-02]	3.38-01	3.33-01
	2	7.31-01	0.39	[-1.47-02, 3.73-02]	4.20-01	4.09-01
	3	4.88-01	0.48	[-2.49-02, 1.18-02]	1.38-01	1.44-01
	4	6.45-01	0.42	[-1.88-02, 7.81-03]	6.59-02	7.14-02
	5	2.37-05	1.00	[-7.75-03, 7.71-03]	2.20-02	2.20-02
	6	6.81-01	0.41	[-4.76-03, 1.88-03]	3.38-03	4.82-03
	7	1.42+00	0.23	[-4.62-03, 1.01-03]	2.11-03	3.92-03
	8	5.78-02	0.81	[-8.95-04, 1.14-03]	4.22-04	3.01-04
	9	5.41-03	0.94	[-1.61-03, 1.50-03]	8.45-04	9.04-04
DL	0	8.91-01	0.35	[-7.88-03, 2.69-03]	9.15-03	1.18-02
	1	4.19-01	0.52	[-1.67-02, 3.31-02]	3.45-01	3.37-01
	2	5.29-02	0.82	[-2.29-02, 2.89-02]	4.19-01	4.16-01
	3	4.74-03	0.95	[-1.88-02, 1.75-02]	1.39-01	1.39-01
	4	1.59+00	0.21	[-2.10-02, 4.46-03]	5.95-02	6.78-02
	5	3.02-02	0.86	[-6.22-03, 7.43-03]	1.75-02	1.69-02
	6	9.49-03	0.92	[-4.22-03, 3.82-03]	5.82-03	6.02-03
	7	3.29-02	0.86	[-3.36-03, 2.79-03]	3.33-03	3.61-03
	8	2.12-03	0.96	[-1.80-03, 1.88-03]	1.25-03	1.21-03
	9	5.26-02	0.82	[-8.92-04, 1.12-03]	4.16-04	3.01-04
L	0	4.67-01	0.49	[-5.31-03, 1.09-02]	2.57-02	2.29-02
	1	2.78-01	0.60	[-1.91-02, 3.32-02]	4.78-01	4.71-01
	2	9.89-03	0.92	[-2.68-02, 2.42-02]	3.87-01	3.88-01
	3	6.78-01	0.41	[-2.10-02, 8.52-03]	8.41-02	9.04-02
	4	4.63-01	0.50	[-9.53-03, 4.58-03]	1.74-02	1.99-02
	5	3.04-04	0.99	[-3.88-03, 3.81-03]	5.39-03	5.42-03
	6	1.54-01	0.69	[-1.40-03, 2.08-03]	1.24-03	9.04-04
	7	6.92-02	0.79	[-2.19-03, 1.67-03]	1.24-03	1.51-03
	9	5.14-02	0.82	[-8.91-04, 1.12-03]	4.14-04	3.01-04
L	0	3.84-02	0.84	[-3.13-03, 3.82-03]	4.56-03	4.22-03
	1	8.22-01	0.36	[-3.48-02, 1.28-02]	2.85-01	2.96-01
	2	6.62-01	0.42	[-1.53-02, 3.70-02]	4.69-01	4.58-01
	3	2.77-01	0.60	[-1.39-02, 2.41-02]	1.57-01	1.52-01
	4	6.65-01	0.41	[-1.78-02, 7.30-03]	5.89-02	6.42-02
	5	2.01-04	0.99	[-6.92-03, 6.82-03]	1.74-02	1.75-02
	6	1.71-03	0.97	[-3.89-03, 4.06-03]	5.81-03	5.72-03
	7	5.16-02	0.82	[-8.91-04, 1.12-03]	4.15-04	3.01-04
	8	1.03-01	0.75	[-1.19-03, 1.65-03]	8.30-04	6.02-04
	9	5.16-02	0.82	[-8.91-04, 1.12-03]	4.15-04	3.01-04
	10	4.79-01	0.49	[-1.80-03, 8.17-04]	4.15-04	9.04-04

Original vs. H-D Imputed PPD > 3 Proportions by Site

Table II.41: Comparison of PPD > 3 Proportions: Original vs. Imputed by Site

Site	Statistic (D)	P-value	Confidence Interval	Original	Imputed
27DV	0.325	0.567	[-1.63e-02, 8.91e-03]	5.90e-02	6.27e-02
27V	0,022	0.881	[-7.61e-03, 8.87e-03]	2.53e-02	2.47e-02
27MV	1.160	0.282	[-2.44e-02, 7.01e-03]	9.46e-02	1.03e-01
27DL	1.067	0.302	[-2.31e-02, 7.09e-03]	8.78e-02	9.58e-02
27L	0.285	0.593	[-1.08e-02, 6.13e-03]	2.57e-02	2.80e-02
27ML	0.485	0.486	[-2.00e-02, 9.47e-03]	8.42e-02	8.95e-02

PPD Statistics Before and After H-D Imputation by Site

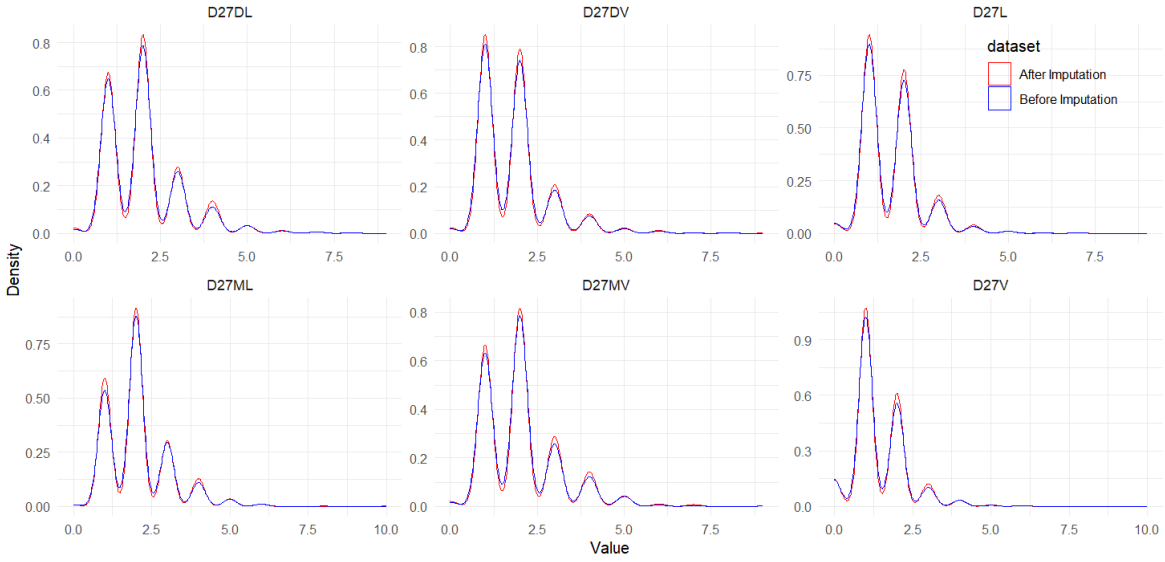
Table II.42: PPD Statistics Comparison: Before and After H-D Imputation by Site

Site	Stats	Tooth 27		Test Results		K-S
		BI	AI	Stats	p	
DV	Median	2	2	W = 3949964	0.508	D = 0.008
	Variance	0.985	1.010	F = 0.155	0.694	p = 1.000
V	Median	1	1	W = 3989478	0.302	D = 0.013
	Variance	0.769	0.764	F = 0.471	0.493	p = 0.975
MV	Median	2	2	W = 3980940	0.369	D = 0.015
	Variance	1.140	1.230	F = 1.853	0.174	p = 0.906
DL	Median	2	2	W = 4029633	0.502	D = 0.009
	Variance	1.160	1.190	F = 0.399	0.528	p = 1.000
L	Median	1	2	W = 4062384	0.315	D = 0.010
	Variance	0.710	0.721	F = 0.066	0.797	p = 0.999
ML	Median	2	2	W = 3975409	0.642	D = 0.011
	Variance	1.030	1.070	F = 0.644	0.422	p = 0.997

Abbreviations: BI – Before Imputation, AI – After Imputation; Stats – Statistics, p – p-value; K-S – Kolmogorov-Smirnov test.

Kernel Density Plot of Imputed and Original

Figure II.14



Kernel Density Plot of Imputed and Original 27 by Site

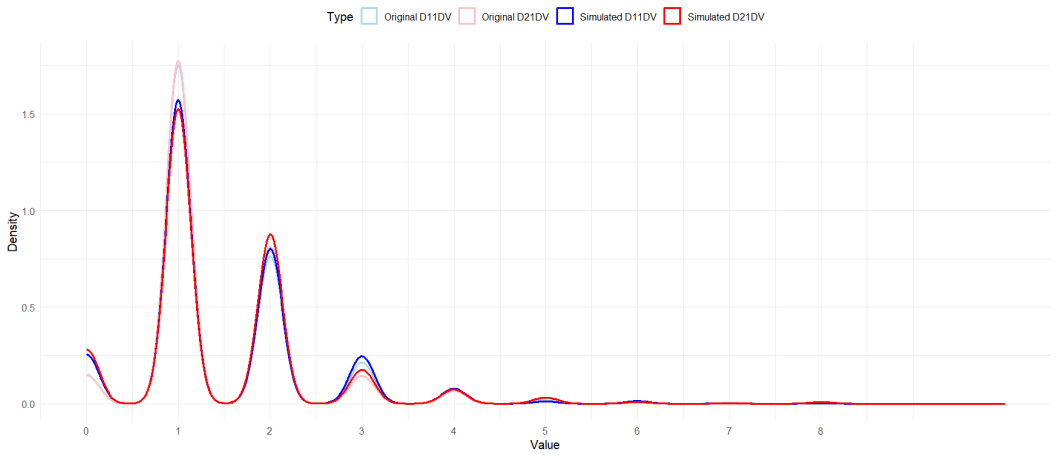
Appendix III

Appendix: Comparison of Original versus Simulated

III.1 Upper Central Incisors - Original *vs* Simulated

Site DV

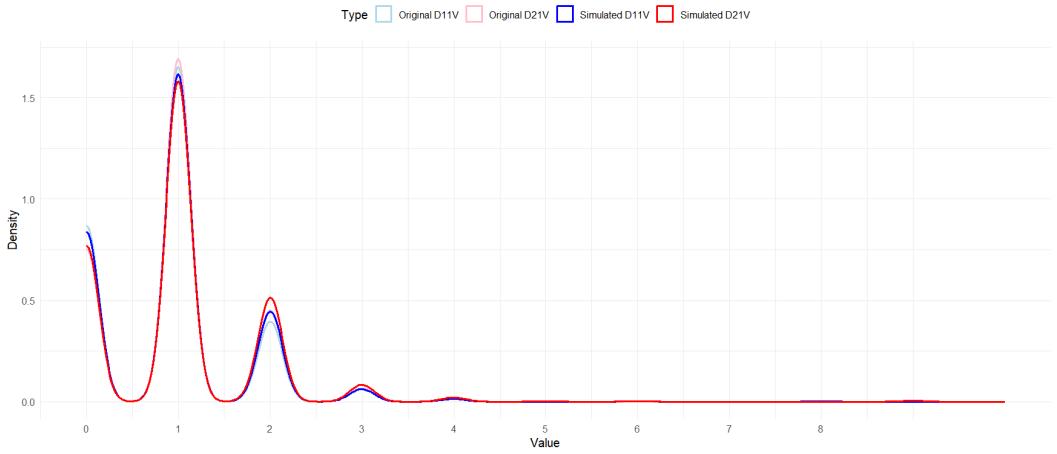
Figure III.1



Kernel Density Plot of Simulated and Original 11DV, 21DV (noise level: 0.4)

Site V

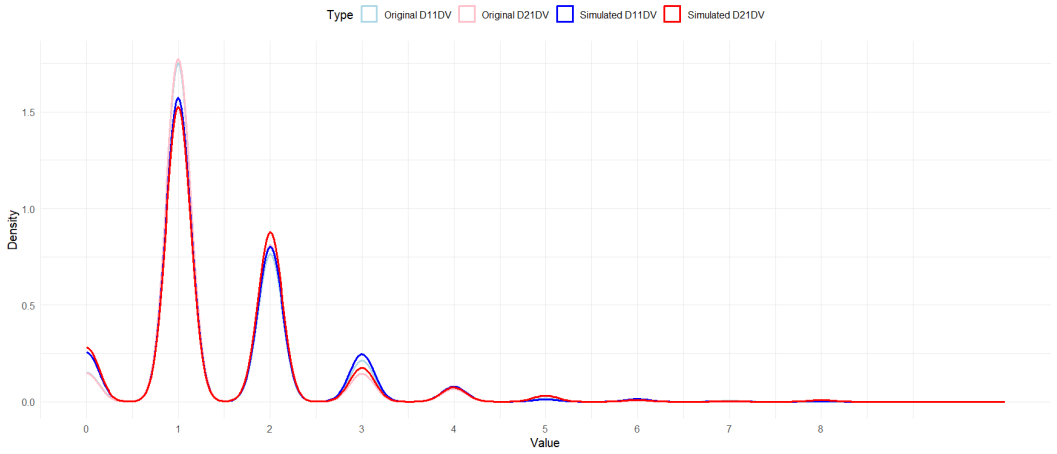
Figure III.2



Kernel Density Plot of Simulated 11V, 21V (noise level: 0.4) and Original 11V, 21V

Site MV

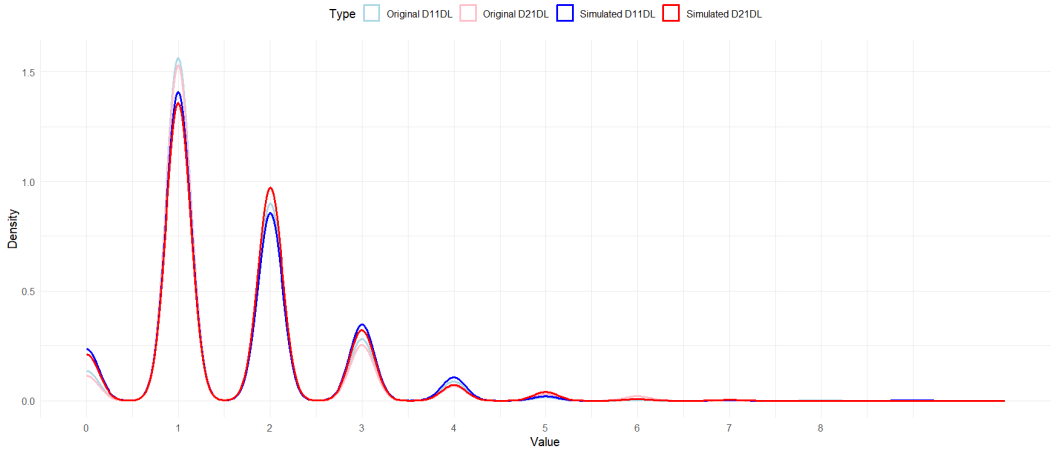
Figure III.3



Kernel Density Plot of Simulated and Original 11MV, 21MV (noise level = 0.4)

Site DL

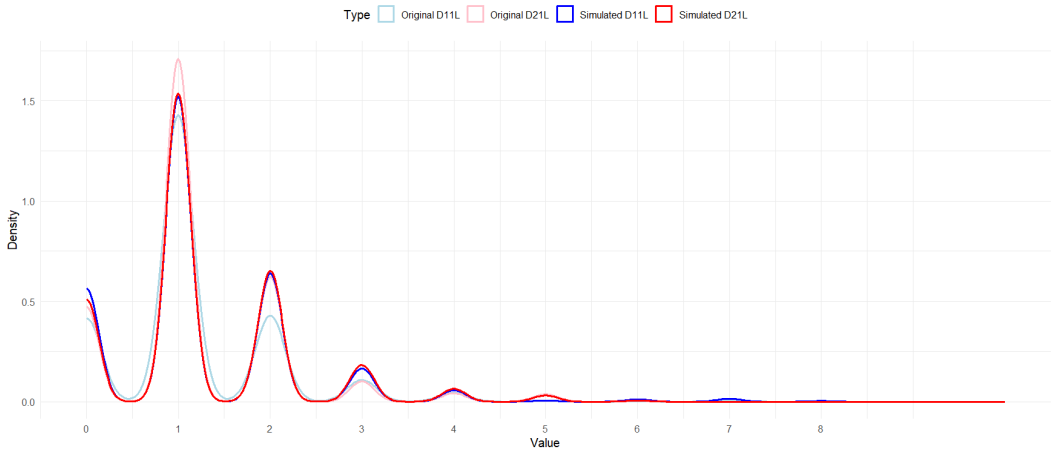
Figure III.4



Kernel Density Plot of Simulated and Original 11DL, 21DL (Noise = 0.4)

Site L

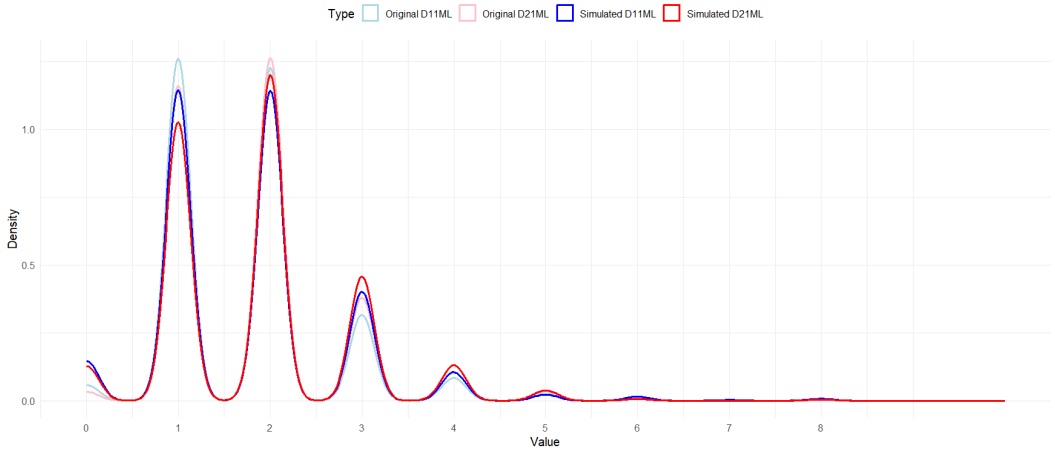
Figure III.5



Kernel Density Plot of Simulated and Original 11L, 21L (noise level = 0.4)

Site ML

Figure III.6

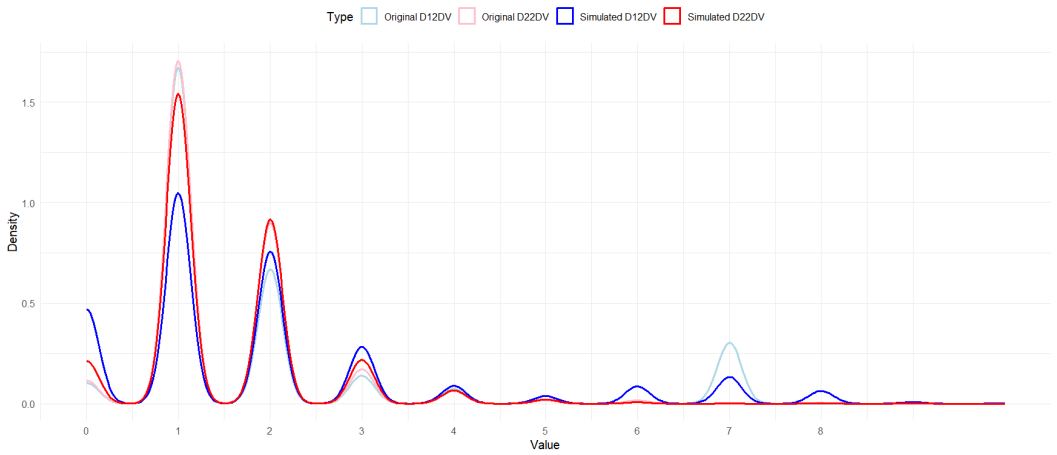


Kernel Density Plot of Simulated and Original 11ML, 21ML (Noise: 0.4)

III.2 Upper Lateral Incisors - Original *vs* Simulated

Site DV

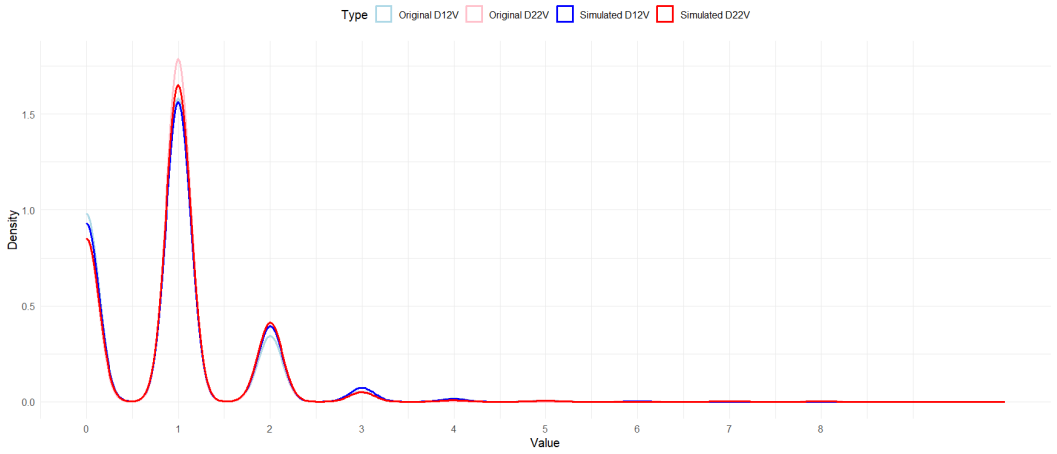
Figure III.7



Kernel Density Plot of Simulated and Original 12DV, 22DV (Noise: 0.4)

Site V

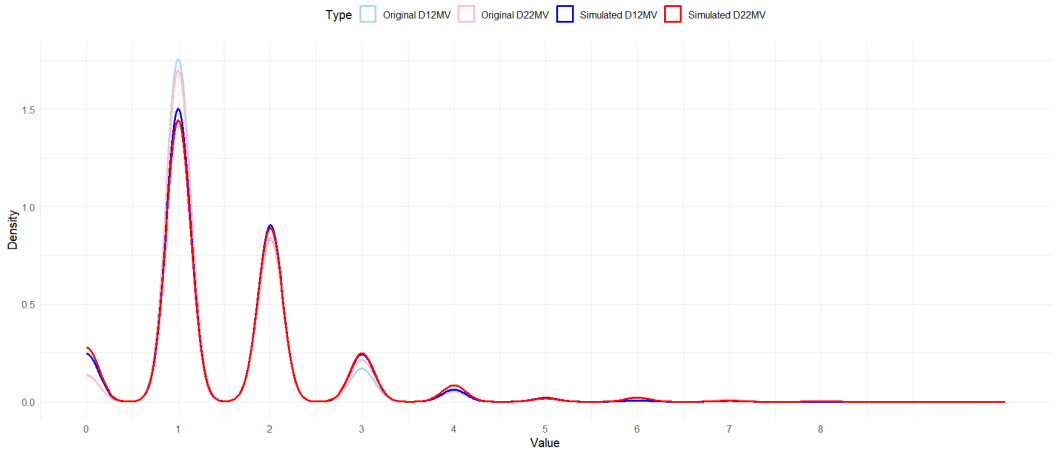
Figure III.8



Kernel Density Plot of Simulated and Original 12V, 22V (Noise: 0.4)

Site MV

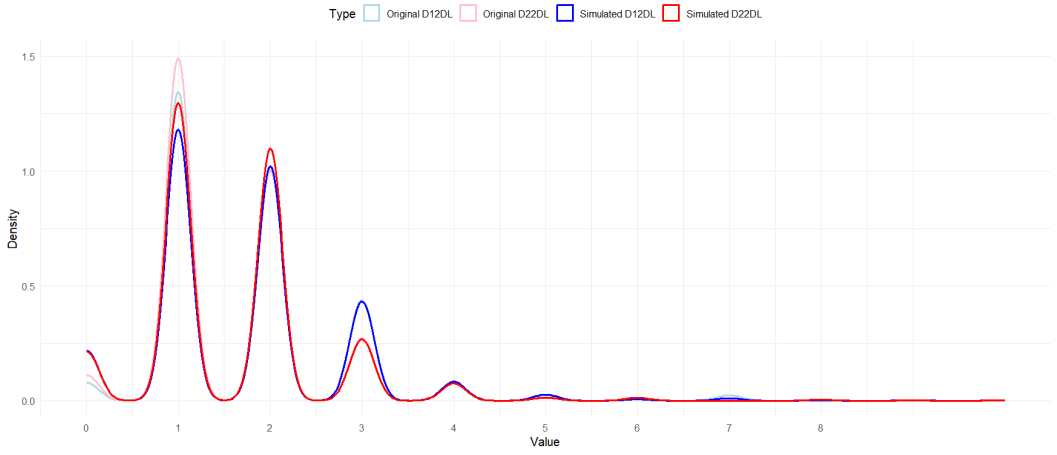
Figure III.9



Kernel Density Plot of Simulated and Original 12MV, 22MV (noise level = 0.4)

Site DL

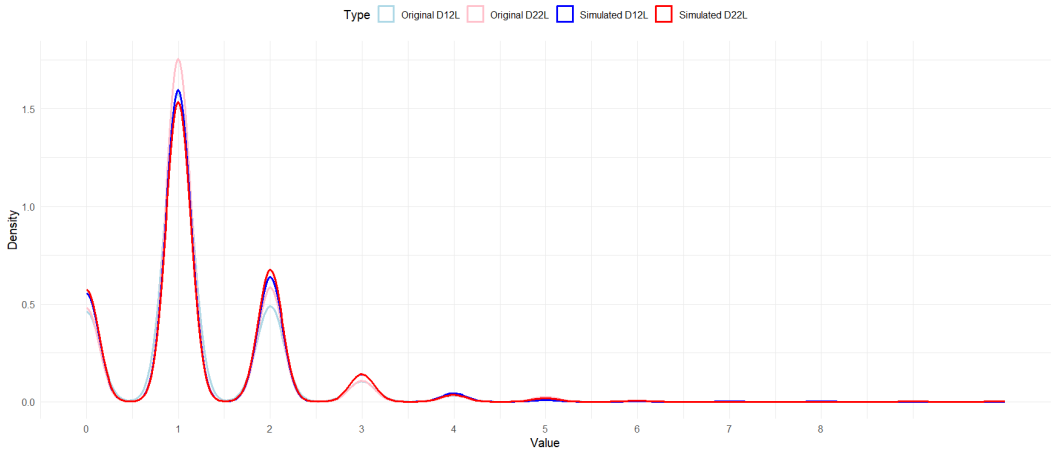
Figure III.10



Kernel Density Plot of Simulated and Original 12DL, 22DL (Noise = 0.4)

Site L

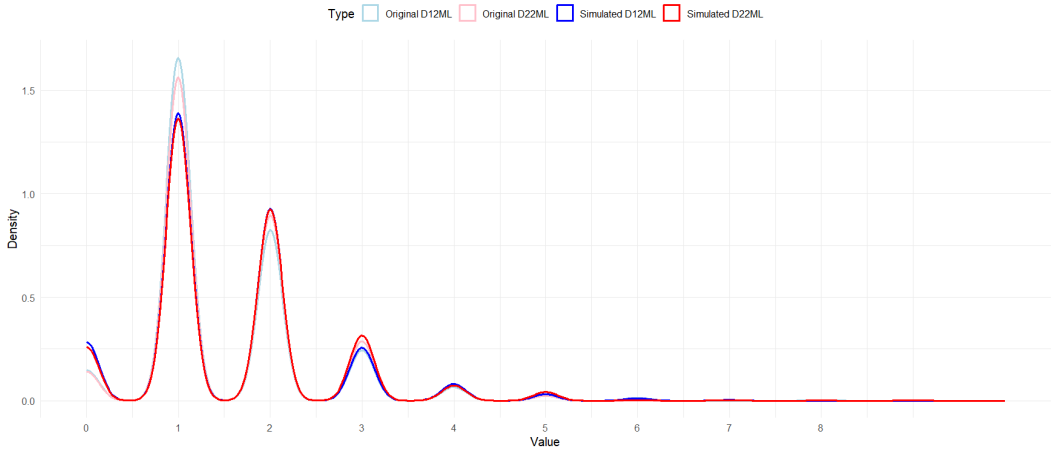
Figure III.11



Kernel Density Plot of Simulated and Original 12L, 22L (noise level = 0.4)

Site ML

Figure III.12

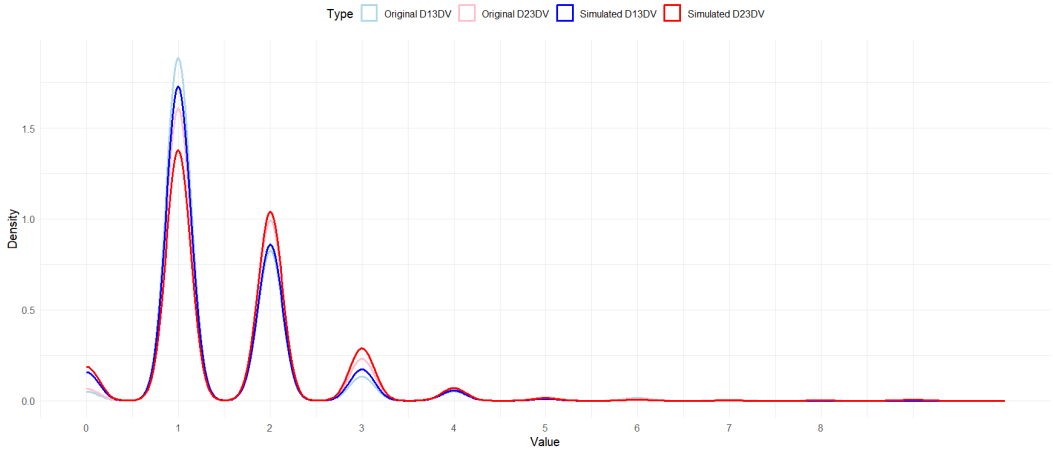


Kernel Density Plot of Simulated and Original 12ML, 22ML (Noise: 0.4)

III.3 Upper Canines - Original *vs* Simulated

Site DV

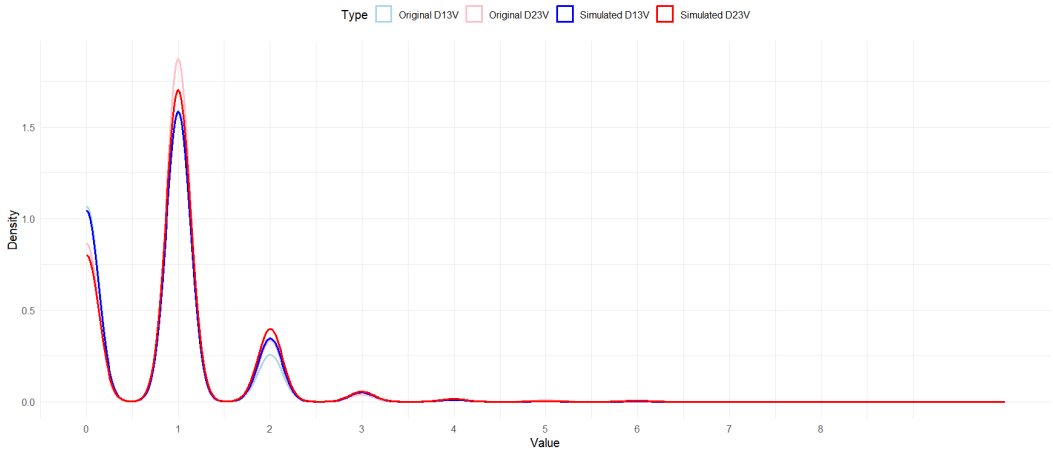
Figure III.13



Kernel Density Plot of Simulated and Original 13DV, 23DV (Noise: 0.4)

Site V

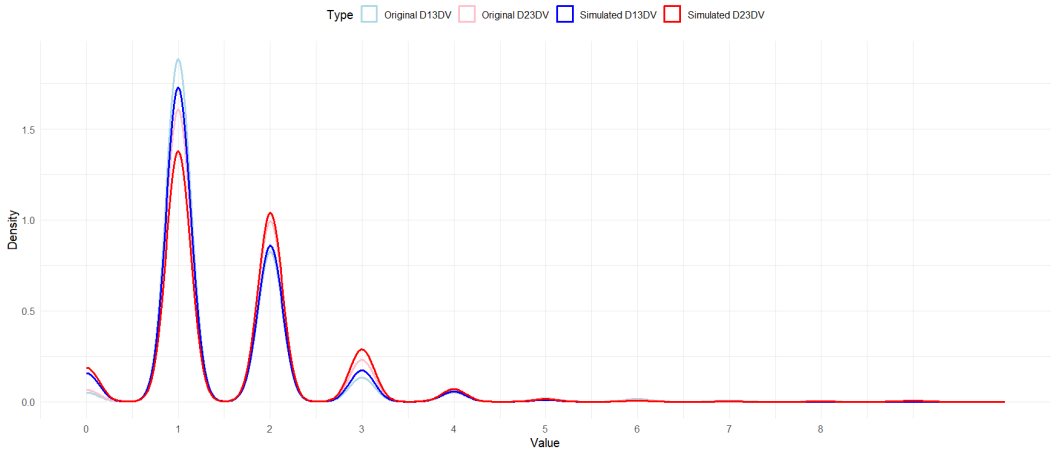
Figure III.14



Kernel Density Plot of Simulated and Original 13V, 23V (Noise: 0.4)

Site MV

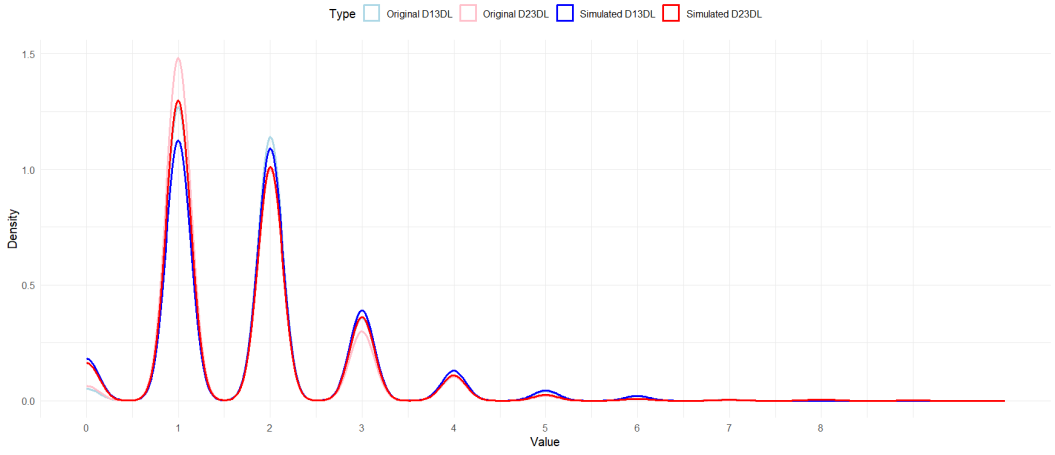
Figure III.15



Kernel Density Plot of Simulated and Original 13MV, 23MV (noise level = 0.4)

Site DL

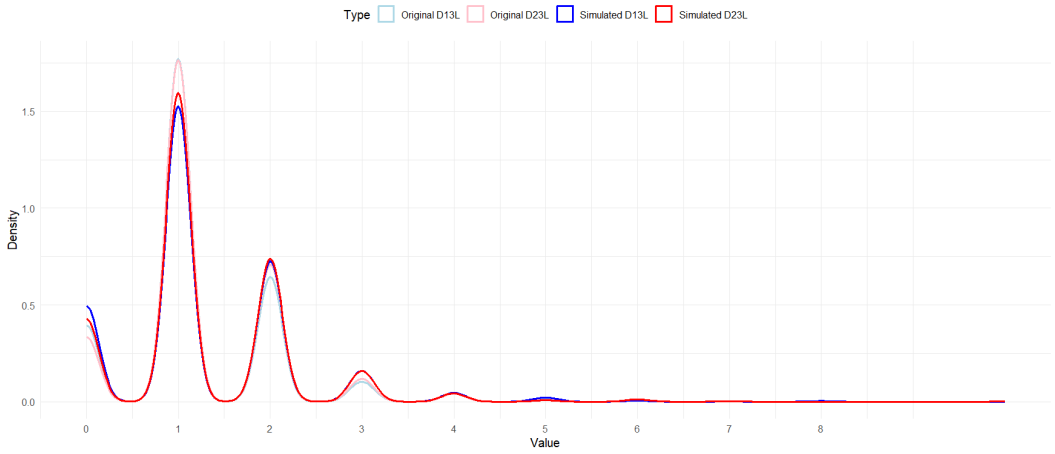
Figure III.16



Kernel Density Plot of Simulated and Original 13DL, 23DL (Noise = 0.4)

Site L

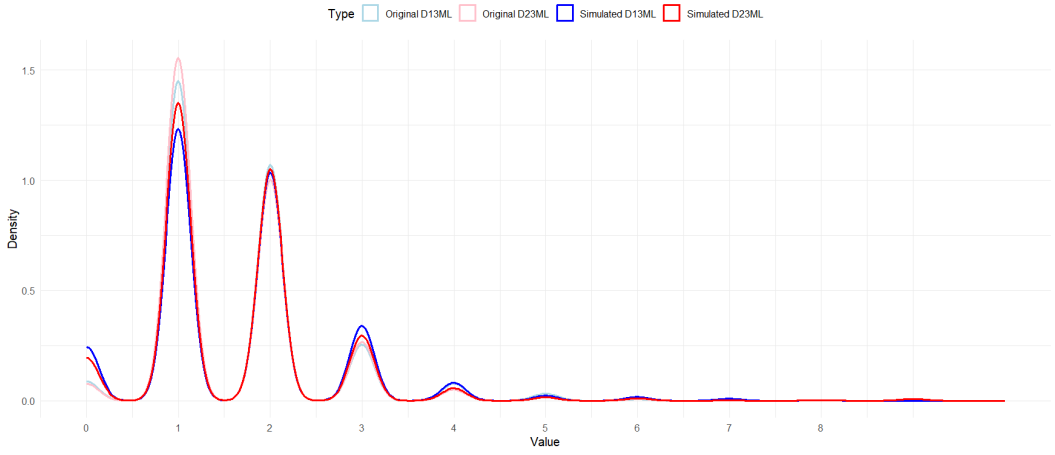
Figure III.17



Kernel Density Plot of Simulated and Original 13L, 23L (noise level = 0.4)

Site ML

Figure III.18

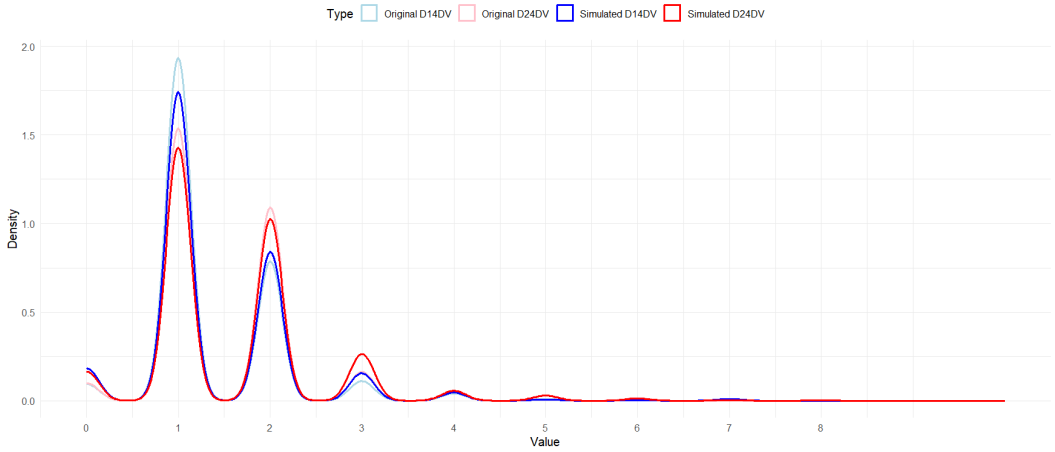


Kernel Density Plot of Simulated and Original 13ML, 23ML (Noise: 0.4)

III.4 Upper First Premolar - Original *vs* Simulated

Site DV

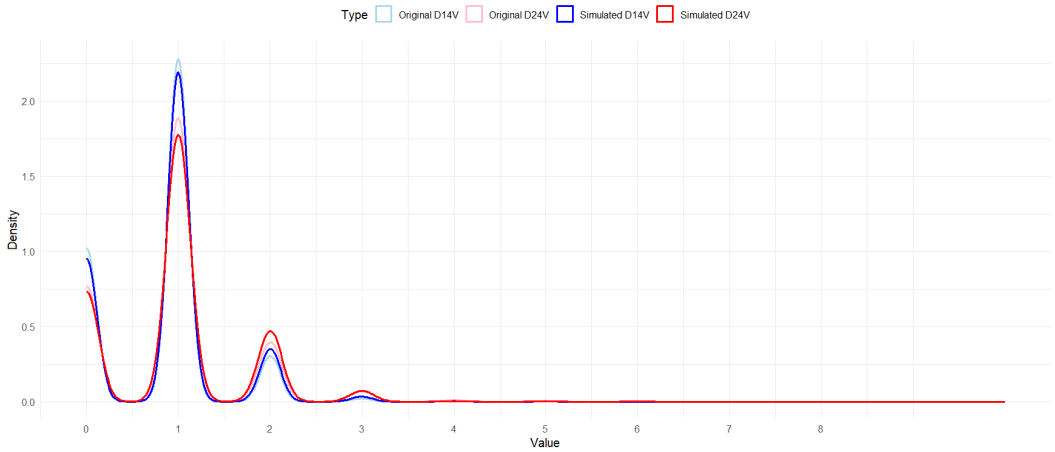
Figure III.19



Kernel Density Plot of Simulated 14DV, 24DV (noise level: 0.4)

Site V

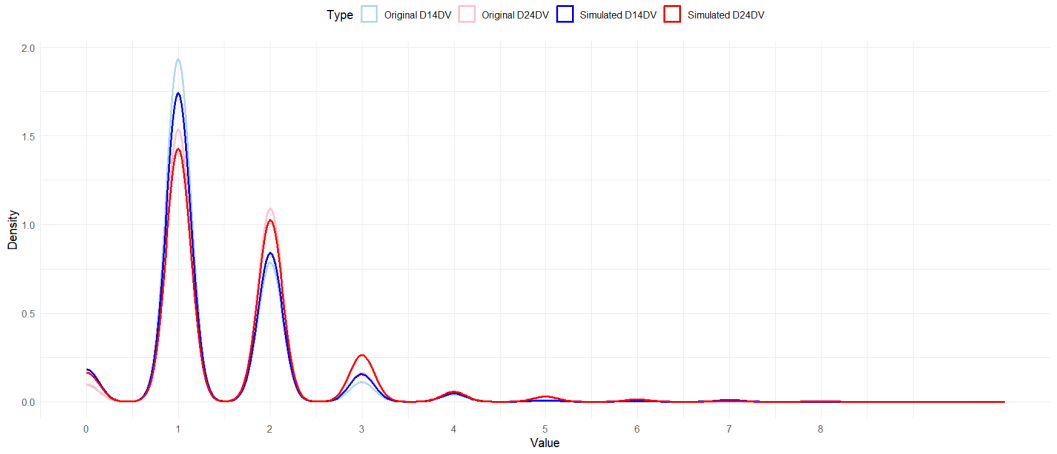
Figure III.20



Kernel Density Plot of Simulated and Original 14V, 24V (Noise: 0.4)

Site MV

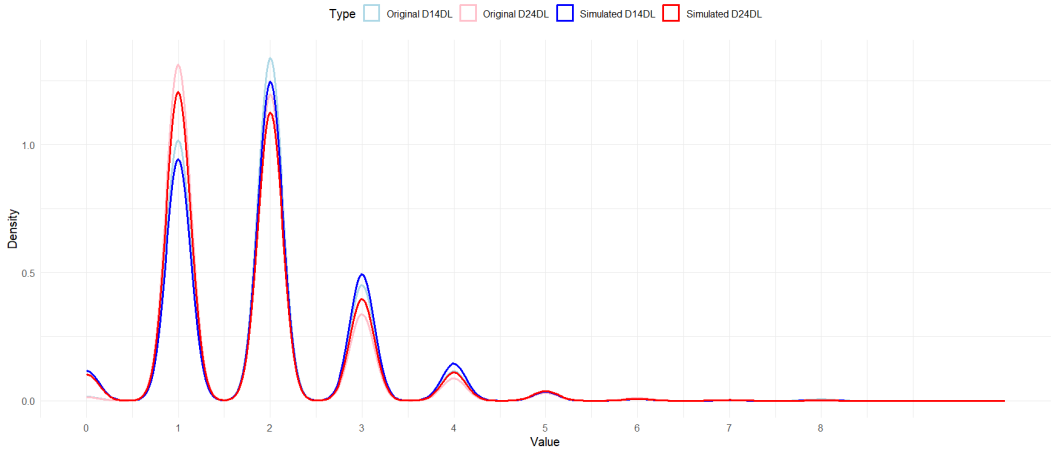
Figure III.21



Kernel Density Plot of Simulated and Original 14MV, 24MV (noise level = 0.4)

Site DL

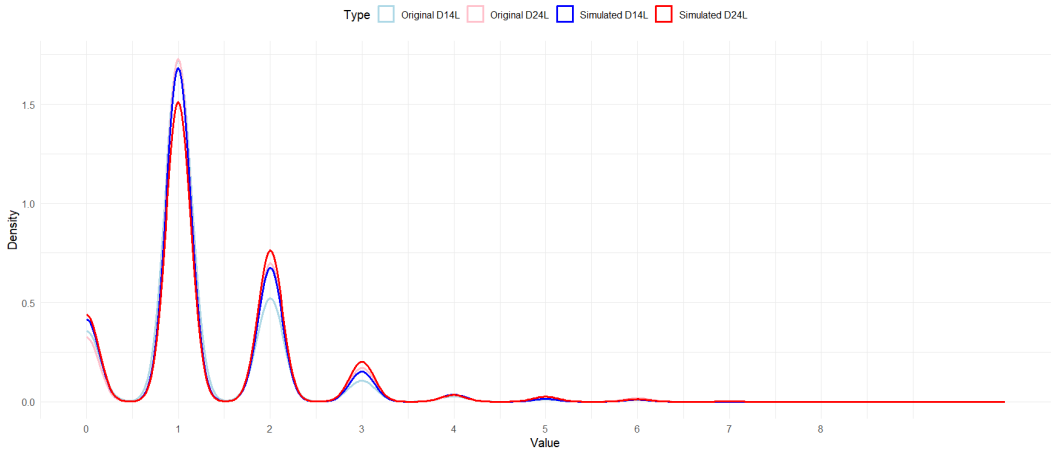
Figure III.22



Kernel Density Plot of Simulated and Original 14DL, 24DL (Noise = 0.4)

Site L

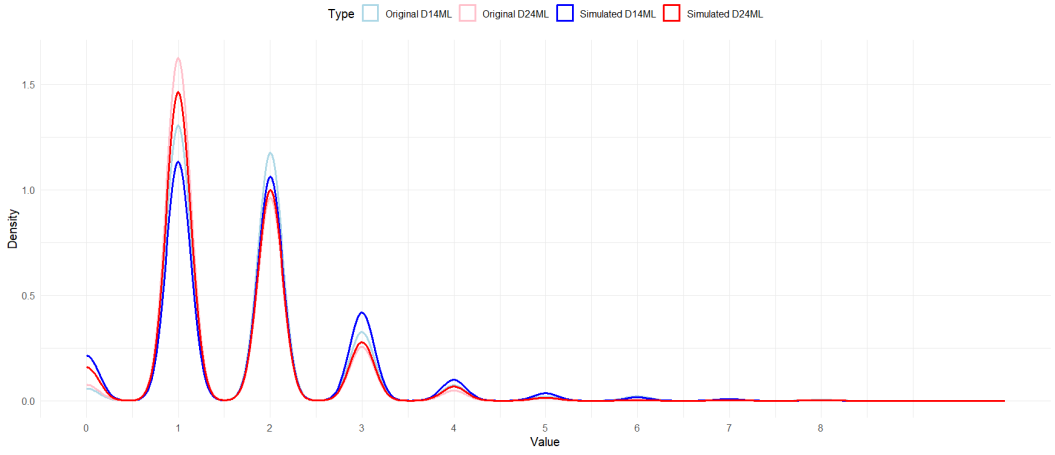
Figure III.23



Kernel Density Plot of Simulated and Original 14L, 24L (noise level = 0.4)

Site ML

Figure III.24

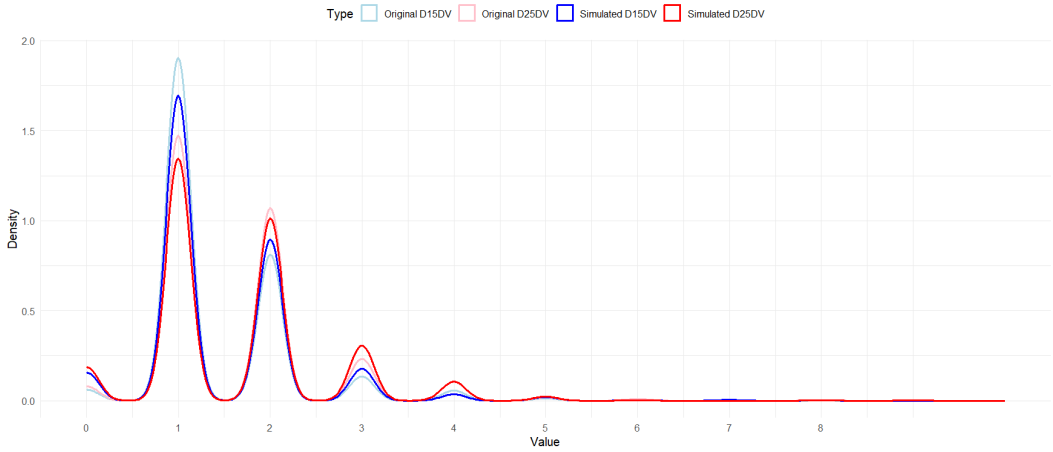


Kernel Density Plot of Simulated and Original 14ML, 24ML (Noise: 0.4)

III.5 Upper Second Premolars - Original *vs* Simulated

Site DV

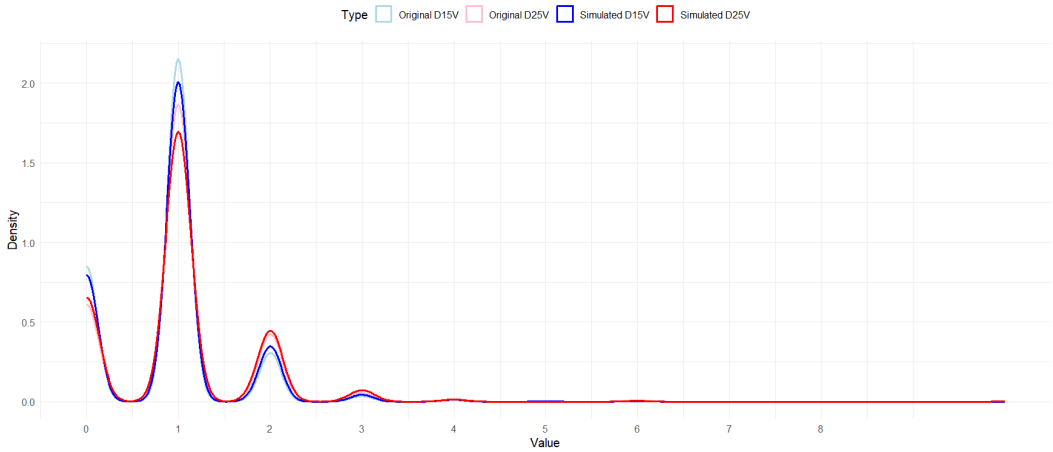
Figure III.25



Kernel Density Plot of Simulated and Original 15DV, 25DV (Noise: 0.4)

Site V

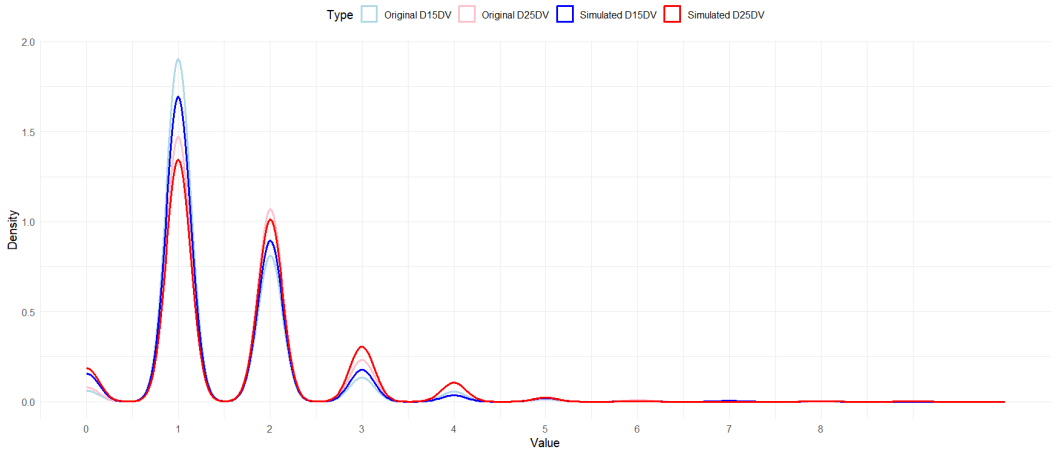
Figure III.26



Kernel Density Plot of Simulated and Original 15V, 25V (Noise: 0.4)

Site MV

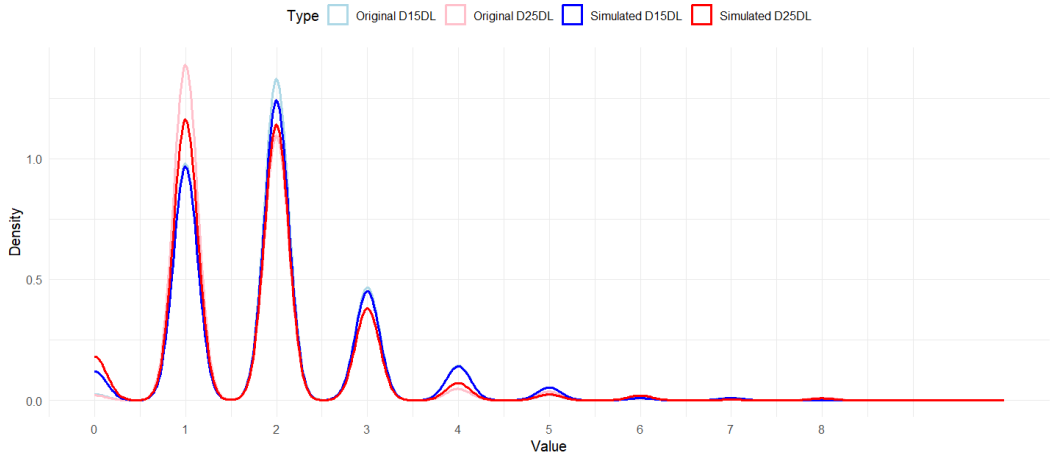
Figure III.27



Kernel Density Plot of Simulated and Original 15MV, 25MV (noise level = 0.4)

Site DL

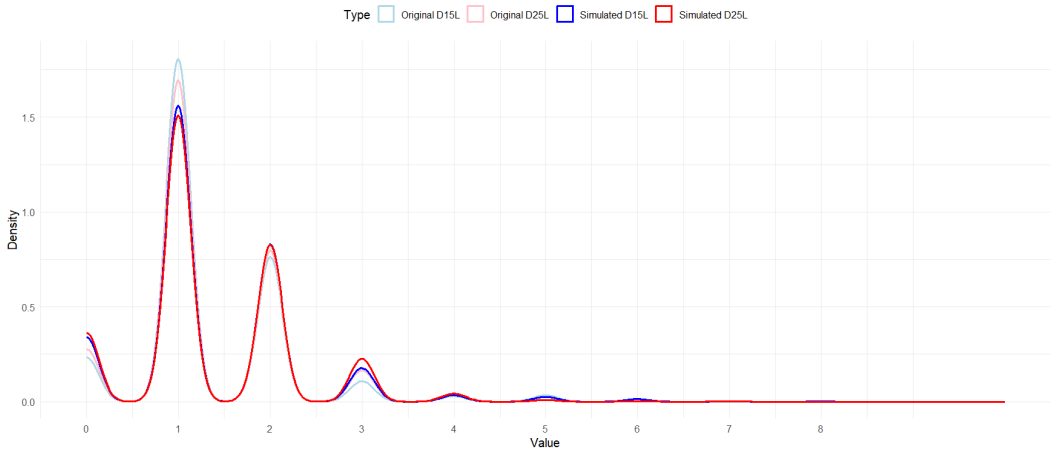
Figure III.28



Kernel Density Plot of Simulated and Original 15DL, 25DL (Noise = 0.4)

Site L

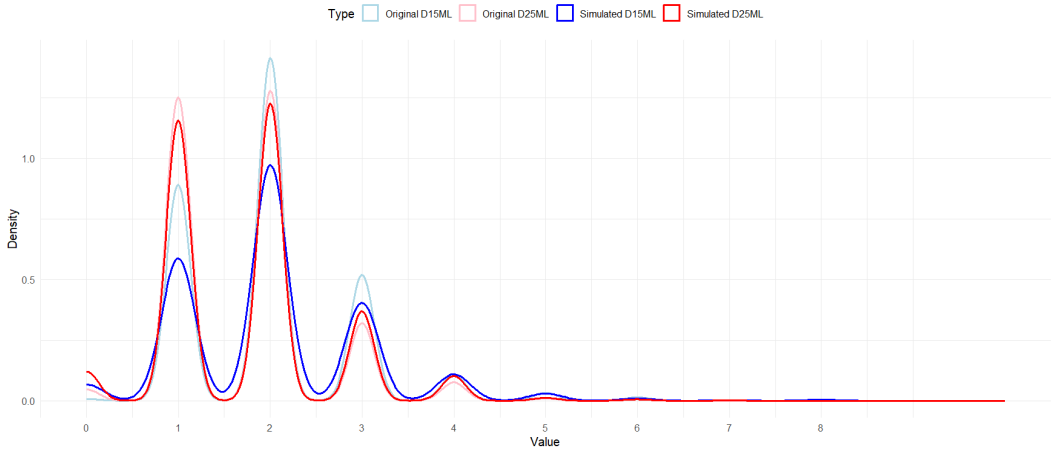
Figure III.29



Kernel Density Plot of Simulated and Original 15L, 25L (noise level = 0.4)

Site ML

Figure III.30

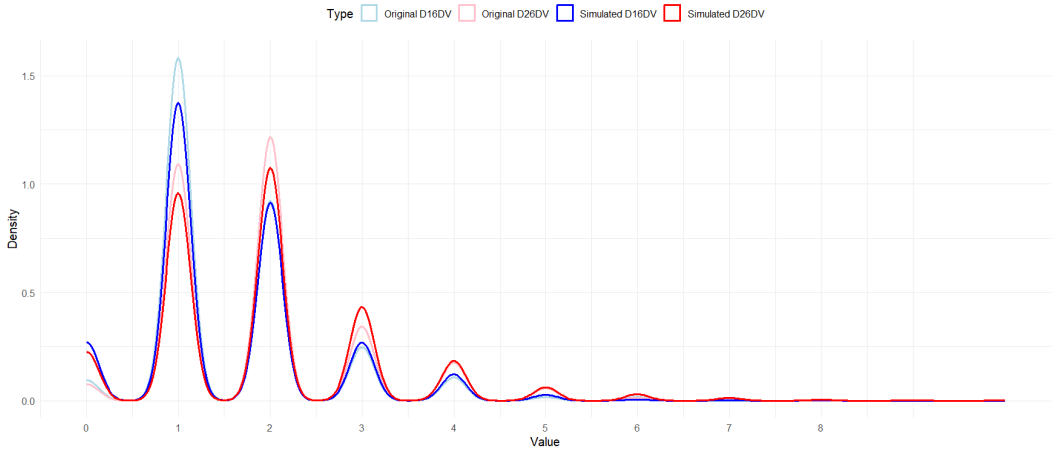


Kernel Density Plot of Simulated and Original 15ML, 25ML (Noise: 0.4)

III.6 Upper First Molars - Original *vs* Simulated

Site DV

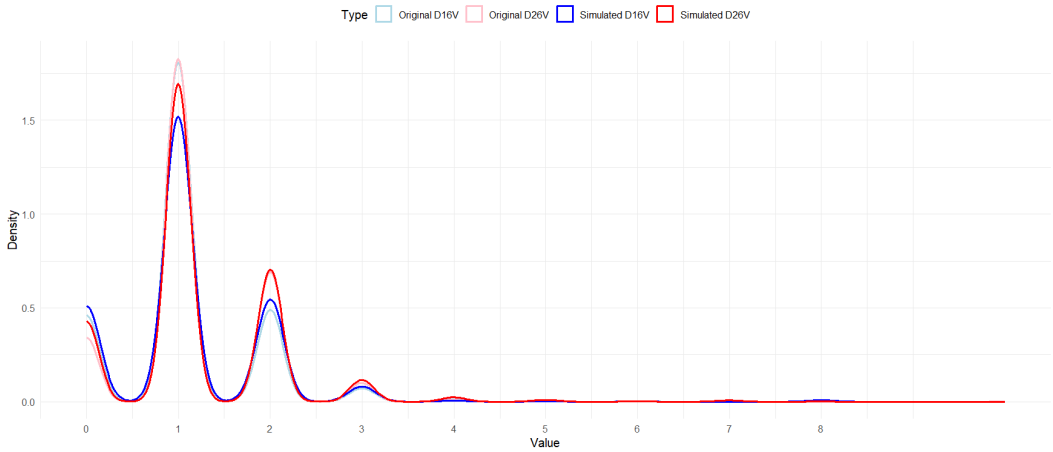
Figure III.31



Kernel Density Plot of Simulated and Original 16DV, 26DV (Noise: 0.4)

Site V

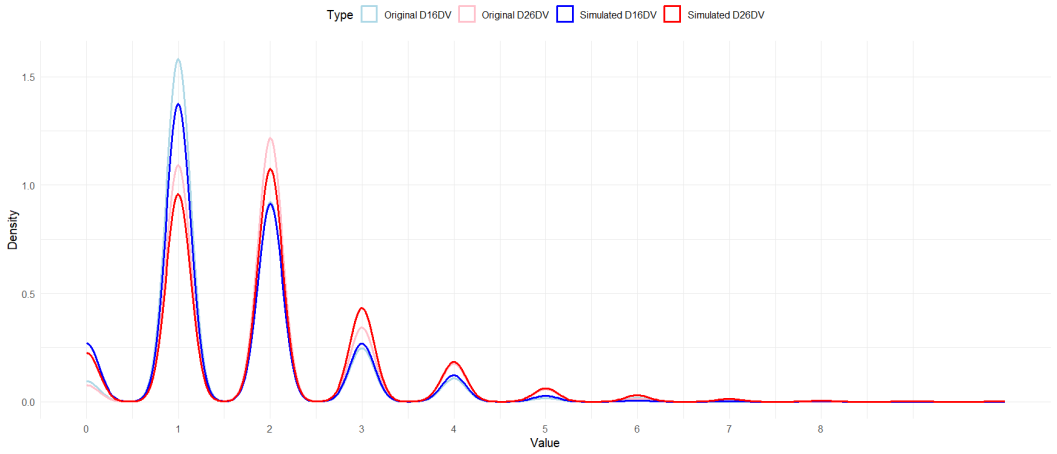
Figure III.32



Kernel Density Plot of Simulated and Original 16V, 26V (Noise: 0.4)

Site MV

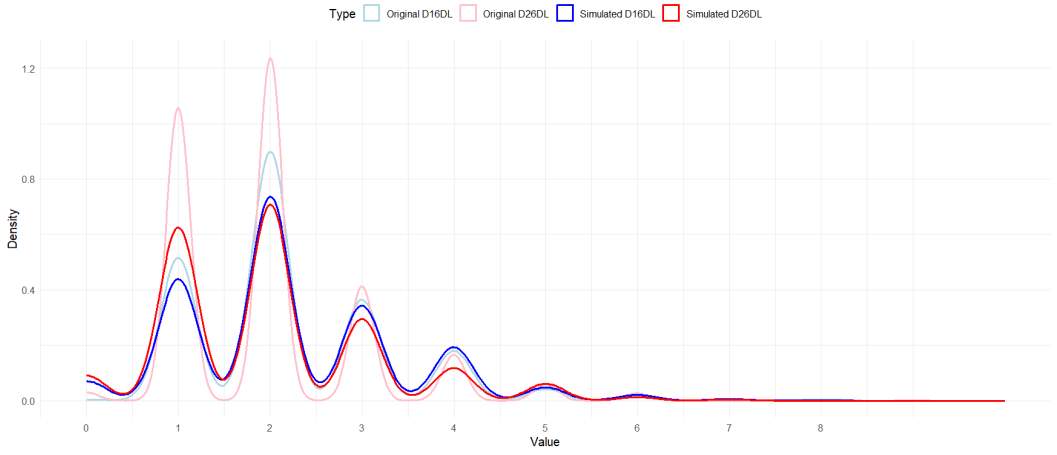
Figure III.33



Kernel Density Plot of Simulated and Original 16MV, 26MV (noise level = 0.4)

Site DL

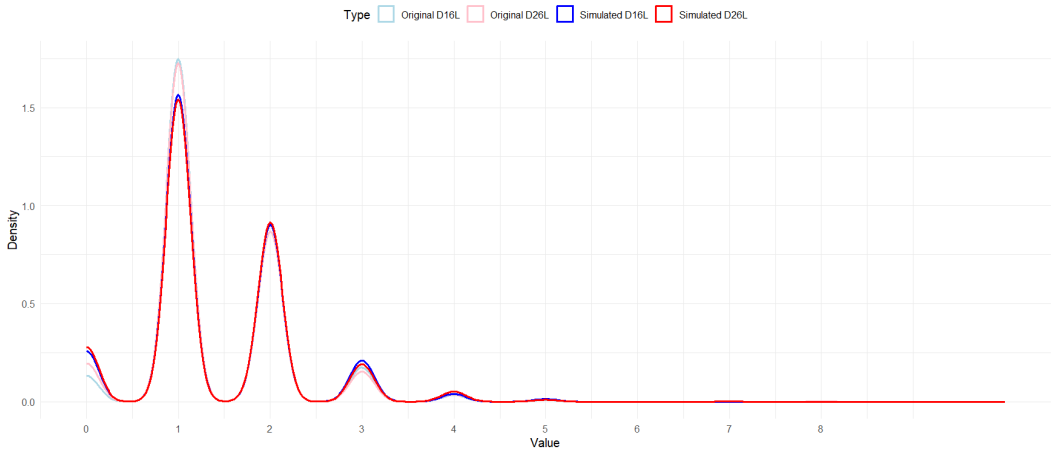
Figure III.34



Kernel Density Plot of Simulated and Original 16DL, 26DL (Noise = 0.4)

Site L

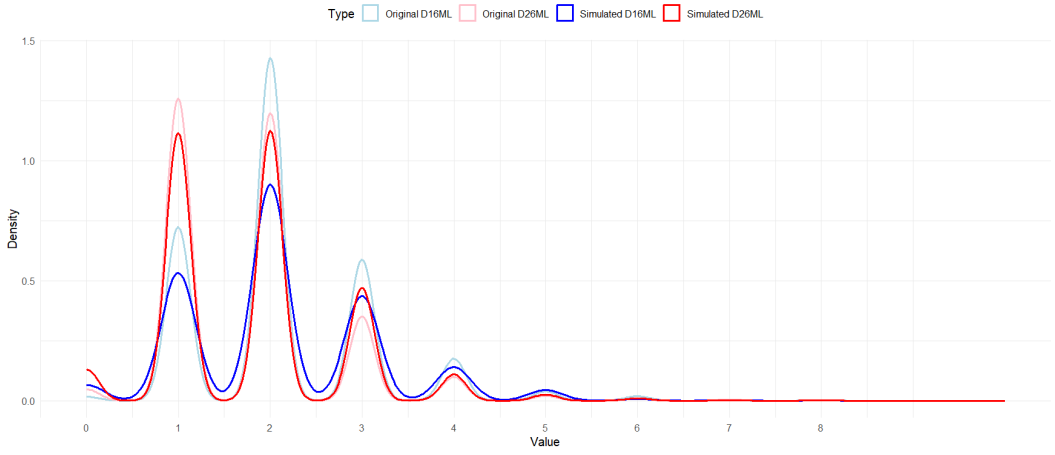
Figure III.35



Kernel Density Plot of Simulated and Original 16L, 26L (noise level = 0.4)

Site ML

Figure III.36

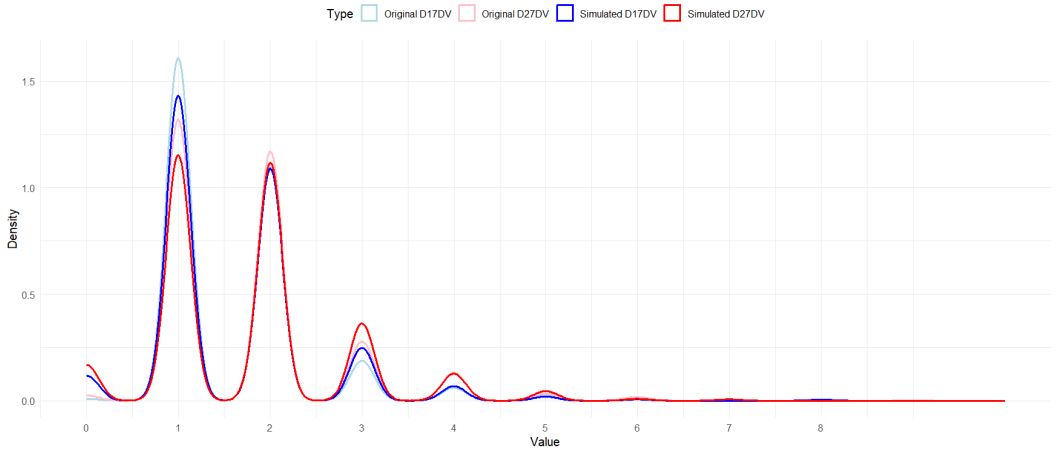


Kernel Density Plot of Simulated and Original 16ML, 26ML (Noise: 0.4)

III.7 Upper Second Molars - Original *vs* Simulated

Site DV

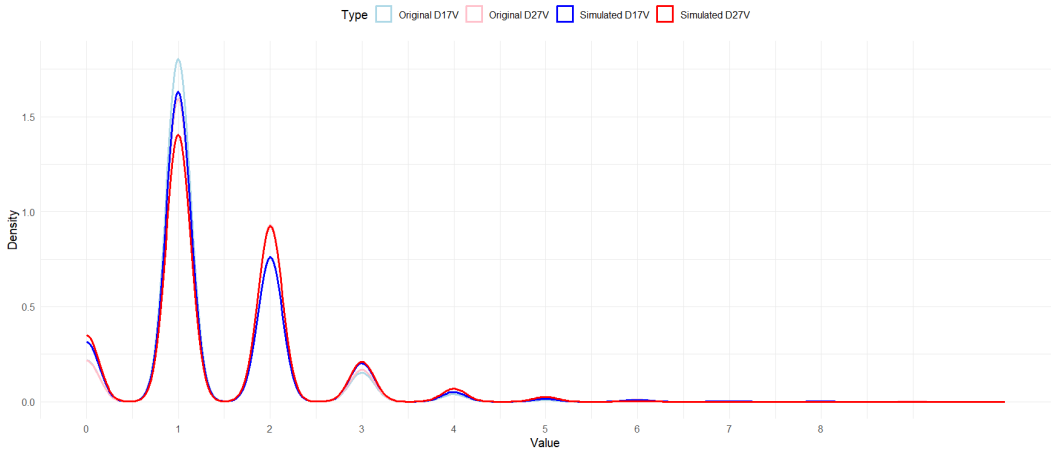
Figure III.37



Kernel Density Plot of Simulated and Original 17DV, 27DV (Noise: 0.4)

Site V

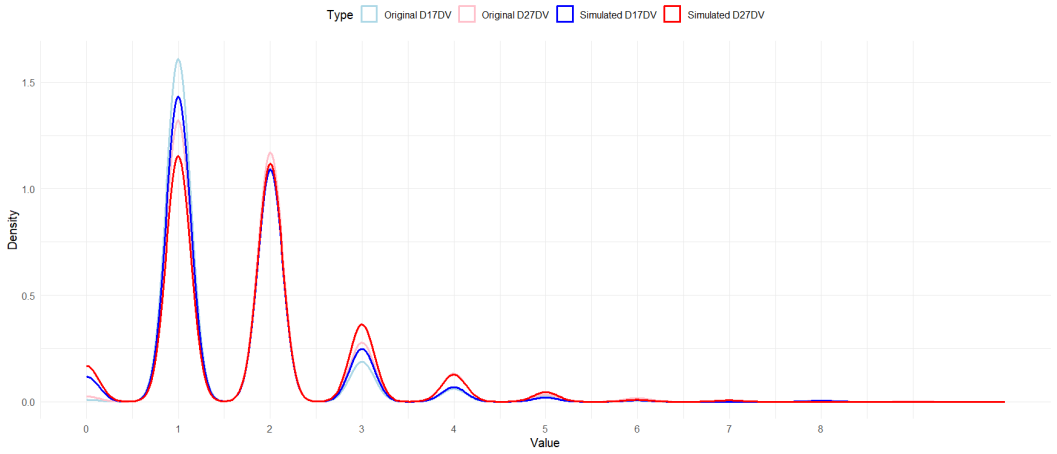
Figure III.38



Kernel Density Plot of Simulated and Original 17V, 27V (Noise: 0.4)

Site MV

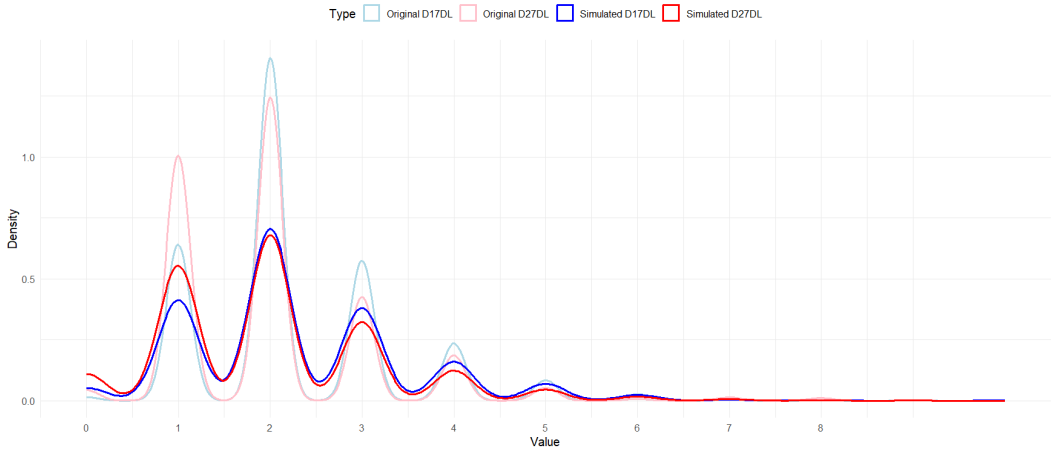
Figure III.39



Kernel Density Plot of Simulated and Original 17MV, 27MV (noise level = 0.4)

Site DL

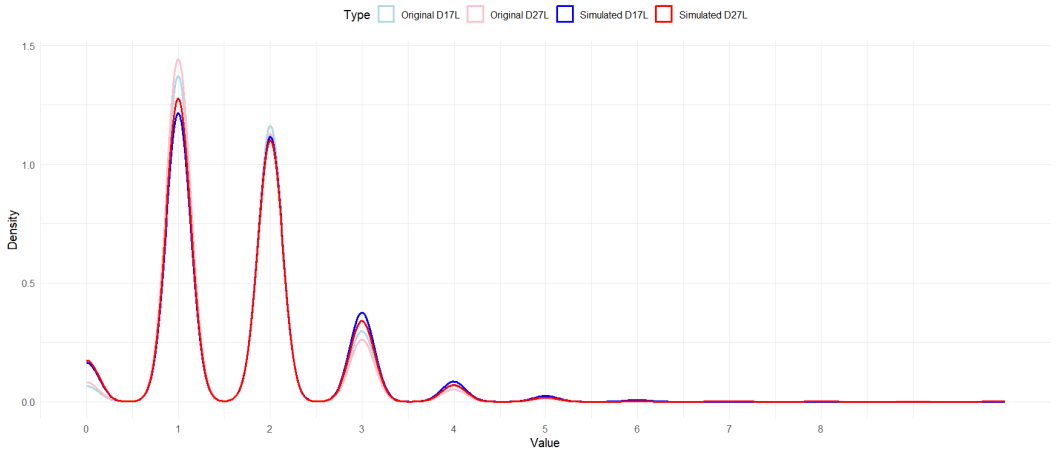
Figure III.40



Kernel Density Plot of Simulated and Original 17DL, 27DL (Noise = 0.4)

Site L

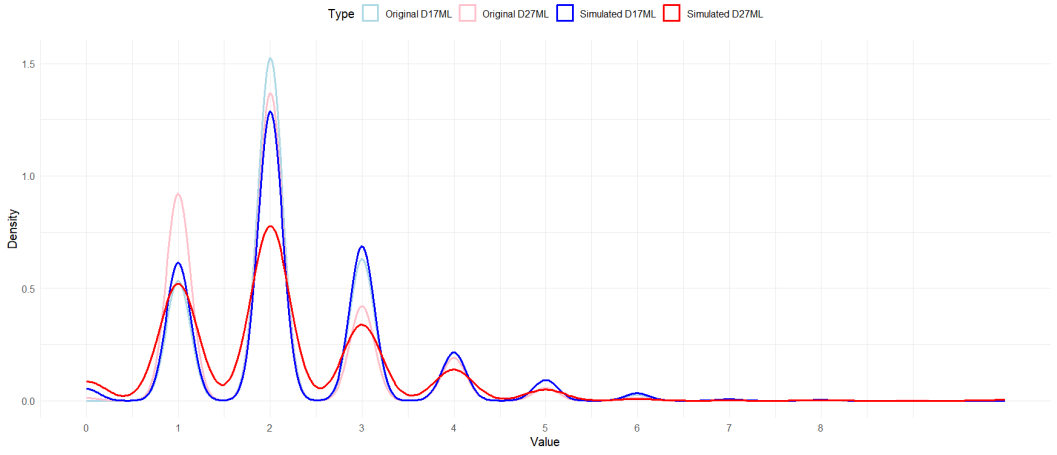
Figure III.41



Kernel Density Plot of Simulated and Original 17L, 27L (noise level = 0.4)

Site ML

Figure III.42



Kernel Density Plot of Simulated and Original 17ML, 27ML (Noise: 0.4)

Appendix IV

Appendix: Mother-Daughter Method Imputation Results

IV.1 Upper Left Central Incisor

Analysis of M21 Mother Models: Characteristics, Performance, and Feature Importance

Table IV.1: Distinctive Characteristics of M21 Mother Models by Site

Metric	M21DV	M21V	M21MV	M21DL	M21L	M21ML
Mother Model Size (Kb)	274.80	146.00	211.80	261.00	366.00	258.80
N.Iter.	261	134	198	253	344	246
Init. Train. RMSE	1.193	0.868	1.152	1.305	1.062	1.476
Final Train. RMSE	1.82e-02	5.94e-02	3.39e-02	2.92e-02	1.60e-02	3.67e-02
Features	γSM.DV 11DV	γSM.V 11V	γSM.MV 11MV	γSM.DL 11DL	γSM.L 11L	γSM.ML 11ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table IV.2: Performance Metrics for the M21 Mother Models by Site

Metric	M21DV	M21V	M21MV	M21DL	M21L	M21ML
RMSE	0.018	0.059	0.034	0.029	0.016	0.037
MAE	0.003	0.007	0.004	0.005	0.002	0.005
MSE	3.30e-04	3.52e-03	1.15e-03	8.50e-04	2.60e-04	1.35e-03
R^2	99.96%	99.48%	99.84%	99.90%	99.97%	99.84%

Table IV.3: Mother Models M21 Features Importance Metrics by Site

Model	Feature	Gain	Cover	Frequency
M21DV	γSM.DV	0.612	0.509	0.615
	Original 11DV	0.388	0.491	0.385
M21V	γSM.V	0.599	0.538	0.651
	Original 11V	0.401	0.462	0.349
M21MV	γSM.MV	0.558	0.474	0.639
	Original 11MV	0.442	0.526	0.361
M21DL	γSM.DL	0.577	0.485	0.592
	Original 11DL	0.423	0.515	0.408
M21L	γSM.L	0.558	0.612	0.690
	Original 11L	0.442	0.388	0.310
M21ML	γSM.ML	0.525	0.512	0.638
	Original 11ML	0.475	0.488	0.362

Analysis of D21 Daughter Models: Characteristics, Performance, and Feature Importance

The daughters models were fitted to the predictors: upper right side PPD values from the simulated data, the soft labels generated by the Mother Models and to five predictions of the New Mother Models

Table IV.4: Distinctive Characteristics of the D21 by Site

Metric	D21DV	D21V	D21MV	D21DL	D21L	21ML
Model Size (Mb)	31.2	32.1	31.2	32.4	32.1	32.5
N.Iter.	30000	30000	30000	30000	30000	30000
Init. Train. RMSE	1.305	0.953	1.287	1.430	1.180	1.590
Final Train. RMSE	0.181	0.142	0.177	0.162	0.161	0.156

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Table IV.5: Performance Metrics by Site for D21

Metric	D21DV	D21V	D21MV	D21DL	D21L	D21ML
RMSE	0.181	0.142	0.177	0.162	0.161	0.156
MAE	0.143	0.112	0.140	0.128	0.128	0.123
MSE	0.033	0.020	0.031	0.026	0.026	0.024
R^2	96.63%	97.44%	96.58%	97.47%	97.37%	97.49%

Table IV.6: Feature Importance Metrics of D21 by Site

Model	Feature	Gain	Cover	Frequency
D21DV	Mother Predictions	0.389	0.273	0.324
	New Mother Predictions	0.332	0.127	0.094
	Simulated 11DV	0.279	0.601	0.582
D21V	Mother Predictions	0.442	0.301	0.300
	New Mother Predictions	0.274	0.088	0.072
	Simulated 11V	0.283	0.611	0.628
D21MV	Mother Predictions	0.380	0.304	0.346
	New Mother Predictions	0.335	0.088	0.066
	Simulated 11MV	0.285	0.608	0.587
D21DL	Mother Predictions	0.381	0.290	0.309
	New Mother Predictions	0.344	0.115	0.088
	Simulated 11DL	0.275	0.596	0.603
D21L	Mother Predictions	0.432	0.261	0.286
	New Mother Predictions	0.290	0.135	0.100
	Simulated 11L	0.278	0.604	0.614
D21ML	Mother Predictions	0.442	0.302	0.322
	New Mother Predictions	0.280	0.093	0.073
	Simulated 11ML	0.278	0.605	0.605

Daughter Model (D21) Predictions

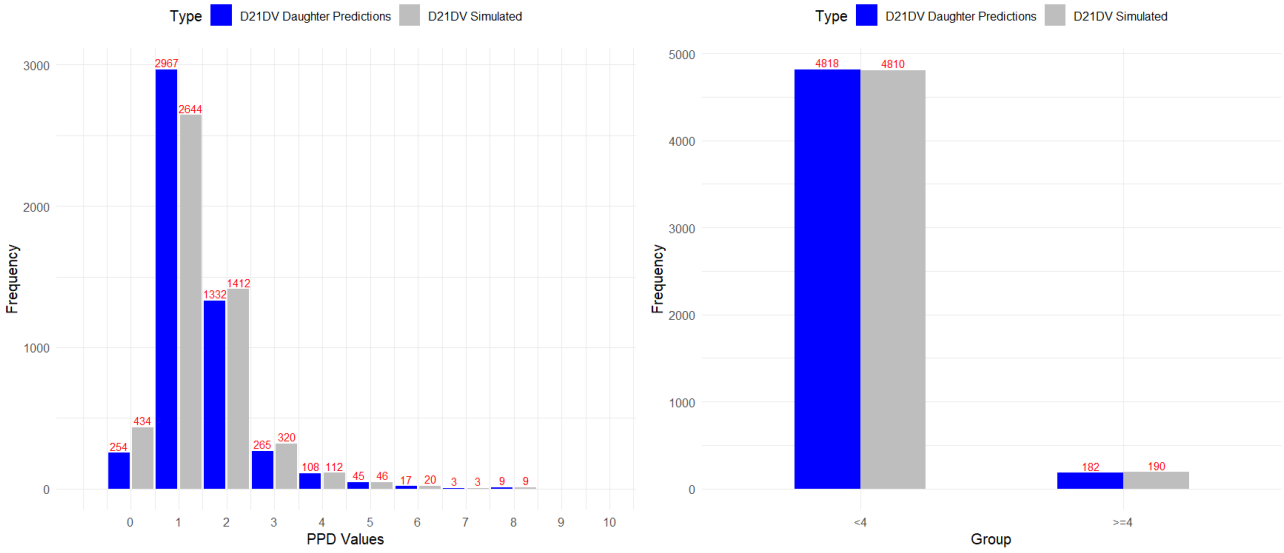
Comparisons of Proportions of PPD Original *vs* D21 Predicted by Site

Site DV

Table IV.7: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
21DV	0	254	434	47.093	**6.77e-12**
	1	2967	2644	18.594	**1.62e-05**
	2	1332	1412	2.332	0.127
	3	265	320	5.171	0.023
	4	108	112	0.073	0.787
	5	45	46	0.011	0.917
	6	17	20	0.243	0.622
	7	3	3	0	1
PPD $\geq$ 4		182	190	0.137	0.711

Figure IV.1



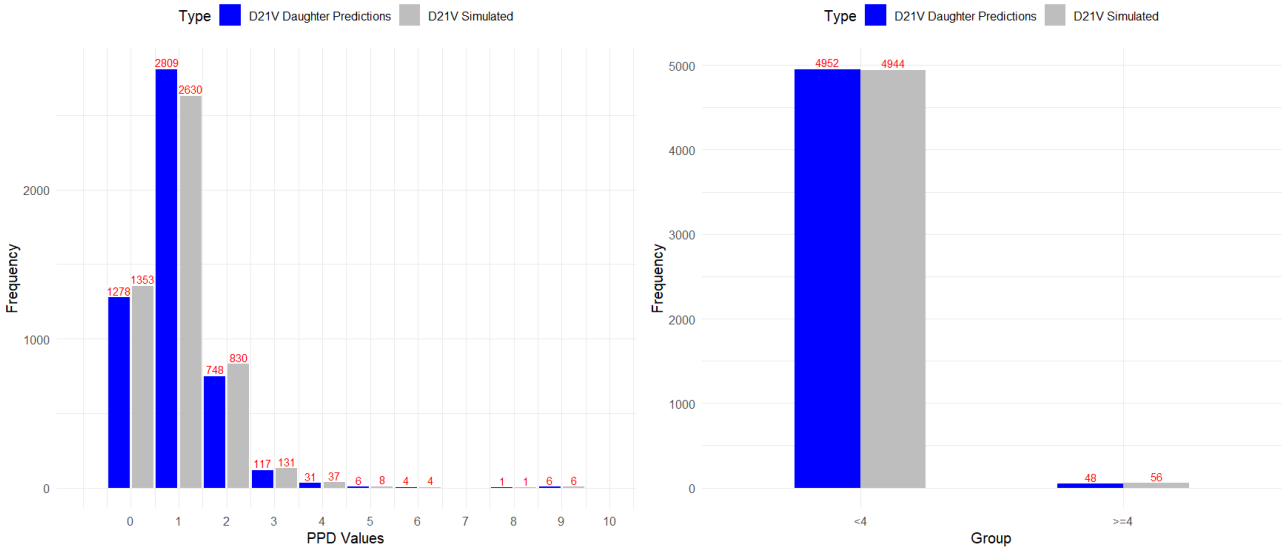
Histograms of Simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

Site V

Table IV.8: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
	0	1278	1353	2.138	0.144
	1	2809	2630	5.891	**0.015**
	2	748	830	4.261	**0.039**
	3	117	131	0.790	0.374
21V	4	31	37	0.529	0.467
	5	6	8	0.286	0.593
	6	4	4	0	1
	8	1	1	0	1
	9	6	6	0	1
PPD ≥ 4		48	56	0.476	0.490

Figure IV.2



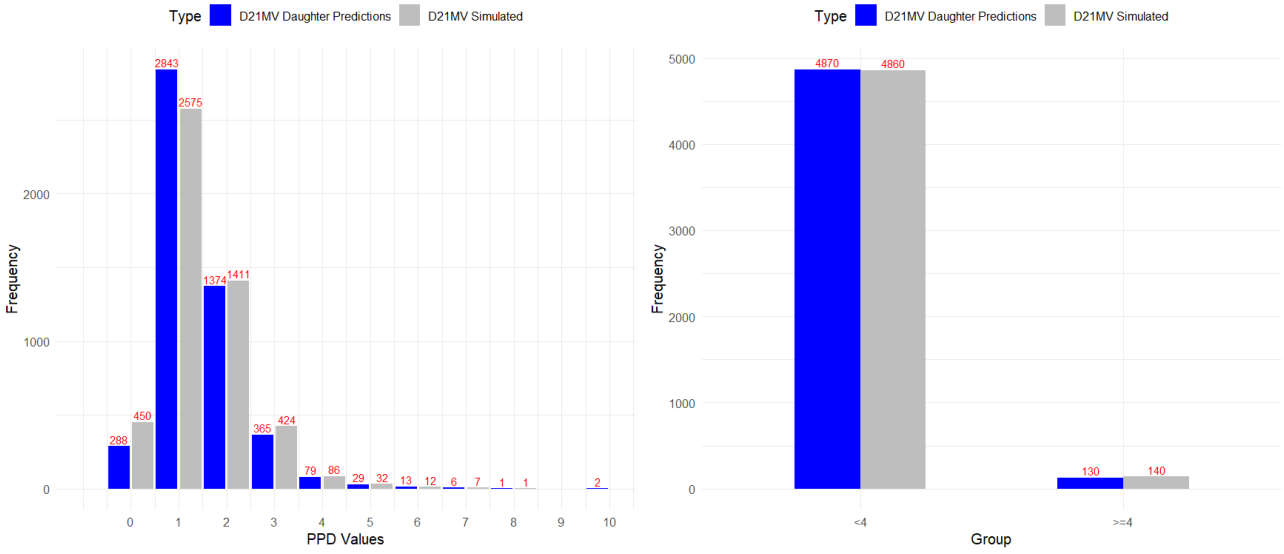
Histograms of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Site MV

Table IV.9: Chi-squared Test Results for Comparison of Simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
21MV	0	288	450	35.561	**2.47e-09**
	1	2843	2575	13.257	**2.70e-04**
	2	1374	1411	0.492	0.483
	3	365	424	4.412	**0.036**
	4	79	86	0.297	0.586
	5	29	32	0.148	0.701
	6	13	12	0.040	0.841
	7	6	7	0.077	0.782
	8	1	1	0	1
	10	2	2	0	1
PPD ≥ 4		130	140	0.308	0.579

Figure IV.3



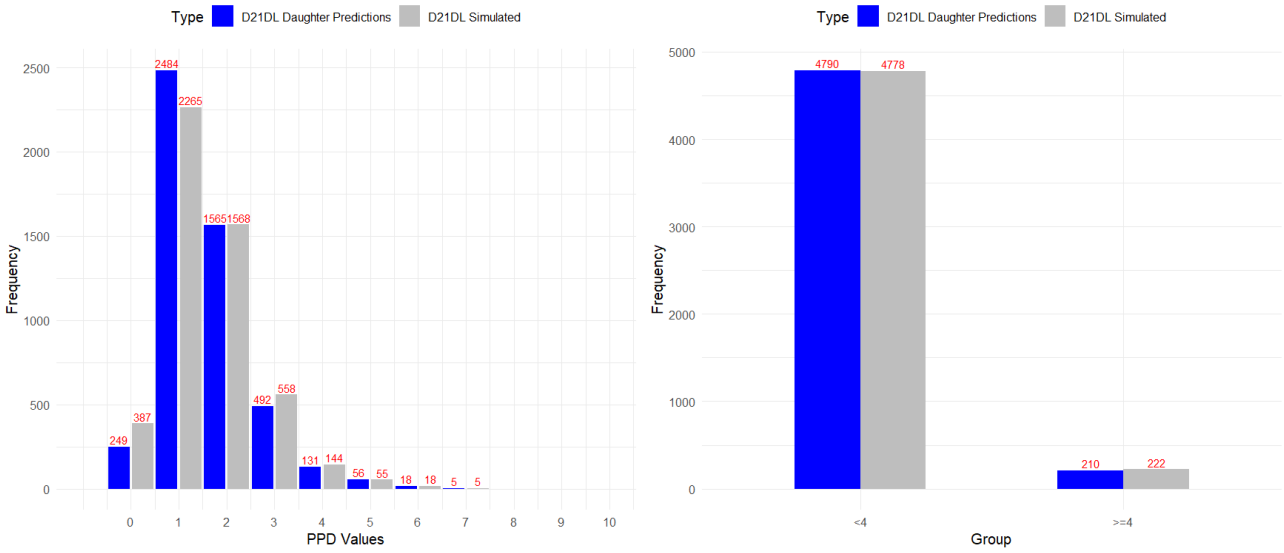
Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Site DL

Table IV.10: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
21DL	0	249	387	29.943	**4.45e-08**
	1	2484	2265	10.099	**1.48e-03**
	2	1565	1568	0.003	0.957
	3	492	558	4.149	**0.042**
	4	131	144	0.615	0.433
	5	56	55	0.009	0.924
	6	18	18	0	1
	7	5	5	0	1
PPD ≥ 4		210	222	0.293	0.589

Figure IV.4



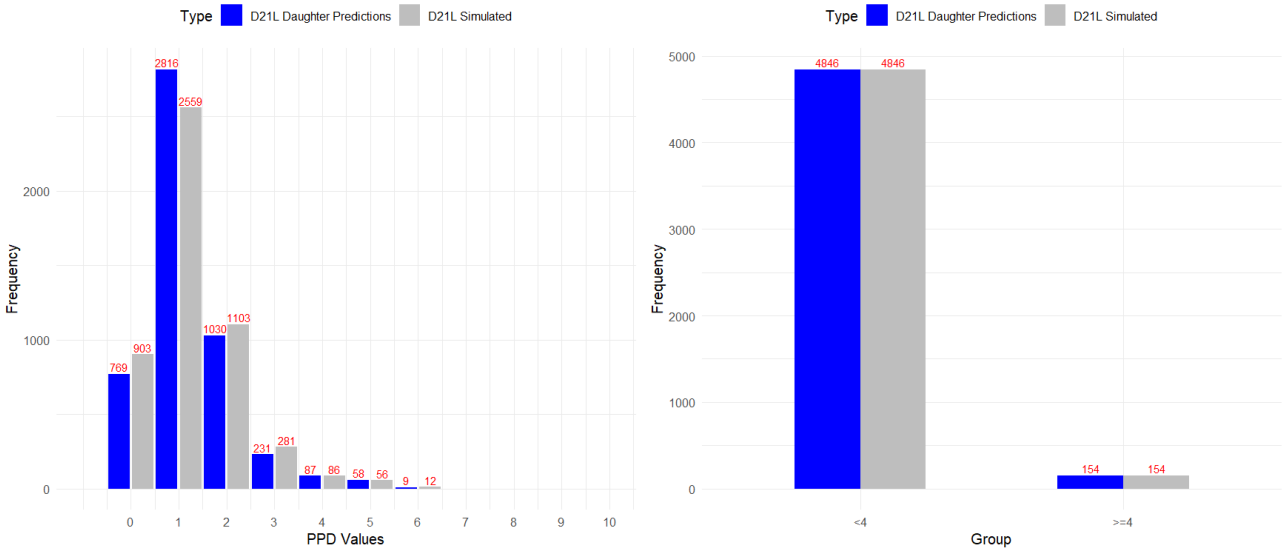
Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Site L

Table IV.11: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
	0	769	903	10.739	**1.05e-03**
	1	2816	2559	12.288	**4.60e-04**
	2	1030	1103	2.498	0.114
21L	3	231	281	4.883	**0.027**
	4	87	86	0.006	0.939
	5	58	56	0.035	0.851
	6	9	12	0.429	0.513
PPD ≥ 4		154	154	0.000	1.000

Figure IV.5



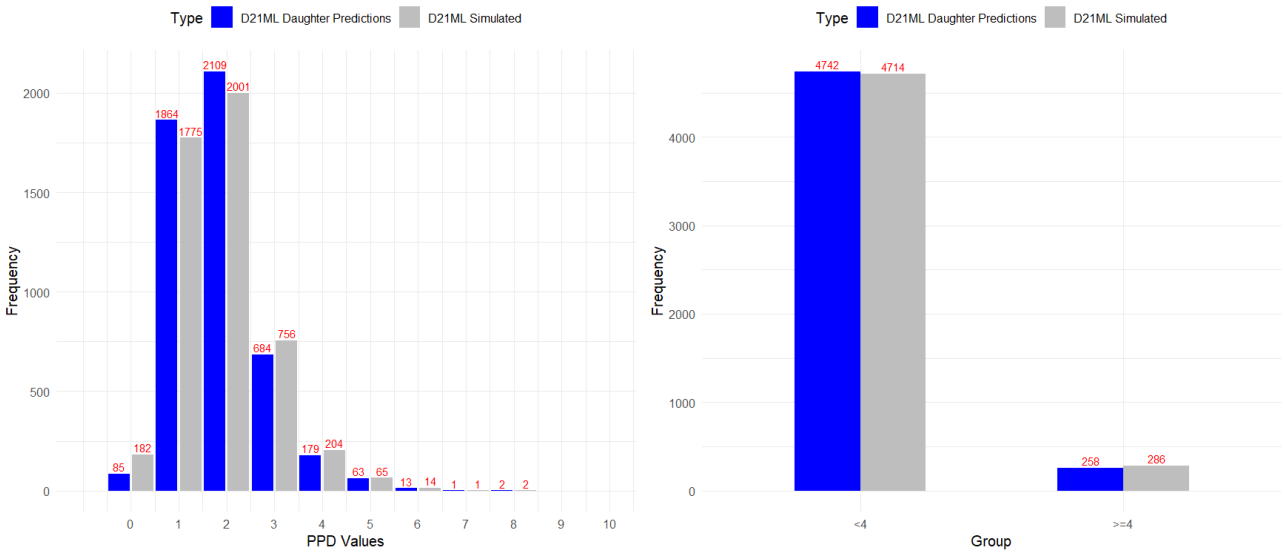
Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Site ML

Table IV.12: Chi-squared Test Results for Comparison of Simulated and Predicted PPD Counts by Value and Site

Site	PPD Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
	0	85	182	35.240	**2.92e-09**
	1	1864	1775	2.177	0.140
	2	2109	2001	2.838	0.092
	3	684	756	3.600	0.058
21ML	4	179	204	1.632	0.201
	5	63	65	0.031	0.860
	6	13	14	0.037	0.847
	7	1	1	0	1
	8	2	2	0	1
	PPD ≥ 4	258 286	1.417	0.234	

Figure IV.6



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

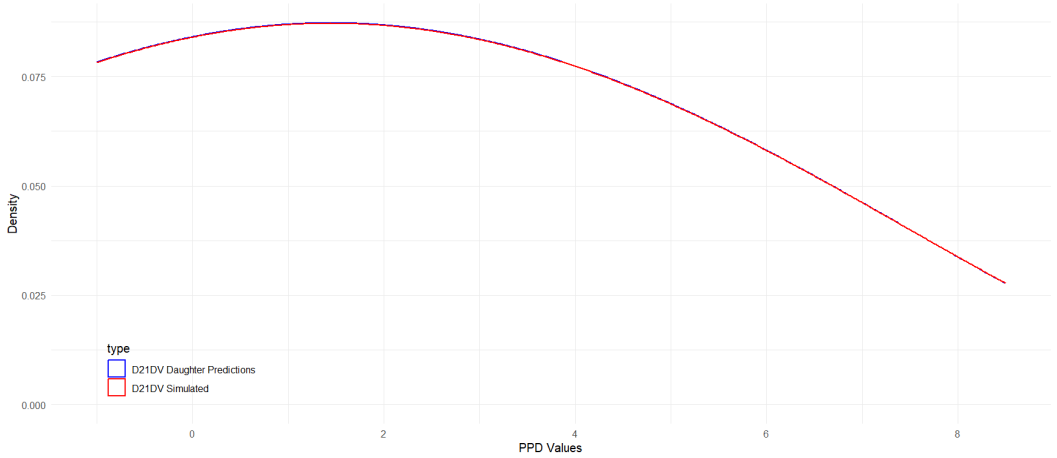
Comparation of Distributions Trough Kernels

Table IV.13: Kernel Density Estimates Differences and Kolmogorov-Smirnov Test

21	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
21DV	Biweight	2	2	2.77e-04	[3.25e-05, 2.47e-03]	0.076	7.77e-13
21V	Cosine	2	2	3.00e-03	[2.00e-03, 6.70e-03]	0.063	4.24e-09
21MV	Biweight	2	2	5.34e-04	[2.09e-04, 1.71e-03]	0.077	2.29e-13
21DL	Biweight	2	2	4.25e-03	[1.01e-04, 5.13e-03]	0.054	8.36e-07
21L	Biweight	2	2	5.00e-03	[4.68e-03, 5.30e-03]	0.066	9.05e-10
21ML	Biweight	2	2	1.00e-03	[2.65e-04, 3.00e-03]	0.042	2.96e-04

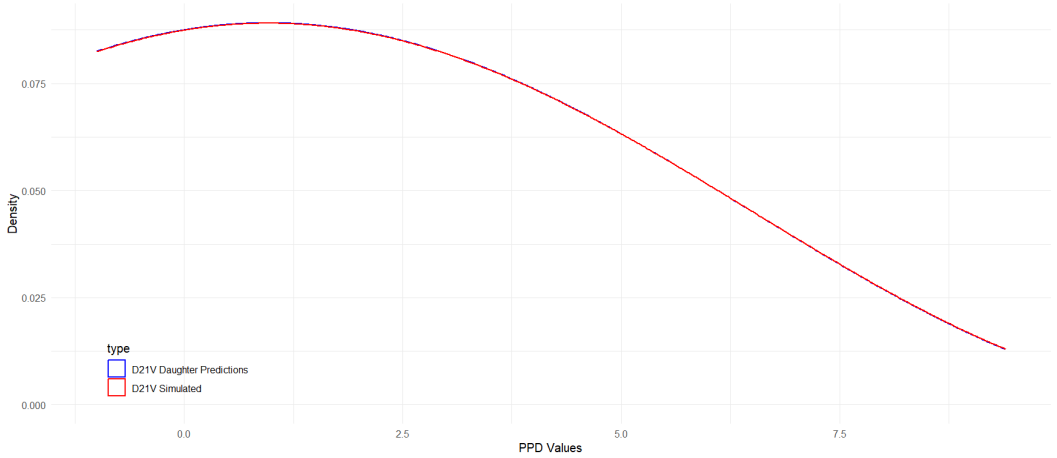
Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

Figure IV.7



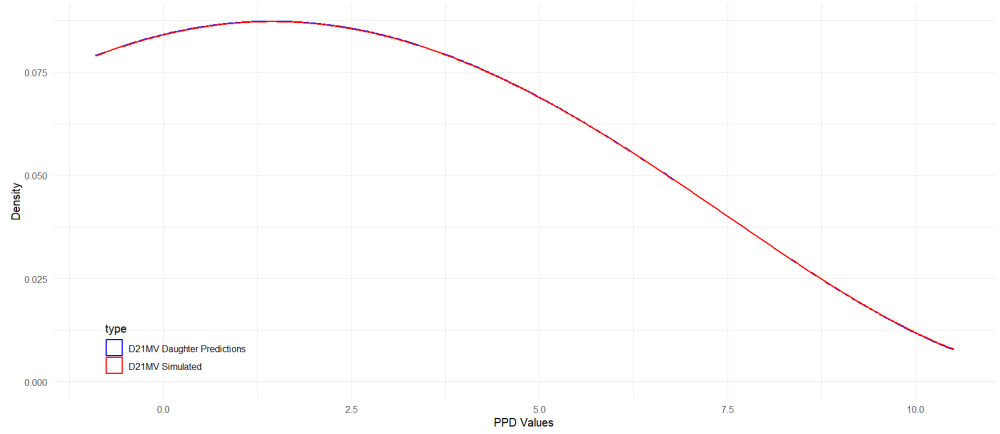
Optimal Kernel Density Plots of 21DV for Simulated and Predicted Data

Figure IV.8



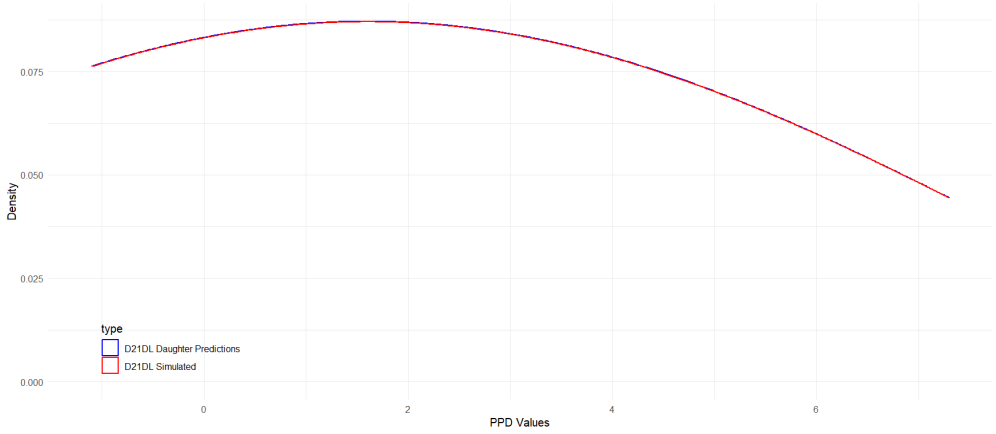
Optimal Kernel Density Plots of 21V for Simulated and Daughter Predicted

Figure IV.9



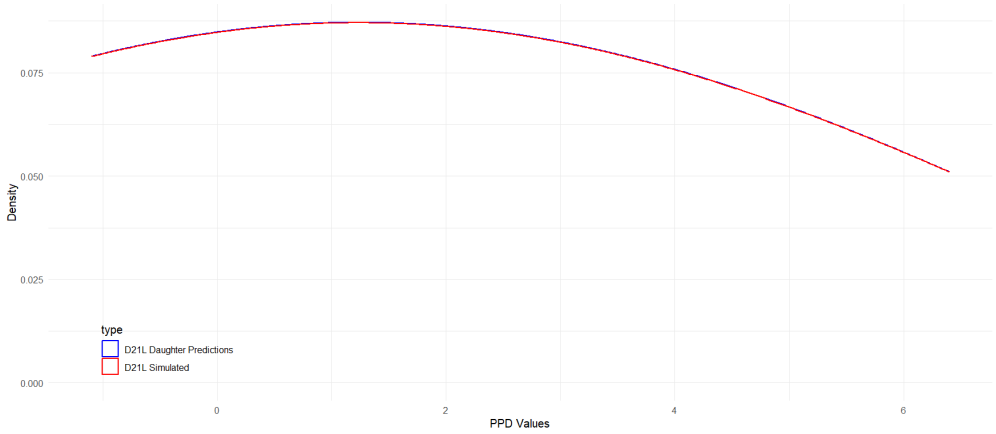
Optimal Kernel Density Plots of 21MV for Simulated and Daughter Predicted

Figure IV.10



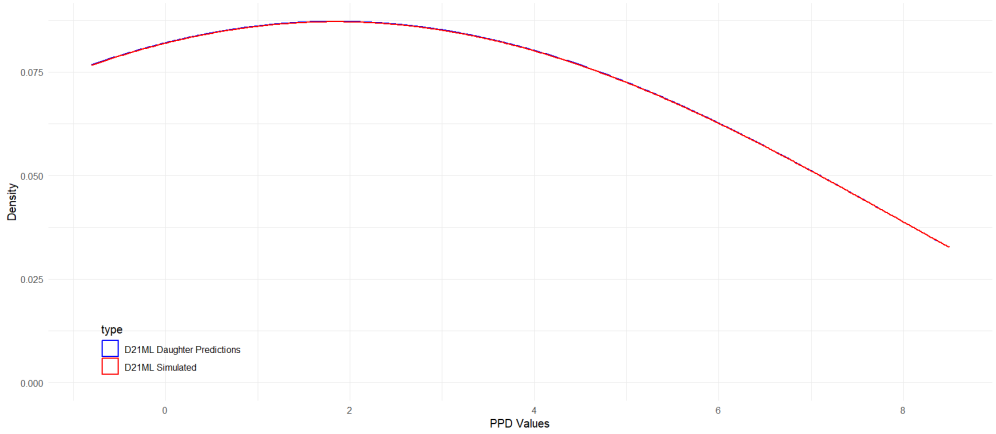
Optimal Kernel Density Plots of 21DL for Simulated and Daughter Predicted

Figure IV.11



Optimal Kernel Density Plots of 21L for Simulated and Daughter Predicted

Figure IV.12



Optimal Kernel Density Plots of 21ML for Simulated and Daughter Predicted

IV.2 Upper Lateral Incisor

Analysis of M22 Mother Models: Characteristics, Performance, and Feature Importance

Table IV.14: Distinctive Characteristics of the M22 Mother Models by Site

Metric	M22DV	M22V	M22MV	M22DL	M22L	M22ML
Mother Model Size (Kb)	386.10	261.00	232.60	270.80	179.80	268.00
N.Iter.	366	253	221	247	168	259
Init. Train. RMSE	1.212	1.305	1.254	1.311	1.018	1.290
Final Train. RMSE	1.86e-02	2.92e-02	2.22e-02	1.68e-02	4.38e-02	2.74e-02
Features	γSM.DV 12DV	γSM.V 12V	γSM.MV 12MV	γSM.DL 12DL	γSM.L 12L	γSM.ML 12ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table IV.15: Performance Metrics for the M22 Mother Models by Site

Metric	M22DV	M22V	M22MV	M22DL	M22L	M22ML
RMSE	0.019	0.029	0.022	0.017	0.044	0.027
MAE	0.003	0.005	0.004	0.003	0.006	0.005
MSE	3.46e-04	8.50e-04	4.95e-04	2.83e-04	1.92e-03	7.53e-04
R^2	99.95%	99.90%	99.94%	99.97%	99.75%	99.91%

Table IV.16: Mother Models features Importance Metrics - M22 by Site

Model	Feature	Gain	Cover	Frequency
M22DV	γSM.DV	0.630	0.702	0.706
	Original 12DV	0.370	0.298	0.294
M22V	γSM.DL	0.577	0.485	0.592
	Original 11DL	0.423	0.515	0.408
M22MV	γSM.MV	0.610	0.489	0.607
	Original 12MV	0.390	0.511	0.393
M22DL	γSM.DL	0.609	0.462	0.609
	Original 12DL	0.391	0.538	0.391
M22L	γSM.L	0.584	0.515	0.599
	Original 12L	0.416	0.485	0.401
M22ML	γSM.ML	0.570	0.583	0.657
	Original 12ML	0.430	0.417	0.343

Analysis of D22 Daughter Models: Characteristics, Performance, and Feature Importance

Table IV.17: Distinctive Characteristics of the D22 Daughter Models by Site

Metric	D22DV	D22V	D22MV	D22DL	D22L	D22ML
Model Size (Mb)	32.4	32.0	32.2	31.9	31.6	32.0
N.Iter.	30000	30000	30000	30000	30000	30000
Init. Train. RMSE	1.300	0.830	1.365	1.414	1.113	1.399
Final Train. RMSE	0.148	0.126	0.161	0.165	0.163	0.163

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Table IV.18: Performance Metrics for the Daughter Models - D22 Data

Metric	D22DV	D22V	D22MV	D22DL	D22L	D22ML
RMSE	0.148	0.126	0.161	0.165	0.297	0.163
MAE	0.116	0.100	0.126	0.130	0.235	0.128
MSE	0.022	0.016	0.026	0.027	0.088	0.026
R^2	97.53%	97.39%	97.49%	97.22%	90.10%	97.37%

Table IV.19: Feature Importance Metrics Of Daughter Models D22 by Site

Model	Feature	Gain	Cover	Frequency
D22DV	Mother Predictions	0.454	0.331	0.336
	New Mother Predictions	0.279	0.088	0.073
	Simulated 12DV	0.266	0.581	0.591
D22V	Mother Predictions	0.440	0.259	0.276
	New Mother Predictions	0.275	0.126	0.090
	Simulated 12V	0.286	0.615	0.634
D22MV	Mother Predictions	0.448	0.309	0.315
	New Mother Predictions	0.276	0.097	0.077
	Simulated 12MV	0.276	0.594	0.608
D22DL	Mother Predictions	0.433	0.301	0.321
	New Mother Predictions	0.292	0.108	0.082
	Simulated 12DL	0.276	0.591	0.597
D22L	Mother Predictions	0.588	0.282	0.301
	New Mother Predictions	0.328	0.137	0.106
	Simulated 12L	0.084	0.581	0.594
D22ML	Mother Predictions	0.434	0.298	0.312
	New Mother Predictions	0.292	0.109	0.084
	Simulated 12ML	0.275	0.593	0.605

Daughter Model D22 Predictions

Table IV.20: Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted for Site 22

Groups	Data	**22DV**	**22V**	**22MV**	**22DL**	**22L**	**22ML**
$\geq$ 4 mm	Simulated	172	18	208	191	107	201
	Predicted	165	17	197	176	102	184
χ^2 statistic		0.1106	0.2927	0.2573	0.5544	0.0782	0.6916
$p - value$		0.740	0.589	0.612	0.457	0.780	0.406

Table IV.21: Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site

Site	Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
D22DV	0	227	365	32.169	**1.41e-08**
	1	2788	2565	9.290	**2.303-03**
	2	1503	1523	0.132	0.716
	3	317	375	4.861	**0.027**
	4	114	119	0.107	0.743
	5	34	36	0.057	0.811
	6	9	9	0	1
	7	2	2	0	1
	8	5	5	0	1
	9	1	1	0	1
D22V	0	1374	1418	0.693	0.405
	1	2875	2765	2.145	0.143
	2	651	705	705	0.957
	3	83	94	0.684	0.408
	4	10	11	0.048	0.433
	5	4	4	0	1
	7	1	1	0	1
	8	2	2	0	1
D22MV	0	289	455	37.038	**1.16e-09**
	1	2659	2401	13.155	**2.87e-04**
	2	1475	1511	0.434	0.510
	3	380	425	2.516	0.113
	4	130	140	0.370	0.543
	5	29	30	0.017	0.896
	6	29	30	0.017	0.896
	7	8	7	0.067	0.796
	8	1	1	0	1
D22DL	0	212	345	31.758	**1.75e-08**
	1	2415	2203	9.732	**1.80e-03**
	2	1795	1788	0.014	0.907
	3	402	473	5.761	**0.016**
	4	117	128	0.494	0.482
	5	27	30	0.158	0.691
	6	18	19	0.027	0.869
	7	4	4	0	1
	8	8	8	0	1
	9	2	1	0.333	0.564
	10	1	NA	NA	NA
D22L	0	817	961	11.663	**6.40e-04**
	1	2910	2594	18.142	**2.05e-05**
	2	989	1110	6.975	**8.03e-03**
	3	182	228	5.161	**0.023**
	4	51	55	0.151	0.698
	5	37	34	0.127	0.722
	6	11	15	0.615	0.433
	7	1	1	0	1
	9	2	1	0.333	0.564
	10	1	NA	NA	NA
D22ML	0	269	406	27.806	**1.34e-07**
	1	2522	2307	9.572	**2.00e-03**
	2	1535	1539	0.005	0.942
	3	490	547	3.133	0.077
	4	116	128	0.590	0.442
	5	53	57	0.145	0.703
	6	7	8	0.067	0.796
	7	6	6	0	1
	8	2	1	0.333	0.564
	9	1	NA	NA	NA

Comparation of Distributions Trough Kernels

Table IV.22: Kernel Density Estimates Differences and Kolmogorov-Smirnov Test

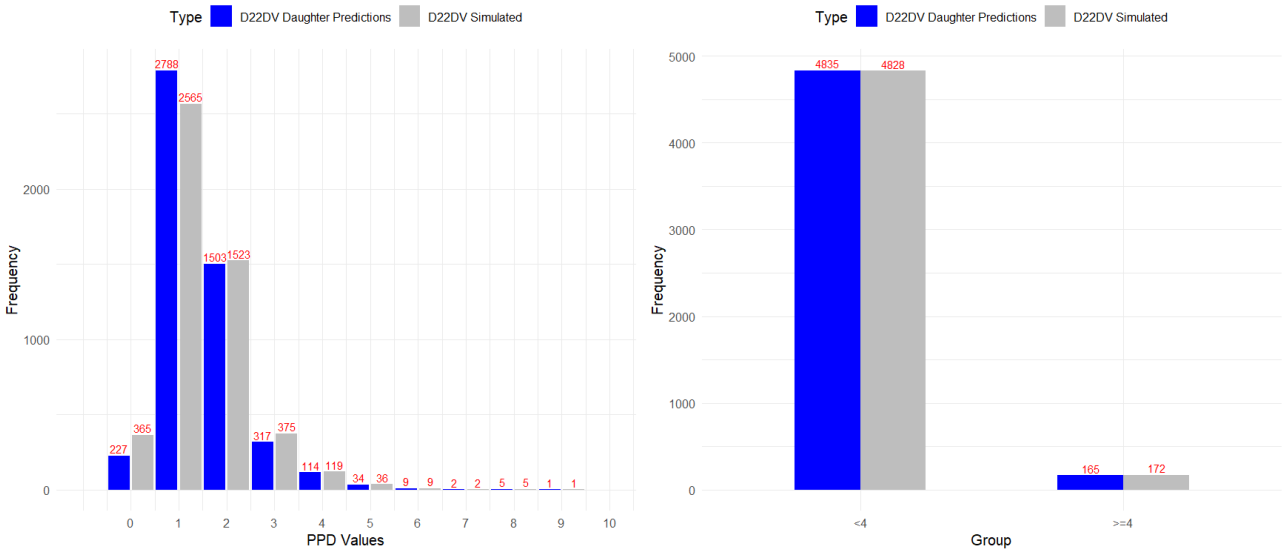
Site	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
22DV	Cosine	2	2	3.39e-03	[3.02e-03, 4.48e-03]	0.059	5.52e-08
22V	Biweight	2	2	4.71e-03	[2.14e-03, 7.32e-03]	0.064	2.90e-09
22MV	Biweight	2	2	2.06e-03	[1.36e-04, 2.53e-03]	0.056	3.10e-07
22DL	Biweight	2	2	9.14e-04	[3.35e-04, 3.79e-03]	0.057	2.21e-07
22L	Biweight	2	2	5.19e-03	[2.05e-03, 1.15e-02]	0.068	2.39e-10
22ML	Biweight	2	2	4.04e-03	[2.27e-04, 4.69e-03]	0.058	7.85e-08

Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

Visual Comparisons of Proportions and Distributions

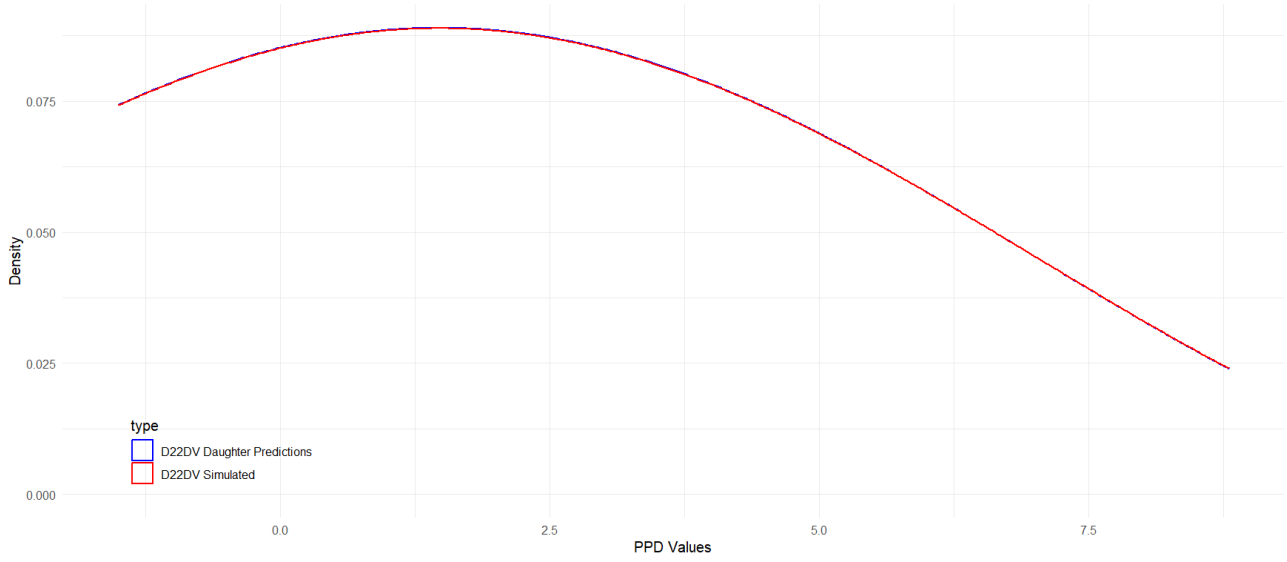
Site DV

Figure IV.13



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

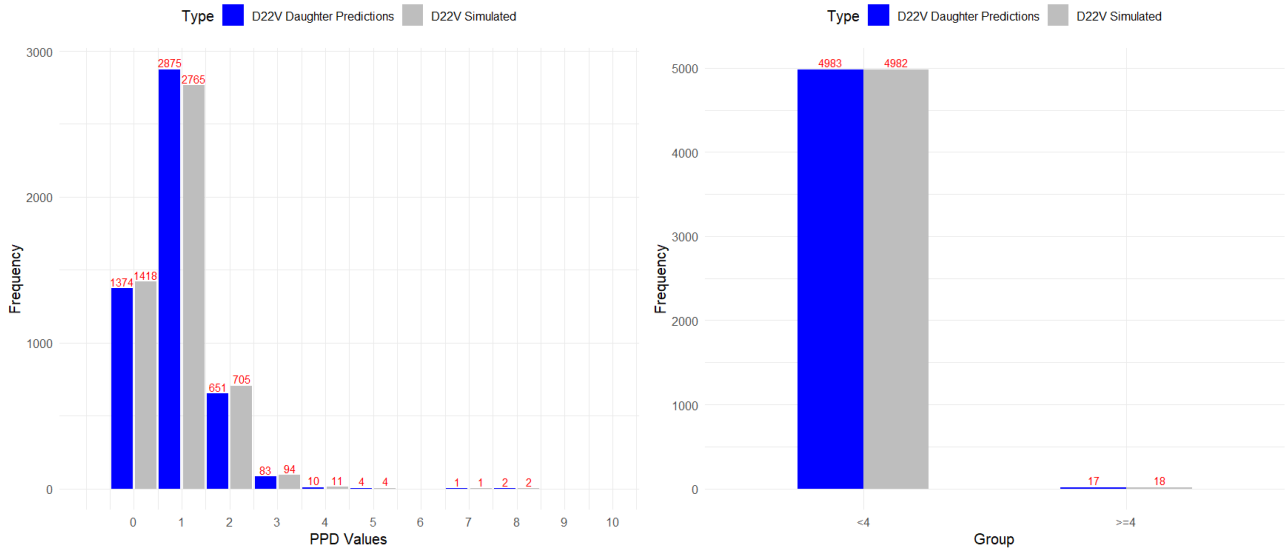
Figure IV.14



Optimal Kernel Density Plots of 22DV for Simulated and Daughter Predicted

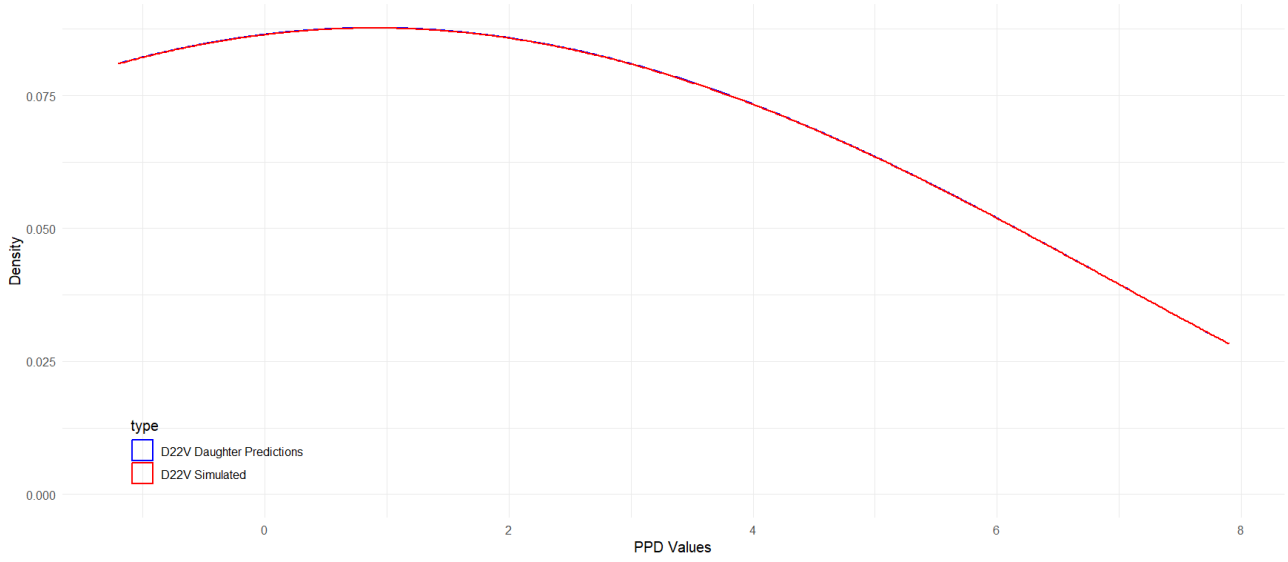
Site V

Figure IV.15



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

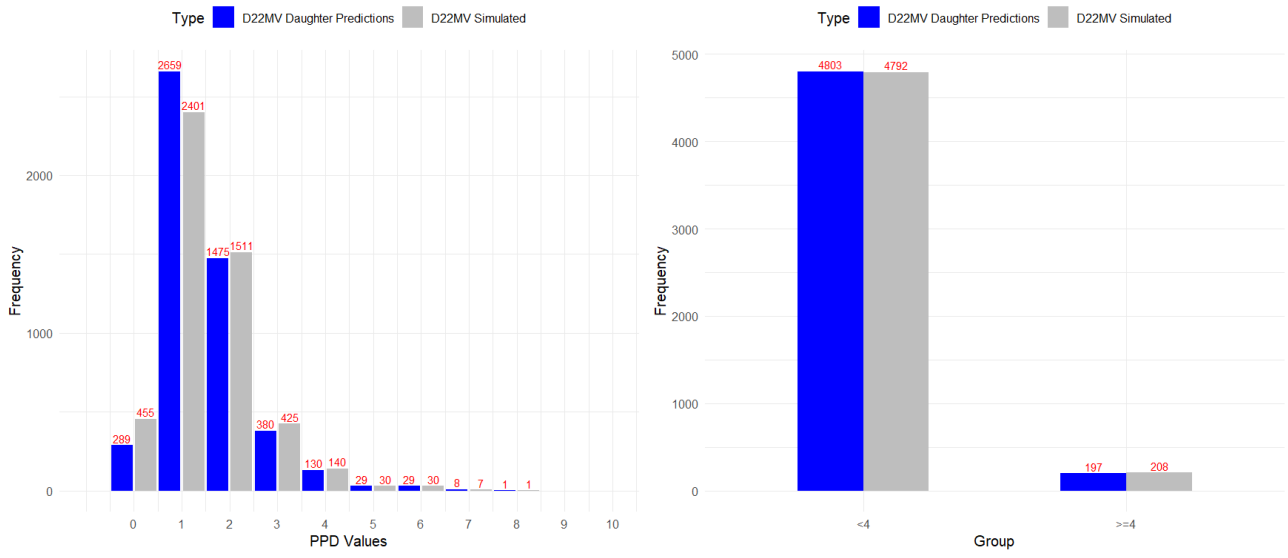
Figure IV.16



Optimal Kernel Density Plots of 22V for Simulated and Daughter Predicted

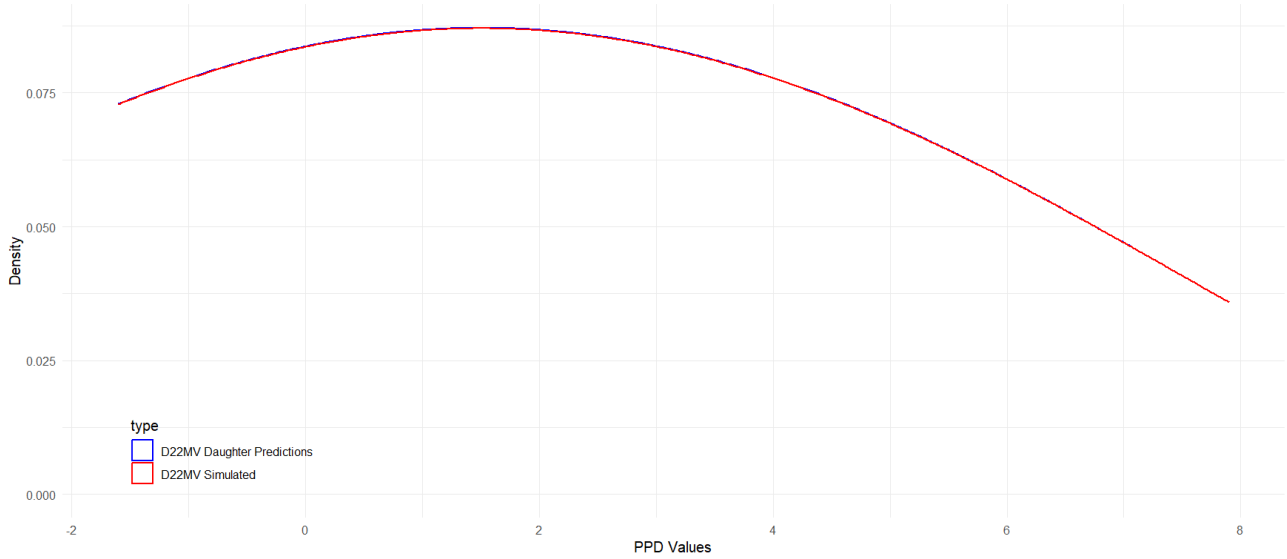
Site MV

Figure IV.17



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

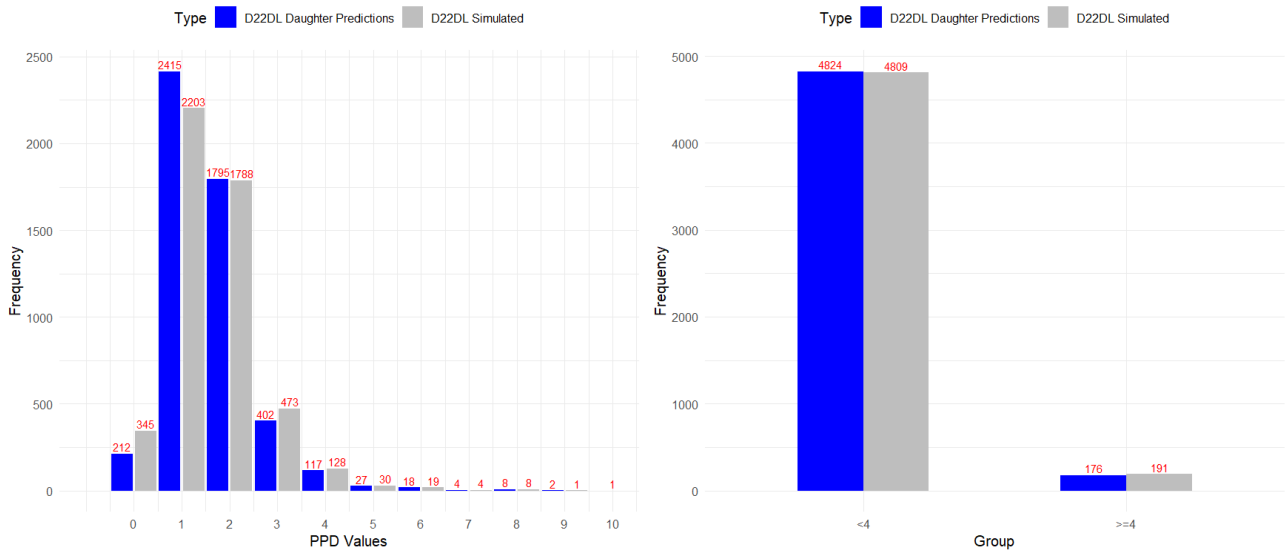
Figure IV.18



Optimal Kernel Density Plots of 22MV for Simulated and Daughter Predicted

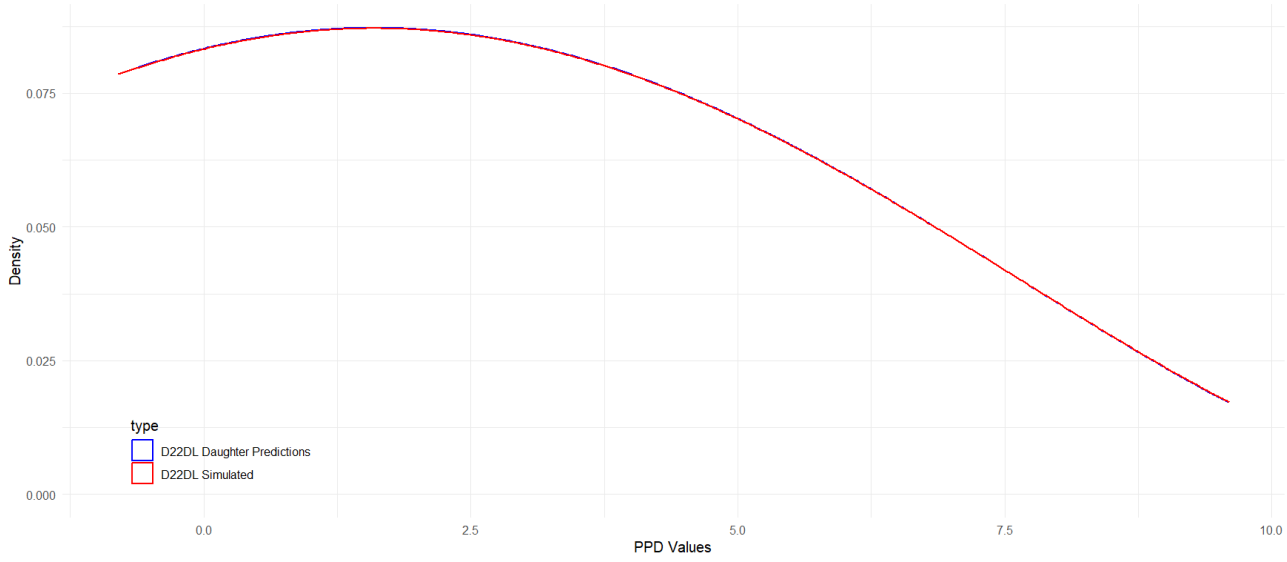
Site DL

Figure IV.19



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

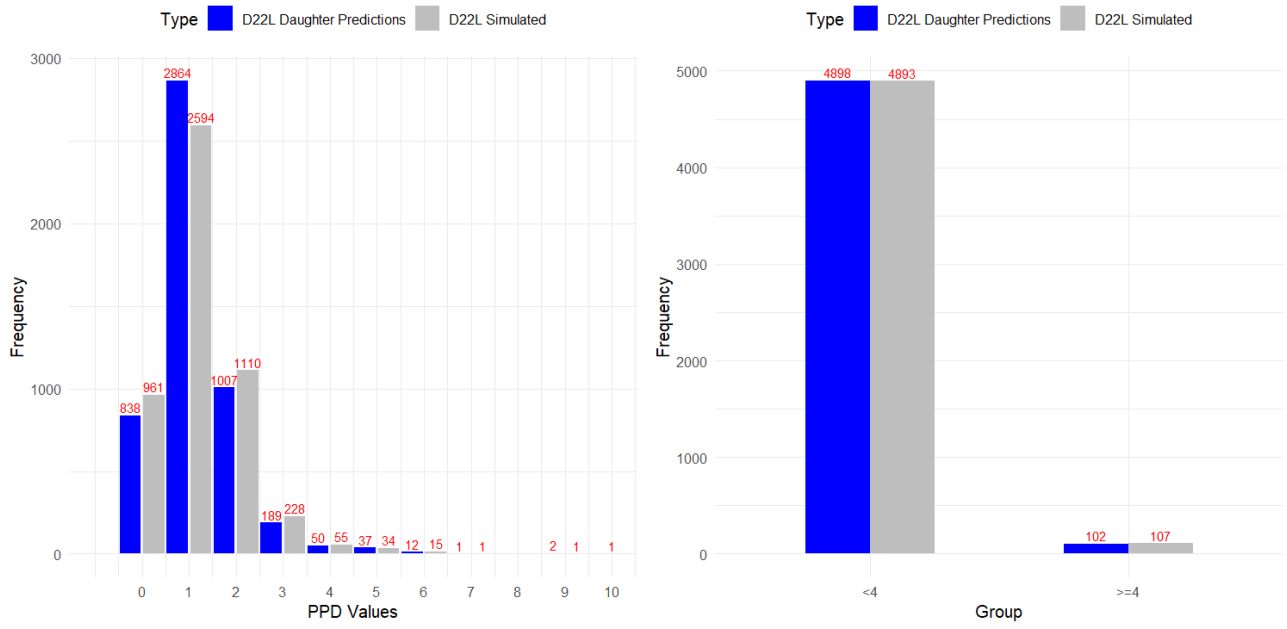
Figure IV.20



Optimal Kernel Density Plots of 22DL for Simulated and Daughter Predicted

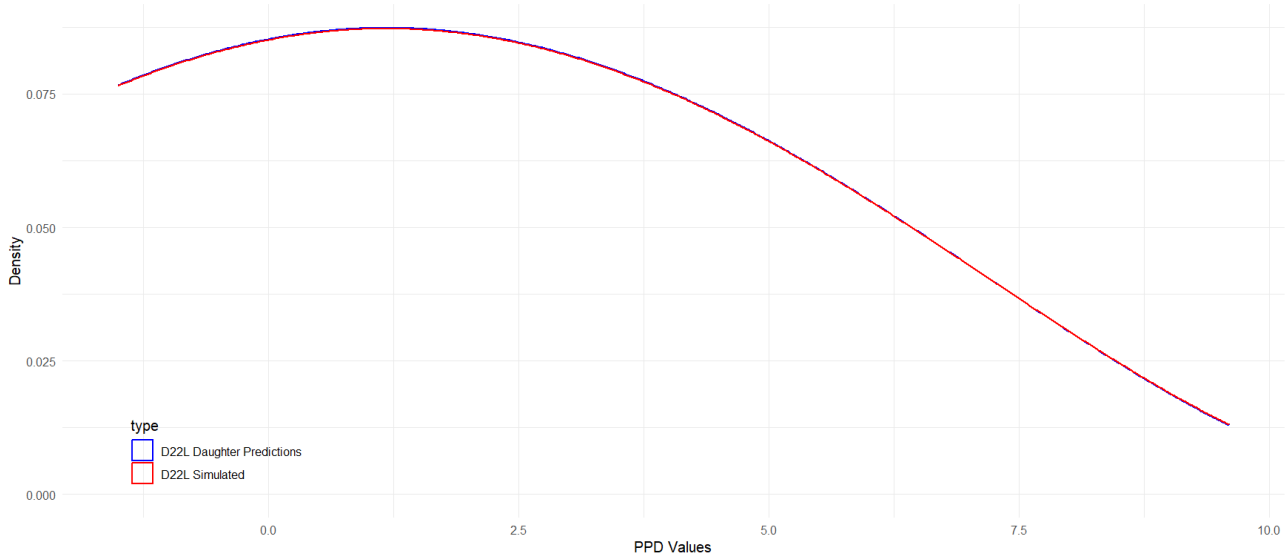
Site L

Figure IV.21



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

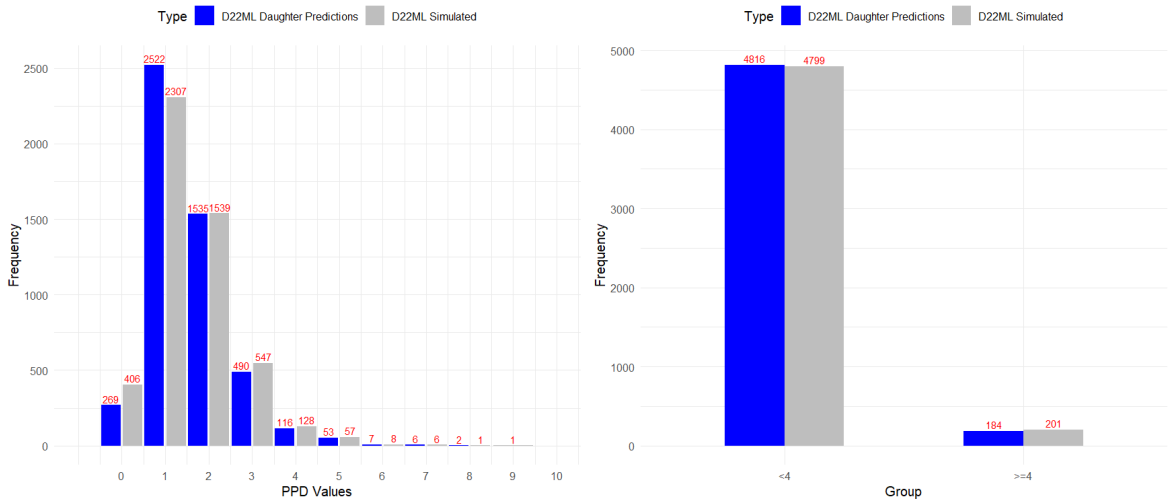
Figure IV.22



Optimal Kernel Density Plots of 22L for Simulated and Daughter Predicted

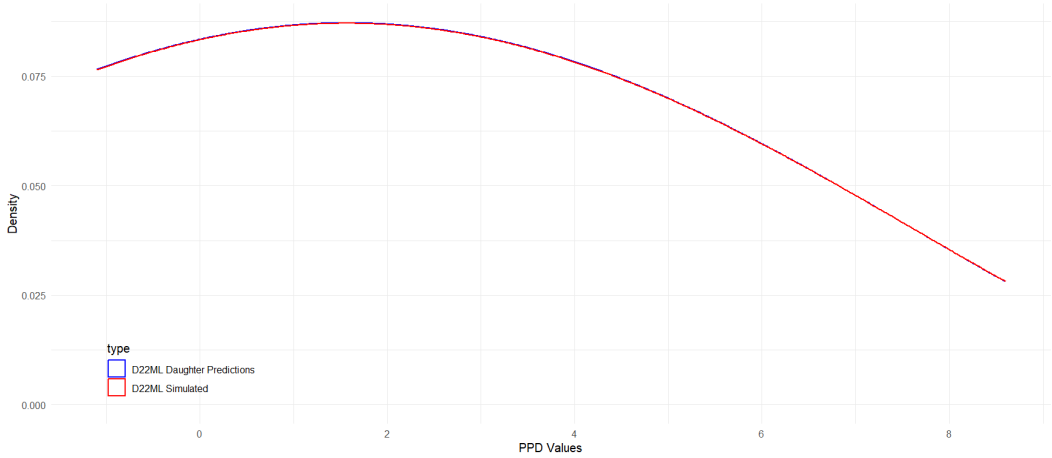
Site ML

Figure IV.23



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

Figure IV.24



Optimal Kernel Density Plots of 22ML for Simulated and Daughter Predicted

IV.3 Upper Left Canine

Analysis of M23 Mother Models: Characteristics, Performance, and Feature Importance

Table IV.23: Distinctive Characteristics of the M23 Mother Models by Site

Metric	M23DV	M23V	M23MV	M23DL	M23L	M23ML
Mother Model Size (Kb)	234.80	215.50	368.20	254.30	386.70	352.10
N.Iter.	228	207	365	239	374	348
Init. Train. RMSE	1.252	0.783	1.182	1.386	1.091	1.300
Final Train. RMSE	4.99e-02	3.81e-02	2.00e-02	3.40e-02	1.76e-02	2.34e-02
Features	γSM.DV 13DV	γSM.V 13V	γSM.MV 13MV	γSM.DL 13DL	γSM.L 13L	γSM.ML 13ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table IV.24: Performance Metrics for the M23 Mother Models by Site

Metric	M23DV	M23V	M23MV	M23DL	M23L	M23ML
RMSE	0.038	0.038	0.020	0.034	0.017	0.023
MAE	0.007	0.004	0.003	0.006	0.003	0.005
MSE	2.49e-03	1.45e-03	3.97e-04	1.16e-03	3.05e-04	5.41e-04
R^2	99.67%	99.75%	99.94%	99.87%	99.96%	99.94%

Table IV.25: Mother Models features Importance Metrics - M23 by Site

Model	Feature	Gain	Cover	Frequency
M23DV	γSM.DV	0.637	0.379	0.546
	Original 13DV	0.363	0.621	0.454
M23V	γSM.V	0.582	0.484	0.625
	Original 13V	0.418	0.516	0.375
M23MV	γSM.MV	0.615	0.488	0.593
	Original 13MV	0.385	0.512	0.407
M23DL	γSM.DL	0.619	0.470	0.625
	Original 13DL	0.381	0.530	0.375
M23L	γSM.L	0.598	0.589	0.656
	Original 13L	0.402	0.411	0.344
M23ML	γSM.ML	0.569	0.494	0.594
	Original 13ML	0.431	0.506	0.406

Analysis of D23 Daughter Models: Characteristics, Performance, and Feature Importance

Table IV.26: Distinctive Characteristics of the D32 Daughter Models by Site

Metric	D23DV	D23V	D23MV	D23DL	D23L	D23ML
Model Size (Mb)	32.3	32.0	26.6	32.5	32.2	32.3
N.Iter.	29999	30000	24739	30000	30000	30000
Init. Train. RMSE	1.365	0.859	1.260	1.486	1.180	1.387
Final Train. RMSE	0.152	0.134	0.156	0.159	0.154	0.155

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Table IV.27: Performance Metrics for the Daughter Models - D23 Data

Metric	**D23DV**	**D23V**	**D23MV**	**D23DL**	**D23L**	**D23ML**
RMSE	0.152	0.134	0.156	0.159	0.154	0.155
MAE	0.120	0.106	0.124	0.125	0.122	0.122
MSE	0.023	0.018	0.024	0.025	0.024	0.024
R^2	97.41%	97.27%	96.91%	97.58%	97.48%	97.51%

Table IV.28: Feature Importance Metrics Of Daughter Models D23 by Site

Model	**Feature**	**Gain**	**Cover**	**Frequency**
D23DV	Mother Predictions	0.415	0.299	0.316
	New Mother Predictions	0.303	0.103	0.074
	Simulated 13DV	0.282	0.598	0.610
D23V	Mother Predictions	0.445	0.258	0.276
	New Mother Predictions	0.271	0.124	0.089
	Simulated 13V	0.284	0.618	0.635
D23MV	Mother Predictions	0.461	0.283	0.301
	New Mother Predictions	0.280	0.116	0.085
	Simulated 13MV	0.260	0.601	0.614
D23DL	Mother Predictions	0.455	0.302	0.325
	New Mother Predictions	0.272	0.106	0.081
	Simulated 13DL	0.273	0.592	0.594
13L	Mother Predictions	0.439	0.299	0.307
	New Mother Predictions	0.286	0.100	0.076
	Simulated 13L	0.275	0.601	0.617
13ML	Mother Predictions	0.452	0.308	0.320
	New Mother Predictions	0.274	0.090	0.073
	Simulated 13ML	0.274	0.603	0.607

Daughter Model M23 Predictions

Table IV.29: Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted for Site 23

Groups	Data	**23DV**	**23V**	**23MV**	**23DL**	**23L**	**23ML**
$\geq$ 4 mm	Simulated	155	44	126	239	117	155
	Predicted	134	42	116	216	115	143
χ^2 statistic		0.6846	0.0117	0.3430	1.1144	0.0044	0.4185
$p - value$		0.408	0.914	0.558	0.291	0.947	0.518

Table IV.30: Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site

Site	Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
23DV	0	147	282	42.483	**7.13e-11**
	1	2567	2380	7.069	**7.84e-03**
	2	1740	1711	0.244	0.623
	3	406	472	4.961	**0.026**
	4	98	109	0.585	0.445
	5	18	21	0.231	0.631
	6	10	11	0.048	0.827
	7	2	2	0	1
	8	7	6	0.077	0.782
	9	4	5	0.111	0.739
	10	1	1	0	1
23V	0	1353	1399	0.769	0.381
	1	2975	2840	3.134	0.077
	2	558	630	4.364	**0.037**
	3	72	87	1.415	0.234
	4	25	27	0.077	0.782
	5	8	8	0	1
	6	6	6	0	1
	7	1	1	0	1
	10	1	1	0	1
	11	1	1	0	1
23MV	0	178	315	38.071	**6.82e-10**
	1	2824	2610	8.428	**3.7e-03**
	2	1556	1567	0.039	0.844
	3	326	382	4.429	**0.035**
	4	84	95	0.676	0.411
	5	15	14	0.034	0.853
	6	10	10	0	1
	7	2	2	0	1
	8	5	4	0.111	0.739
	9	1	NA	NA	NA
23DL	0	148	287	44.416	**2.65e-11**
	1	2343	2157	7.688	**5.6e-03**
	2	1777	1746	0.273	0.601
	3	516	571	2.783	0.095
	4	148	168	1.266	0.261
	5	37	39	0.053	0.819
	6	15	15	0	1
	7	5	6	0.091	0.763
	8	10	10	0	1
	9	1	1	0	1
23L	0	598	744	15.884	**7.40e-05**
	1	2923	2660	12.389	**4.30e-04**
	2	1146	1204	1.431	0.232
	3	218	275	6.590	**0.010**
	4	76	75	0.007	0.935
	5	11	14	0.360	0.549
	6	21	20	0.024	0.876
	7	2	3	0.200	0.655
	9	1	1	0	1
	10	1	1	0	1
	11	3	2	0.200	0.655
	12	1	NA	NA	NA
23ML	0	189	336	41.160	**1.40e-10**
	1	2550	2333	9.643	**1.9e-03**
	2	1681	1669	0.043	0.836
	3	437	507	5.191	**0.023**
	4	80	90	0.588	0.443
	5	37	39	0.053	0.819
	6	10	10	0	1
	7	4	4	0	1
	8	6	5	0.091	0.763
	9	5	6	0.091	0.763
	10	1	1	0	1

Comparation of Distributions Trough Kernels

Table IV.31: Kernel Density Estimates Differences and Kolmogorov-Smirnov Test

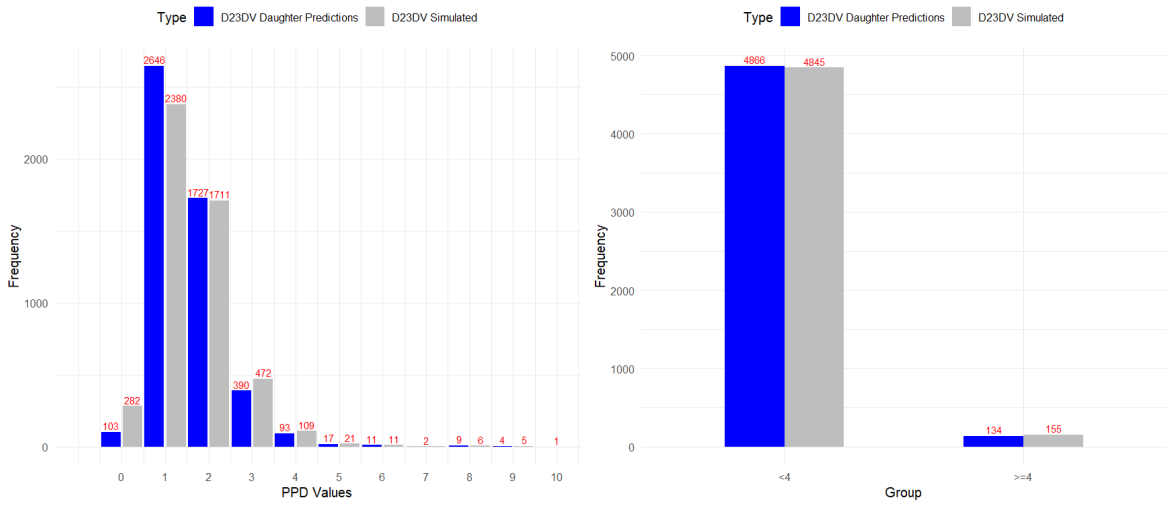
23	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
M23DV	Cosine	2	2	4.85e-03	[3.66e-03, 6.65e-03]	0.0588	6.21e-08
M23V	Biweight	2	2	1.51e-03	[4.98e-04, 4.75e-03]	0.0678	2.09e-10
M23MV	Biweight	2	2	2.16e-03	[8.23e-05, 5.28e-03]	0.0674	2.73e-10
M23DL	Biweight	2	2	4.39e-03	[2.70e-03, 5.63e-03]	0.0566	7.78e-07
M23L	Biweight	2	2	7.16e-03	[4.03e-03, 8.18e-03]	0.0618	1.02e-08
M23ML	Biweight	2	2	2.84e-03	[2.26e-03, 4.99e-03]	0.0544	7.50e-07

Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

Visual Comparisons of Proportions and Distributions

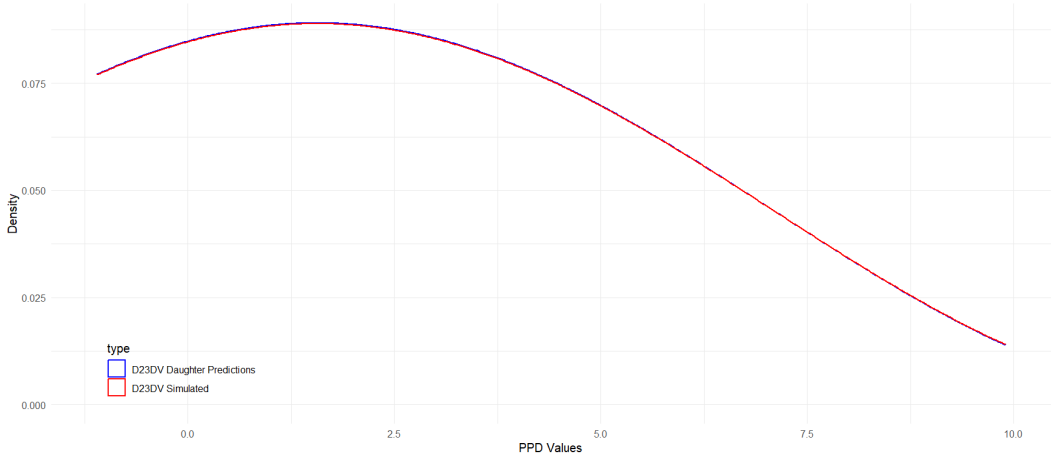
Site DV

Figure IV.25



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

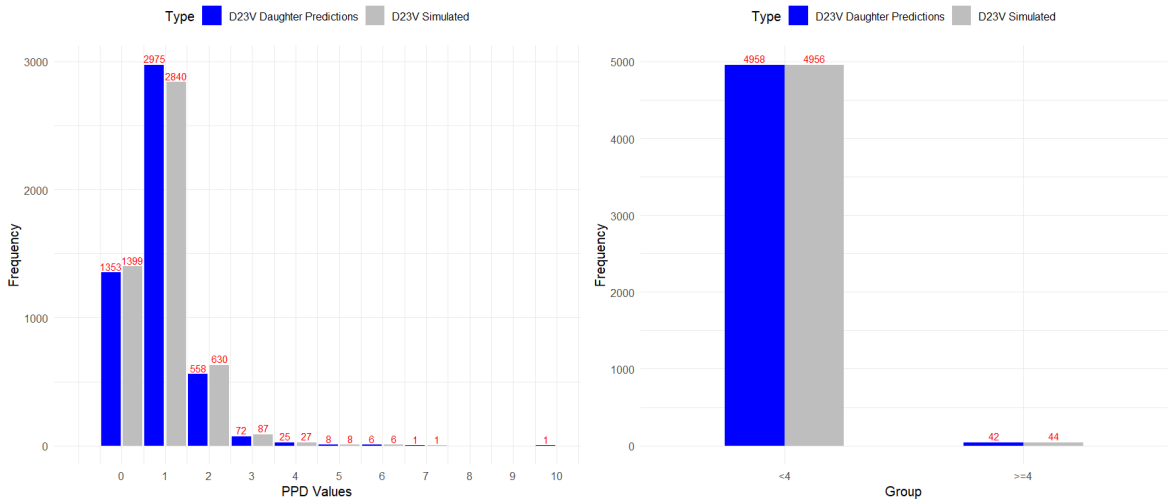
Figure IV.26



Optimal Kernel Density Plots of 23DV for Simulated and Daughter Predicted

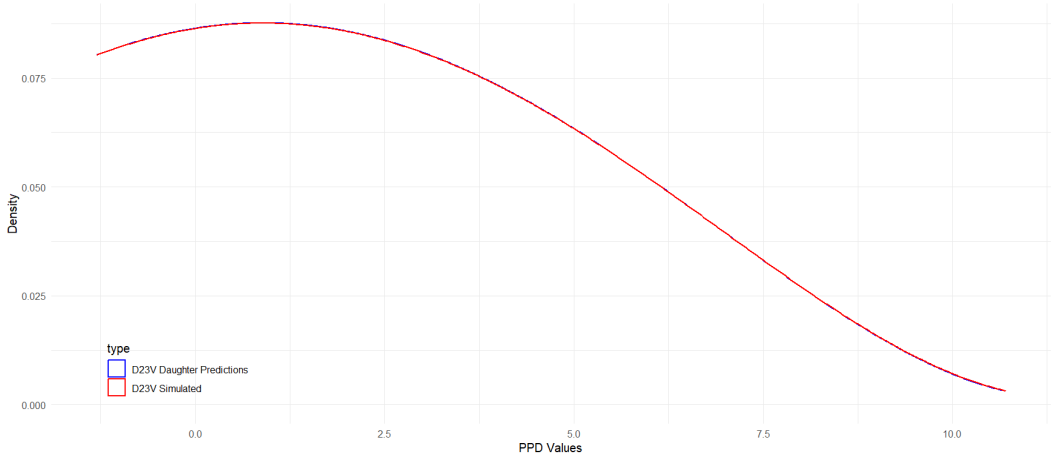
Site V

Figure IV.27



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

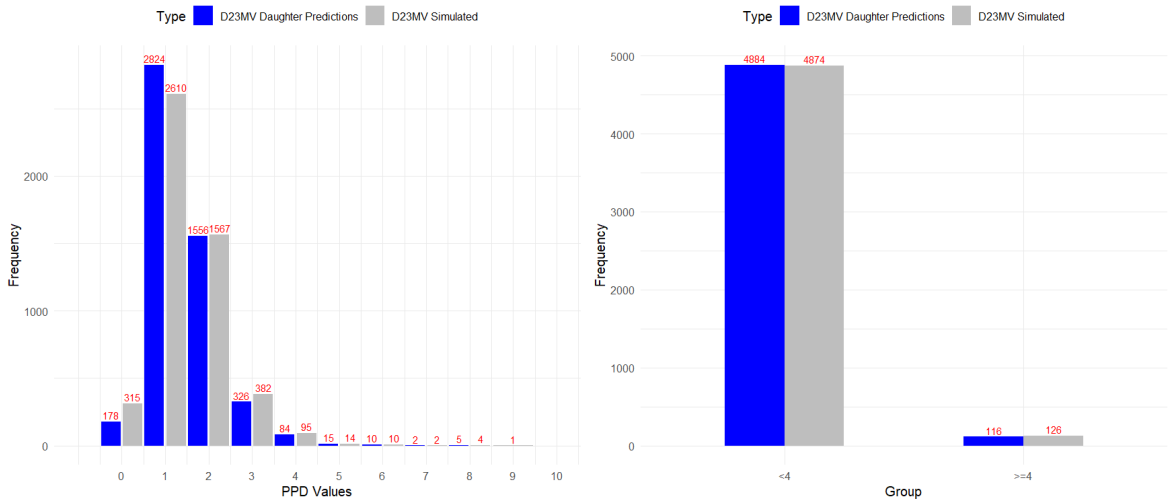
Figure IV.28



Optimal Kernel Density Plots of 23V for Simulated and Daughter Predicted

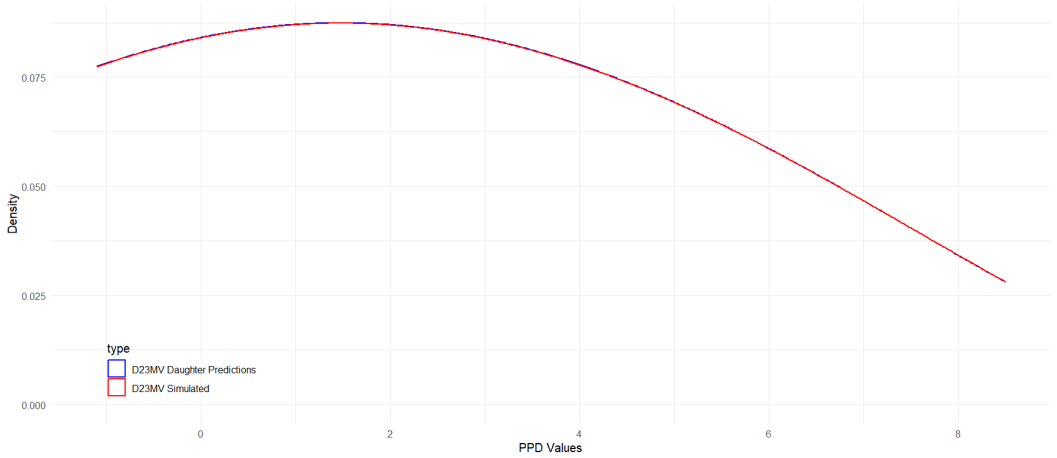
Site MV

Figure IV.29



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

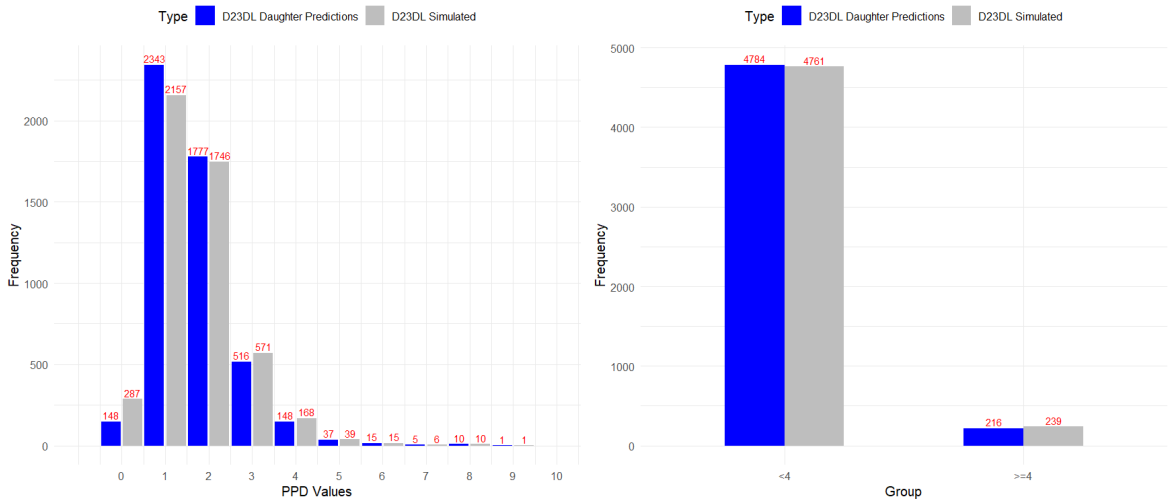
Figure IV.30



Optimal Kernel Density Plots of 23MV for Simulated and Daughter Predicted

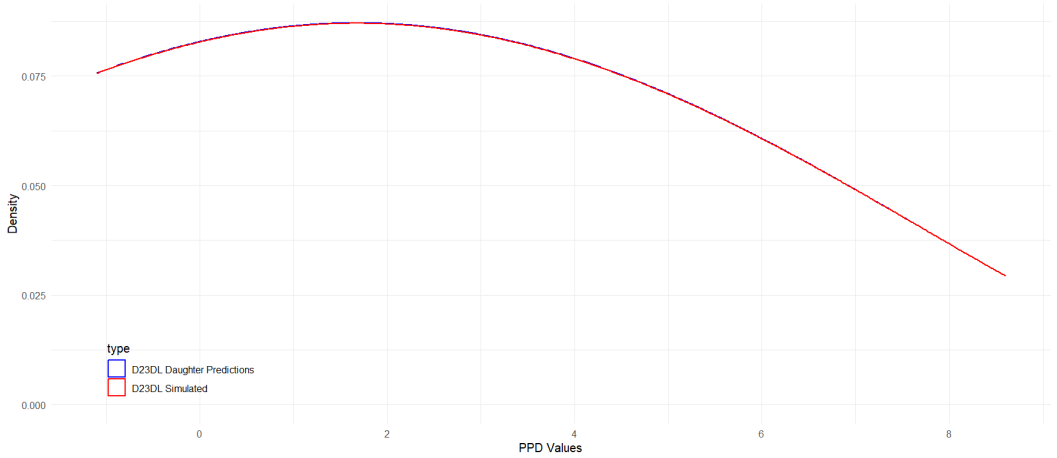
Site DL

Figure IV.31



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

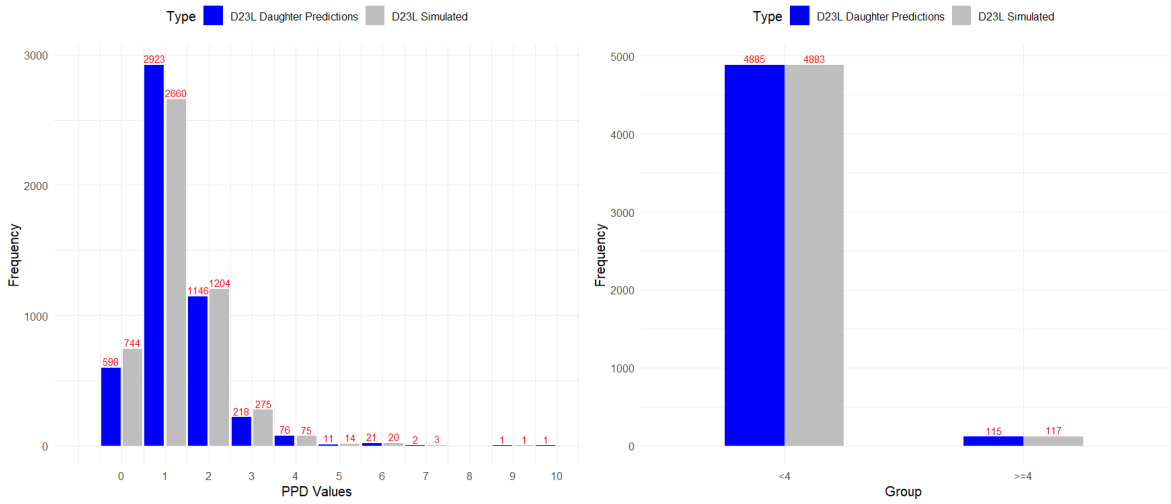
Figure IV.32



Optimal Kernel Density Plots of 23DL for Simulated and Daughter Predicted

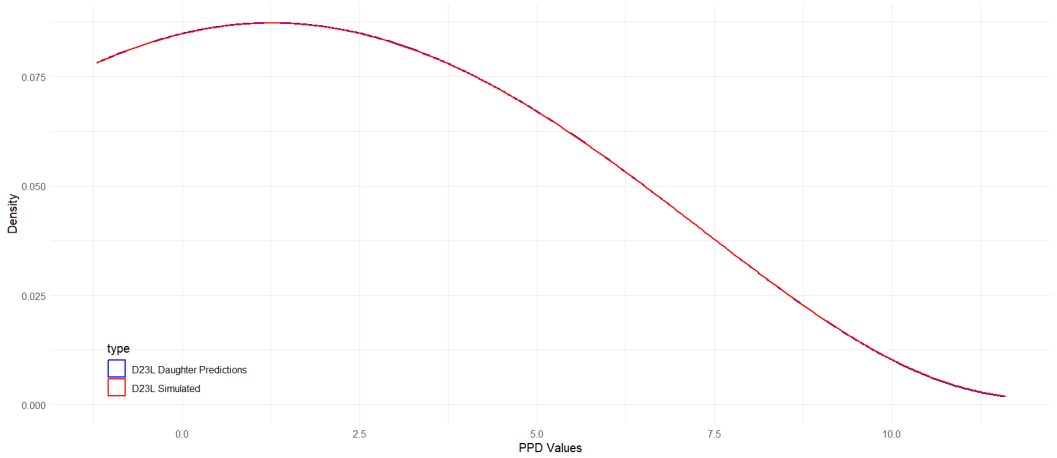
Site L

Figure IV.33



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

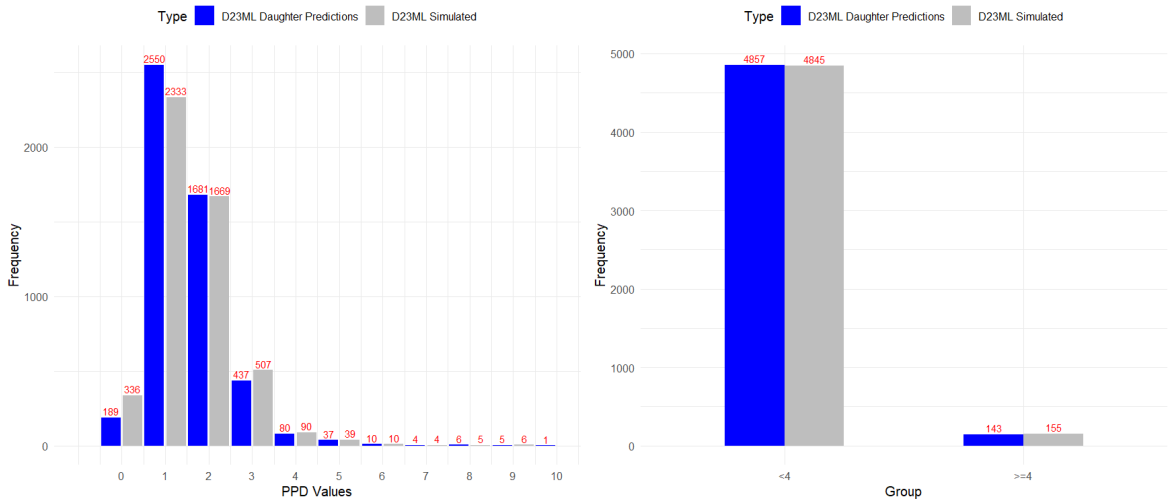
Figure IV.34



Optimal Kernel Density Plots of 23L for Simulated and Daughter Predicted

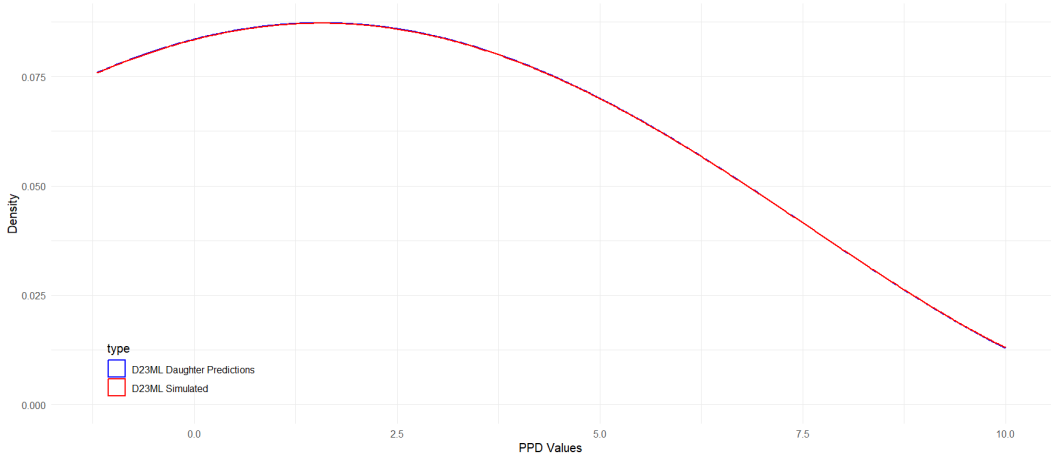
Site ML

Figure IV.35



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Figure IV.36



Optimal Kernel Density Plots of 23ML for Simulated and Daughter Predicted

IV.4 Upper First Premolar

Analysis of M24 Mother Models: Characteristics, Performance, and Feature Importance

Table IV.32: Distinctive Characteristics of the M24 Mother Models by Site

Metric	M24DV	M24V	M24MV	M24DL	M24L	M24ML
Mother Model Size (Kb)	205.30	163.50	317.70	319.00	418.50	242.80
N.Iter.	196	158	320	313	396	226
Init. Train. RMSE	1.255	0.761	1.315	1.412	1.074	1.224
Final Train. RMSE	3.29e-02	3.13e-02	3.12e-02	2.37e-02	4.78e-03	3.30e-02
Features	γSM.DV 14DV	γSM.V 14V	γSM.MV 14V	γSM.DL 14DL	γSM.L 14L	γSM.ML 14ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table IV.33: Performance Metrics for the M24 Mother Models by Site

Metric	M24DV	M24V	M24MV	M24DL	M24L	M24ML
RMSE	0.033	0.031	0.031	0.024	0.005	0.033
MAE	0.005	0.004	0.005	0.004	0.001	0.005
MSE	1.08e-03	9.81e-04	9.74e-04	5.61e-04	2.28e-05	1.09e-03
R^2	99.86%	99.81%	99.87%	99.93%	99.99%	99.85%

Table IV.34: Mother Models features Importance Metrics - M24 by Site

Model	Feature	Gain	Cover	Frequency
M24DV	γSM.DV	0.655	0.462	0.585
	Original 14DV	0.345	0.538	0.415
M24V	γSM.V	0.606	0.467	0.588
	Original 14V	0.394	0.533	0.413
M24MV	γSM.MV	0.646	0.484	0.556
	Original 14MV	0.354	0.516	0.444
M24DL	γSM.DL	0.622	0.561	0.643
	Original 14DL	0.378	0.439	0.357
M24L	γSM.L	0.638	0.568	0.671
	Original 14L	0.362	0.432	0.329
M24ML	γSM.ML	0.647	0.521	0.640
	Original 14ML	0.353	0.479	0.360

Analysis of D24 Daughter Models: Characteristics, Performance, and Feature Importance

Table IV.35: Distinctive Characteristics of the D42 Daughter Models by Site

Metric	D42DV	D42V	D42MV	D42DL	D42L	D42ML
Model Size (Mb)	32.4	15.9	16.1	32.5	27.6	32.5
N.Iter.	30000	15000	15000	30000	25935	30000
Init. Train. RMSE	1.341	0.831	1.402	1.503	1.182	1.313
Final Train. RMSE	0.150	0.165	0.207	0.149	0.167	0.141

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Table IV.36: Performance Metrics for the Daughter Models - D24 Data

Metric	**D24DV**	**D24V**	**D24MV**	**D24DL**	**D24L**	**D24ML**
RMSE	0.150	0.241	0.154	0.149	0.167	0.141
MAE	0.119	0.192	0.122	0.118	0.131	0.111
MSE	0.023	0.058	0.024	0.022	0.028	0.020
R^2	97.41%	89.56%	97.29%	97.53%	96.88%	97.57%

Table IV.37: Feature Importance Metrics of Daughter Models D24 by Site

Model	**Feature**	**Gain**	**Cover**	**Frequency**
D24DV	Mother Predictions	0.446	0.251	0.287
	14DV	0.279	0.596	0.609
	New Mother Predictions	0.275	0.153	0.104
D24V	Mother Predictions	0.572	0.298	0.281
	New Mother Predictions	0.361	0.132	0.111
	14V	0.067	0.571	0.608
D24MV	Mother Predictions	0.447	0.280	0.307
	New Mother Predictions	0.278	0.128	0.093
	14MV	0.275	0.592	0.600
D24DL	Mother Predictions	0.449	0.297	0.320
	New Mother Predictions	0.276	0.112	0.084
	14DL	0.275	0.591	0.596
D24L	Mother Predictions	0.455	0.256	0.288
	New Mother Predictions	0.284	0.144	0.105
	14L	0.261	0.599	0.607
D24ML	Mother Predictions	0.449	0.269	0.304
	New Mother Predictions	0.279	0.140	0.099
	14ML	0.273	0.591	0.597

Daughter Model D24 Predictions

Table IV.38: Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted for Site 24

Groups	Data	**24DV**	**24V**	**24MV**	**24DL**	**24L**	**24ML**
$\geq$ 4 mm	Simulated	173	19	191	237	106	142
	Predicted	162	18	176	217	104	128
χ^2 statistic		0.3089	0.0000	0.7279	1.0133	0.0049	0.6433
$p - value$		0.578	1.000	0.394	0.314	0.944	0.423

Table IV.39: Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site

Site	Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
24DV	0	174	308	37.253	**1.04e-09**
	1	2573	2377	7.761	**5.34e-03**
	2	1786	1768	0.091	0.763
	3	305	374	7.012	**8.10e-03**
	4	93	100	0.254	0.614
	5	48	53	0.248	0.619
	6	16	14	0.133	0.715
	7	3	4	0.143	0.705
	8	2	2	0	1
24V	0	1170	1243	2.208	0.137
	1	3067	2918	3.709	0.054
	2	647	702	2.242	0.134
	3	98	118	1.852	0.174
	4	12	13	0.040	0.841
	5	5	5	0	1
	6	1	1	0	1
24MV	0	94	243	65.878	**4.80e-16**
	1	2464	2263	8.547	**3.50e-03**
	2	1835	1782	0.777	0.378
	3	433	521	8.117	**4.40e-03**
	4	117	129	0.585	0.444
	5	44	46	0.044	0.833
	6	11	13	0.167	0.683
	7	1	2	0.333	0.564
	8	1	1	0	1
24DL	0	52	161	55.779	**8.11e-14**
	1	2145	2017	3.937	**0.047**
	2	1999	1948	0.659	0.417
	3	587	635	1.885	0.170
	4	155	173	0.988	0.320
	5	44	47	0.099	0.753
	6	13	14	0.037	0.847
	7	4	4	0	1
	8	1	1	0	1
24L	0	519	677	20.873	**4.91e-6**
	1	2952	2666	14.560	**1.40e-04**
	2	1146	1218	2.193	0.139
	3	279	333	4.765	**0.029**
	4	46	46	0	1
	5	35	37	0.056	0.814
	6	22	21	0.023	0.879
	7	1	2	0.333	0.564
24ML	0	144	259	32.816	**1.01e-08**
	1	2715	2517	7.493	**6.20e-03**
	2	1584	1623	0.474	0.491
	3	429	459	1.014	0.314
	4	92	101	0.420	0.517
	5	25	29	0.296	0.586
	6	4	5	0.111	0.739
	7	4	4	0	1
	8	3	3	0	1

Comparation of Distributions Trough Kernels

Table IV.40: Kernel Density Estimates Differences and Kolmogorov-Smirnov Test

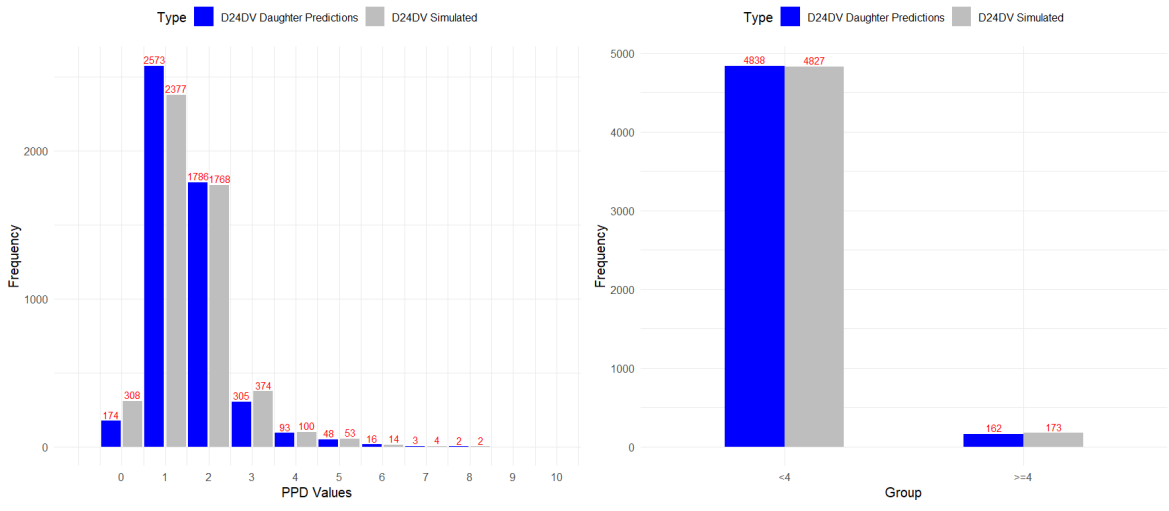
24	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
M24DV	Cosine	2	2	1.41e-03	[8.50e-04, 1.75e-03]	0.059	6.99e-08
M24V	Cosine	2	2	2.91e-03	[5.77e-04, 5.09e-03]	0.099	0
M24MV	Biweight	2	2	6.98e-03	[2.08e-03, 1.35e-02]	0.077	1.96e-13
M24DL	Biweight	2	2	1.73e-03	[4.88e-04, 5.68e-03]	0.043	1.77e-04
M24L	Cosine	2	2	1.46e-03	[2.82e-04, 2.35e-03]	0.069	1.21e-10
M24ML	Biweight	2	2	1.30e-03	[7.49e-04, 3.49e-03]	0.054	9.31e-07

Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

Visual Comparisons of Proportions and Distributions

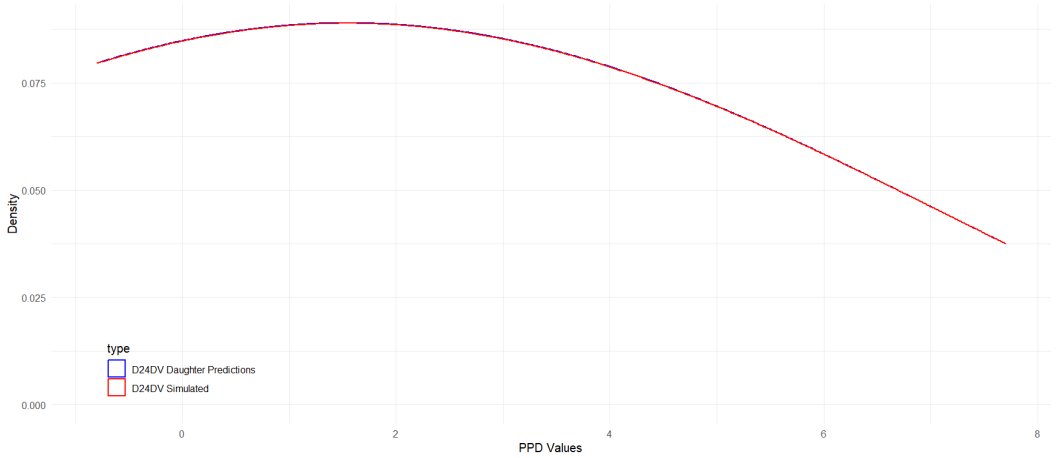
Site DV

Figure IV.37



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

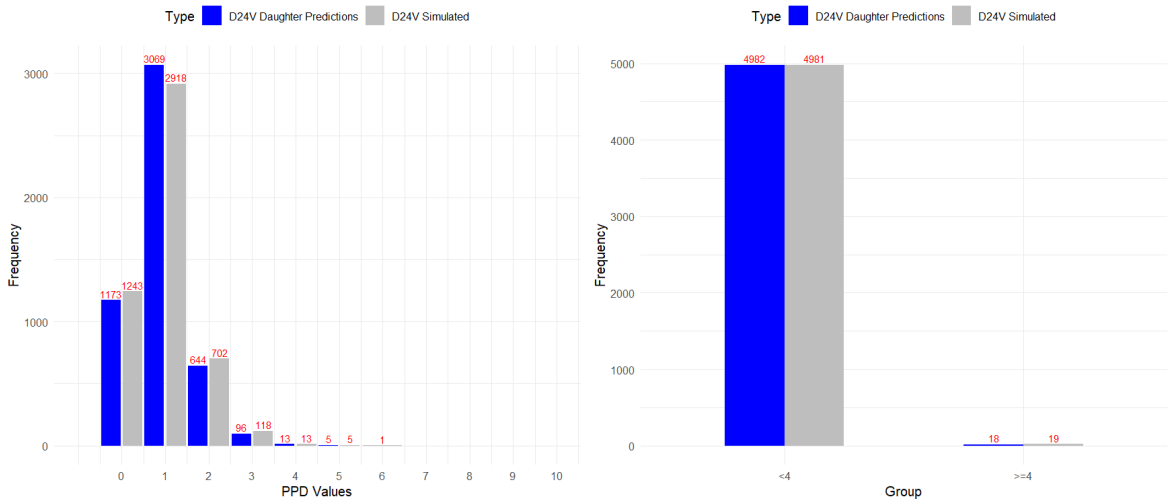
Figure IV.38



Optimal Kernel Density Plots of 24DV for Simulated and Daughter Predicted

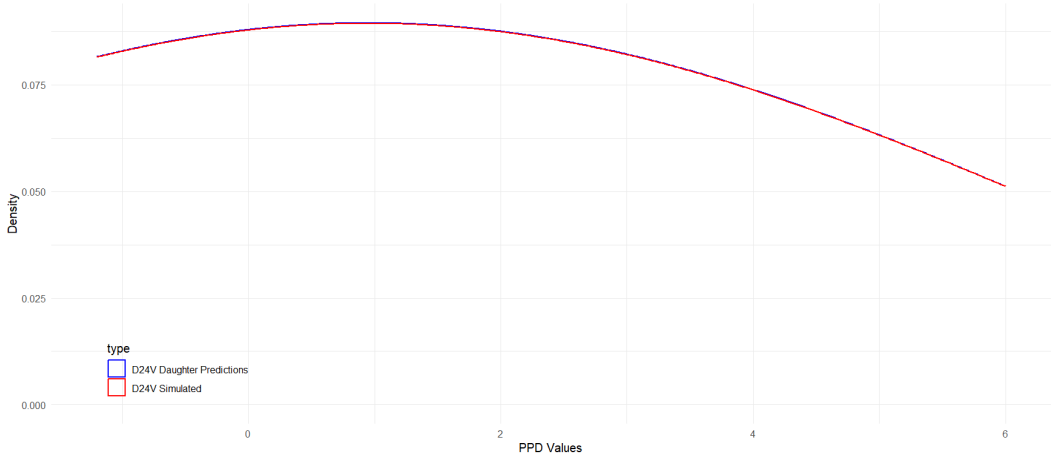
Site V

Figure IV.39



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

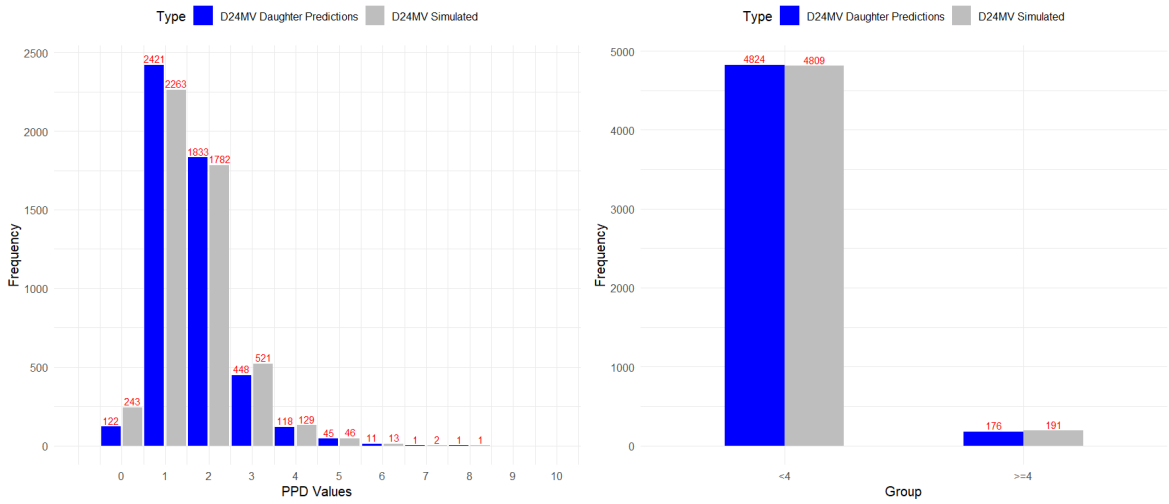
Figure IV.40



Optimal Kernel Density Plots of 24V for Simulated and Daughter Predicted

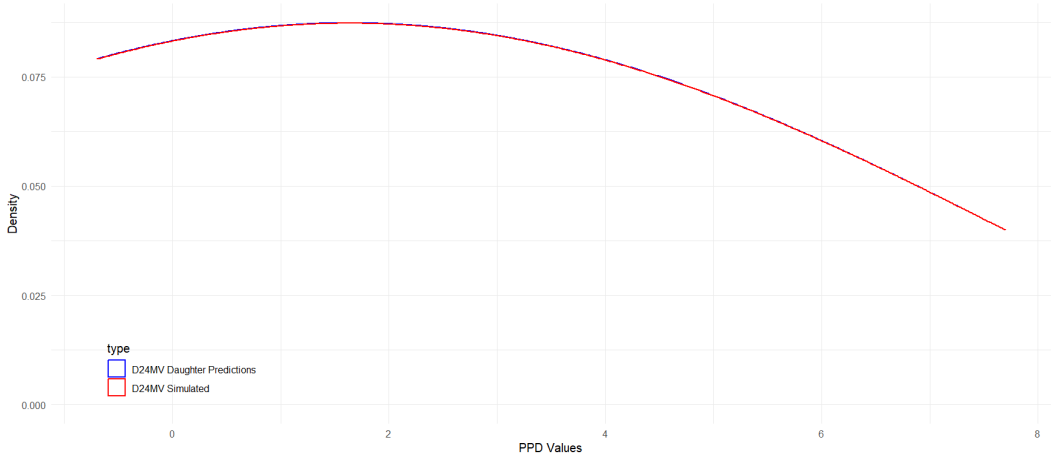
Site MV

Figure IV.41



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

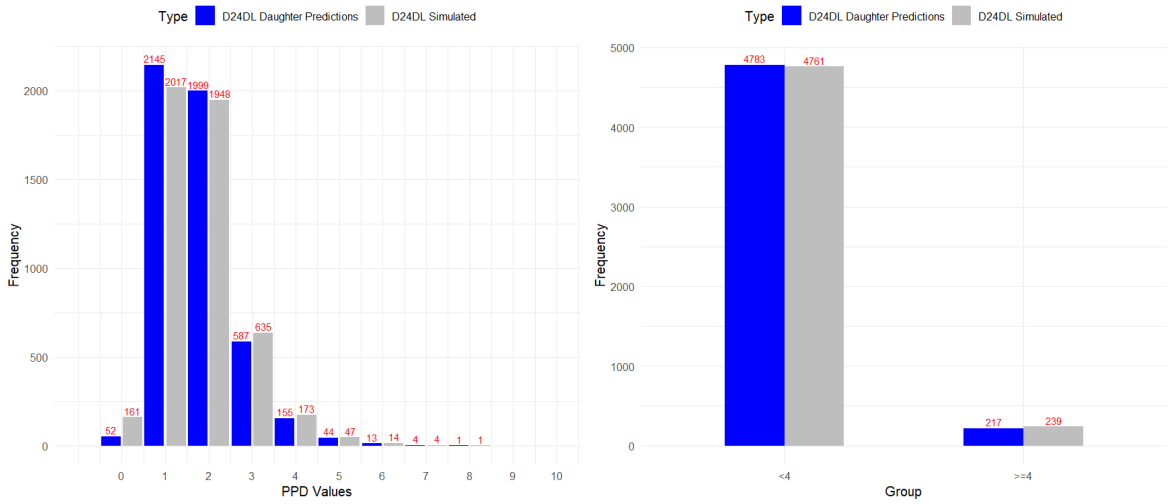
Figure IV.42



Optimal Kernel Density Plots of 24MV for Simulated and Daughter Predicted

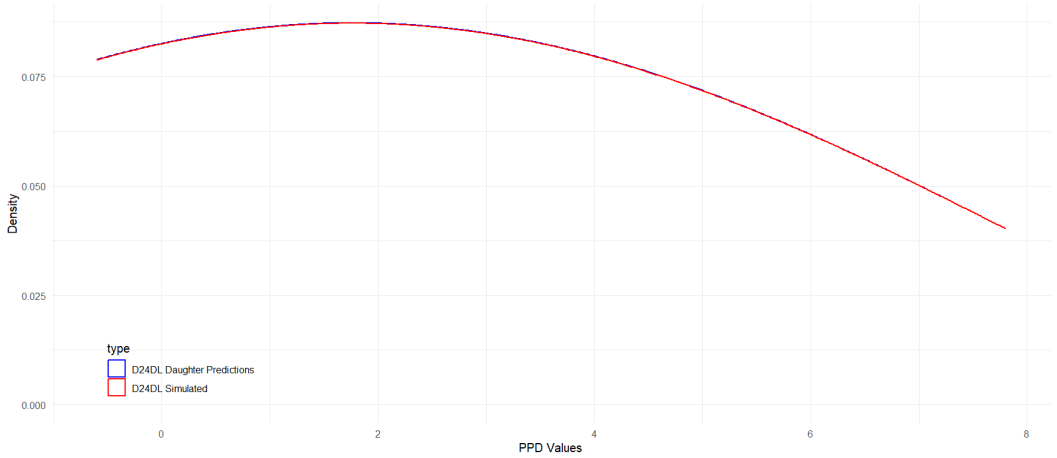
Site DL

Figure IV.43



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

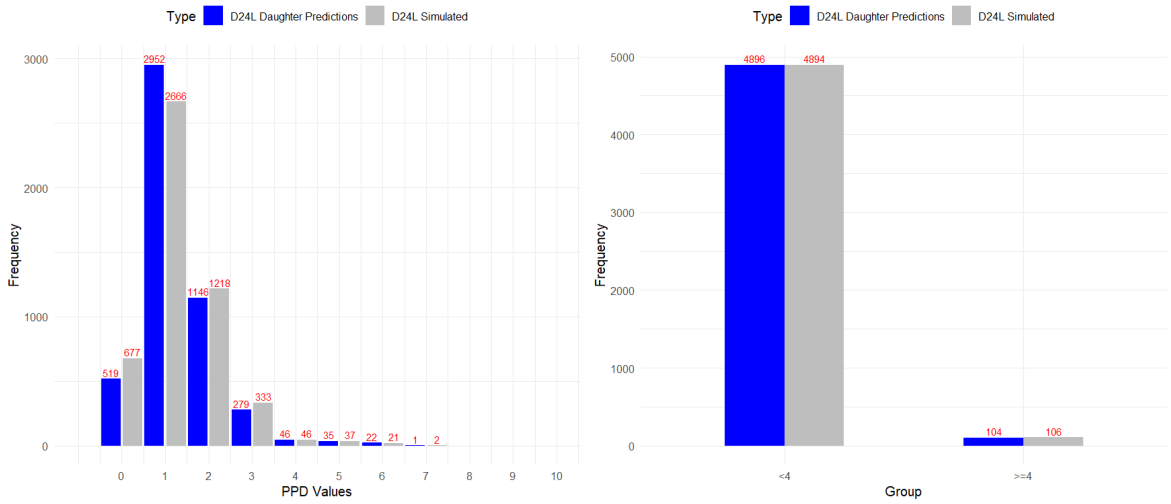
Figure IV.44



Optimal Kernel Density Plots of 24DL for Simulated and Daughter Predicted

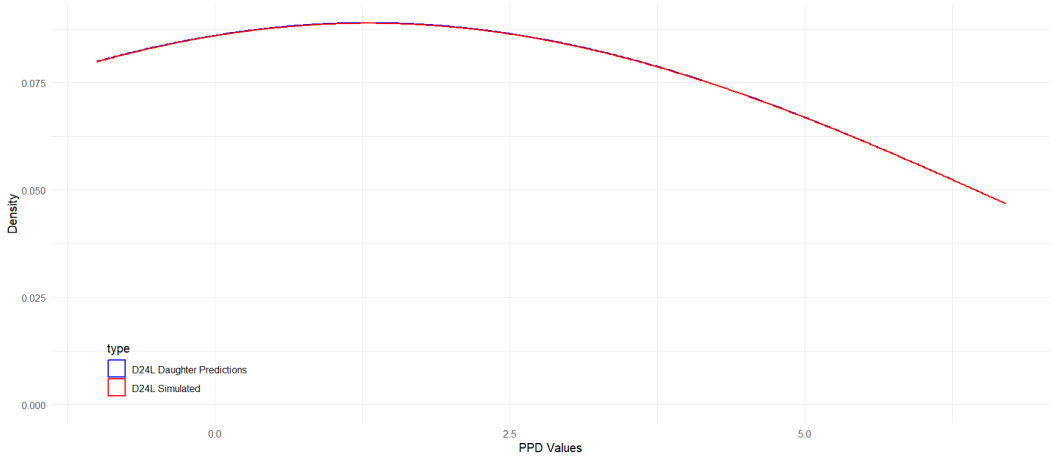
Site L

Figure IV.45



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

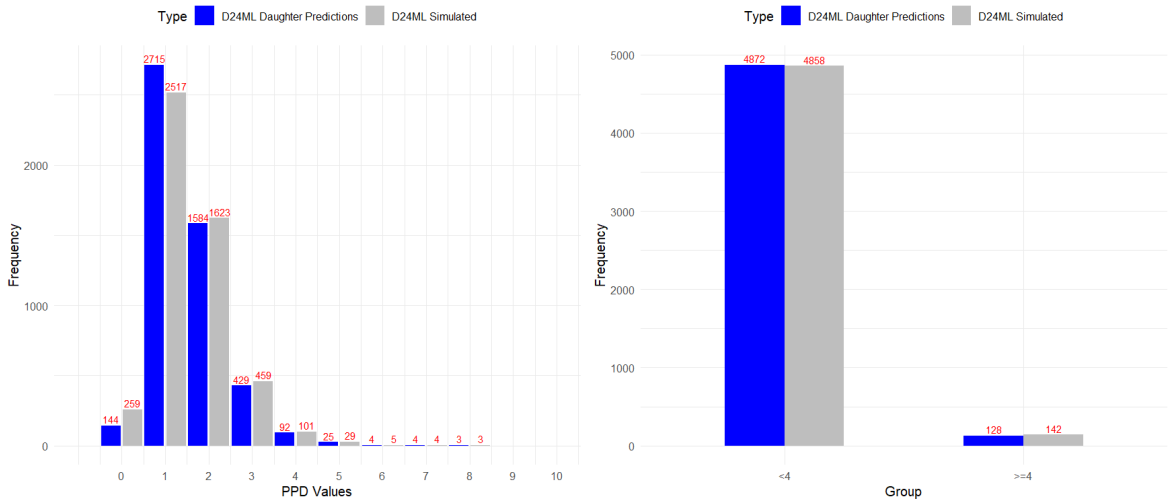
Figure IV.46



Optimal Kernel Density Plots of 24L for Simulated and Daughter Predicted

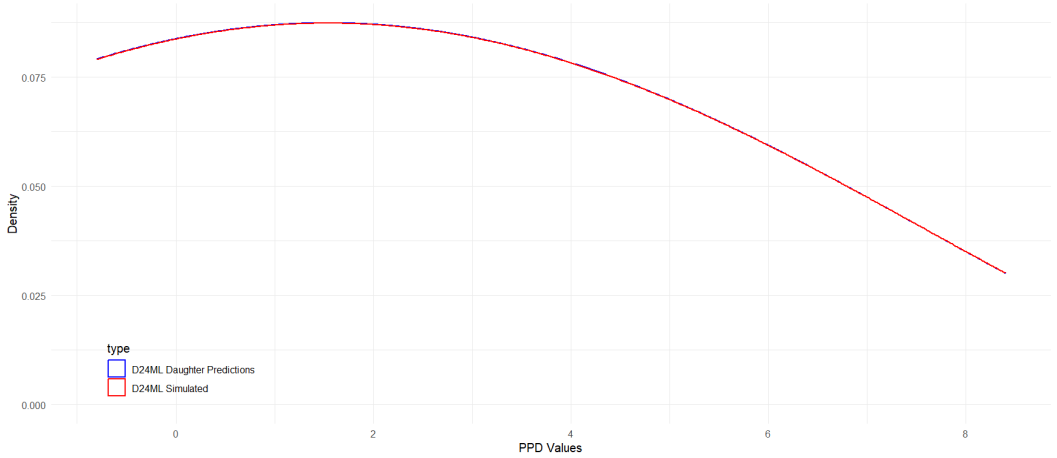
Site ML

Figure IV.47



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Figure IV.48



Optimal Kernel Density Plots of 24ML for Simulated and Daughter Predicted

IV.5 Upper Second Premolar

Analysis of M25 Mother Models: Characteristics, Performance, and Feature Importance

Table IV.41: Distinctive Characteristics of the M25 Mother Models by Site

Metric	M25DV	M25V	M25MV	M25DL	M25L	M25ML
Mother Model Size (Kb)	552.20	133.10	274.20	201.80	615.30	599.10
N.Iter.	551	123	263	185	591	579
Init. Train. RMSE	1.319	0.824	1.328	1.441	1.100	1.364
Final Train. RMSE	1.50e-02	3.54e-02	3.03e-02	2.97e-02	5.70e-03	8.10e-03
Features	γSM.DV 15DV	γSM.V 15V	γSM.MV 15MV	γSM.DL 15DL	γSM.L 15L	γSM.ML 15ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table IV.42: Performance Metrics for the M25 Mother Models by Site

Metric	M25DV	M25V	M25MV	M25DL	M25L	M25ML
RMSE	0.015	0.035	0.030	0.030	0.006	0.008
MAE	0.003	0.008	0.005	0.005	0.001	0.001
MSE	2.25e-04	1.25e-03	9.15e-04	8.84e-04	3.25e-05	6.56e-05
R^2	99.97%	99.79%	99.88%	99.90%	99.99%	99.99%

Table IV.43: Mother Models features Importance Metrics - M25 by Site

Model	Feature	Gain	Cover	Frequency
D25DV	γSM.DV	0.722	0.478	0.596
	Original 15DV	0.278	0.522	0.404
D25V	γSM.V	0.645	0.540	0.664
	Original 15V	0.355	0.460	0.336
D25MV	γSM.MV	0.697	0.460	0.613
	Original 15MV	0.303	0.540	0.387
D25DL	γSM.DL	0.673	0.543	0.644
	Original 15DL	0.327	0.457	0.356
D25L	γSM.L	0.636	0.563	0.685
	Original 15L	0.364	0.437	0.315
D25ML	γSM.ML	0.631	0.533	0.632
	Original 15ML	0.369	0.467	0.368

Analysis of D25 Daughter Models: Characteristics, Performance, and Feature Importance

Table IV.44: Distinctive Characteristics of the D25 Daughter Models by Site

Metric	D25DV	D25V	D25MV	D25DL	D25L	D25ML
Model Size (Mb)	32.4	32.4	32.4	32.3	25.8	32.2
N.Iter.	30000	30000	30000	30000	23978	30000
Init. Train. RMSE	1.427	1.442	1.442	1.585	1.225	1.481
Final Train. RMSE	0.158	0.152	0.152	0.170	0.168	0.152

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Table IV.45: Performance Metrics for the Daughter Models - D25 Data

Metric	**D25DV**	**D25V**	**D25MV**	**D25DL**	**D25L**	**D25ML**
RMSE	0.158	0.152	0.152	0.170	0.168	0.152
MAE	0.124	0.120	0.120	0.135	0.132	0.120
MSE	0.025	0.023	0.023	0.029	0.028	0.023
R^2	97.46%	96.50%	97.46%	97.53%	96.82%	97.28%

Table IV.46: Feature Importance Metrics Of Daughter Models D25 by Site

Model	**Feature**	**Gain**	**Cover**	**Frequency**
D25DV	Mother Predictions	0.428	0.278	0.302
	New Mother Predictions	0.295	0.126	0.090
	15DV	0.277	0.597	0.608
D25V	Mother Predictions	0.449	0.312	0.326
	New Mother Predictions	0.274	0.091	0.071
	15V	0.068	0.593	0.597
D25MV	Mother Predictions	0.449	0.312	0.326
	New Mother Predictions	0.274	0.091	0.071
	15MV	0.277	0.597	0.603
D25DL	Mother Predictions	0.418	0.290	0.316
	New Mother Predictions	0.317	0.126	0.092
	15DL	0.266	0.584	0.592
D25L	Mother Predictions	0.454	0.244	0.283
	New Mother Predictions	0.292	0.160	0.113
	15L	0.254	0.596	0.605
D25ML	Mother Predictions	0.427	0.264	0.306
	New Mother Predictions	0.301	0.145	0.107
	15ML	0.272	0.591	0.587

Daughter Model M25 Predictions

Table IV.47: Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted for Site 25

Groups	Data	**25DV**	**25V**	**25MV**	**25DL**	**25L**	**25ML**
$\geq$ 4 mm	Simulated	235	37	201	249	124	212
	Predicted	226	35	180	224	112	187
χ^2 statistic		0.1455	0.014’	1.0915	1.2782	0.5251	1.5036
$p - value$		0.703	0.906	0.296	0.258	0.469	0.220

Table IV.48: Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site

Site	Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
25DV	0	188	323	35.665	**2.34e-09**
	1	2409	2225	7.306	**6.87e-03**
	2	1775	1725	0.714	0.398
	3	402	492	9.060	**2.61e-03**
	4	173	173	0	1
	5	35	42	0.636	0.425
	6	11	13	0.167	0.683
	8	6	6	0	1
	9	1	1	0	1
25V	0	1001	1107	5.330	**0.021**
	1	3136	2909	8.524	**3.5e-03**
	2	726	819	5.598	**0.018**
	3	102	128	2.939	0.086
	4	24	26	0.080	0.777
	5	4	3	0.143	0.705
	6	5	6	0.091	0.763
	10	2	2	0	1
25MV	0	133	256	38.892	**4.48e-10**
	1	2218	2075	4.763	**0.029**
	2	2039	1964	1.405	0.236
	3	430	504	5.863	**0.015**
	4	116	136	1.587	0.208
	5	33	33	0	1
	6	22	23	0.022	0.881
	7	7	7	0	1
	8	2	2	0	1
25DL	0	101	269	76.281	**2.46e-18**
	1	2217	2001	11.061	**8.80e-04**
	2	1846	1798	0.632	0.427
	3	612	683	3.893	**0.048**
	4	118	141	2.042	0.153
	5	56	58	0.035	0.851
	6	19	18	0.027	0.869
	7	12	12	0	1
	8	17	18	0.029	0.866
	9	2	2	0	1
25L	0	496	642	18.731	**1.51e-05**
	1	2708	2440	13.952	**1.90e-04**
	2	1381	1426	0.721	0.396
	3	303	368	6.297	**0.012**
	4	75	88	1.037	0.309
	5	25	22	0.191	0.662
	6	7	8	0.067	0.796
	7	5	6	0.091	0.763
25ML	0	105	206	32.801	**1.02e-08**
	1	1979	1867	3.262	0.071
	2	2195	2118	1.375	0.241
	3	534	597	3.509	0.061
	4	149	170	1.382	0.240
	5	21	26	0.532	0.466
	6	12	10	0.182	0.670
	7	2	3	0.200	0.655
	8	3	3	0	1

Comparation of Distributions Trough Kernels

Table IV.49: Kernel Density Estimates Differences and Kolmogorov-Smirnov Test

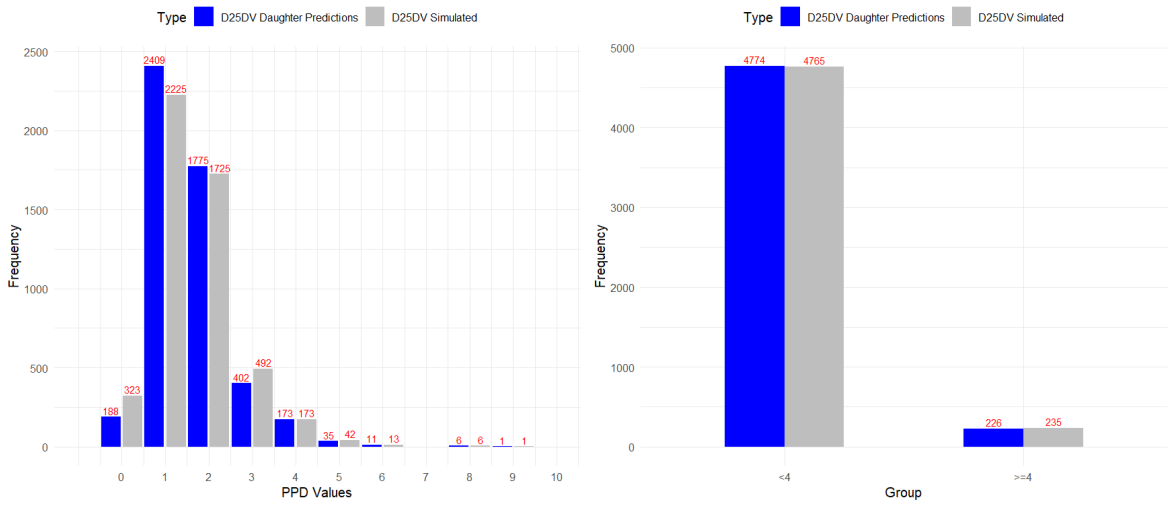
Site	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
D25DV	Cosine	2	2	1.71e-03	[8.86e-04, 3.95e-03]	0.0532	1.43e-06
D25V	Biweight	2	2	5.57e-03	[3.31e-04, 8.67e-03]	0.1868	0
D25MV	Biweight	2	2	1.88e-03	[1.32e-04, 5.17e-03]	0.0496	9.10e-06
D25DL	Cosine	2	2	3.37e-03	[1.96e-03, 5.91e-03]	0.048	1.99e-05
D25L	Biweight	2	2	1.77e-03	[4.55e-05, 4.72e-03]	0.0696	6.05e-11
D25ML	Cosine	2	2	2.40e-03	[9.78e-04, 6.59e-03]	0.0466	3.85e-05

Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

Visual Comparisons of Proportions and Distributions

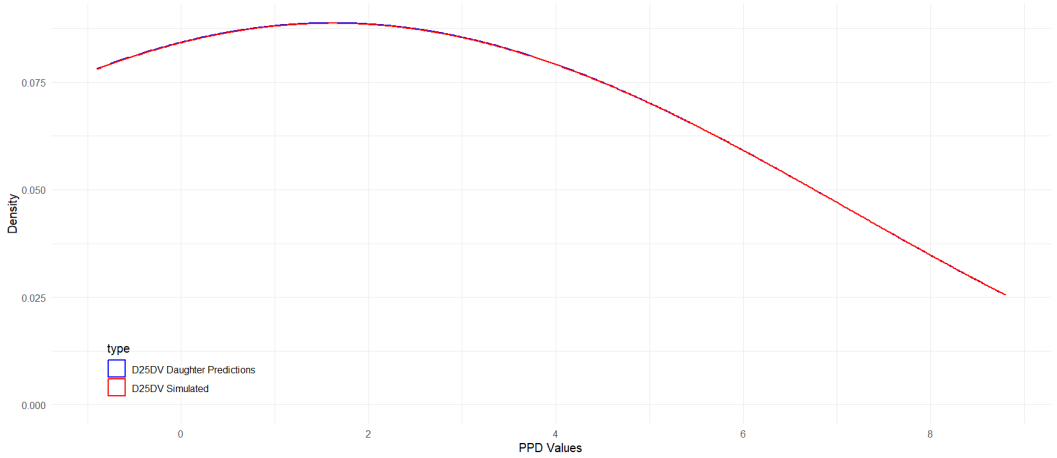
Site DV

Figure IV.49



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

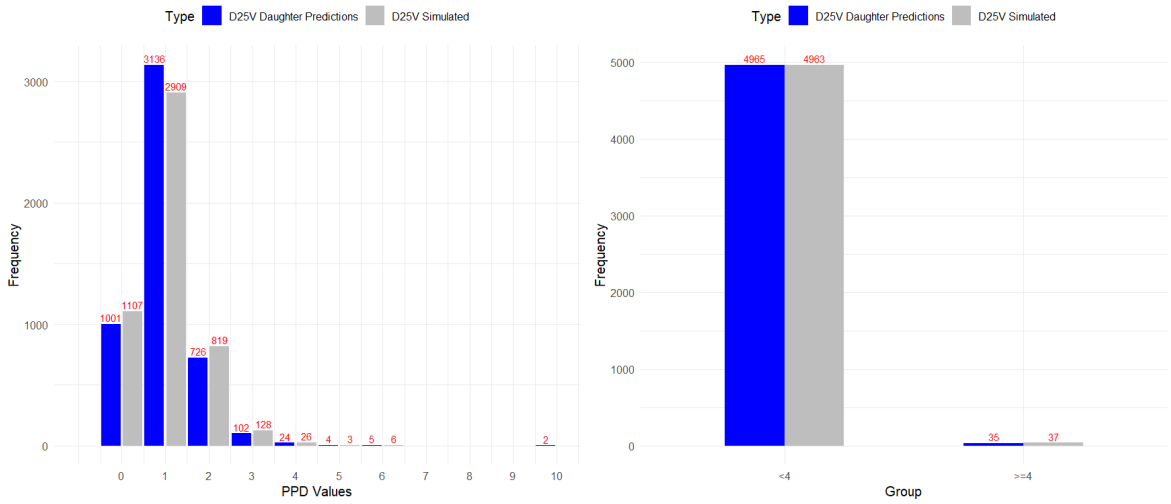
Figure IV.50



Optimal Kernel Density Plots of 25DV for Simulated and Daughter Predicted

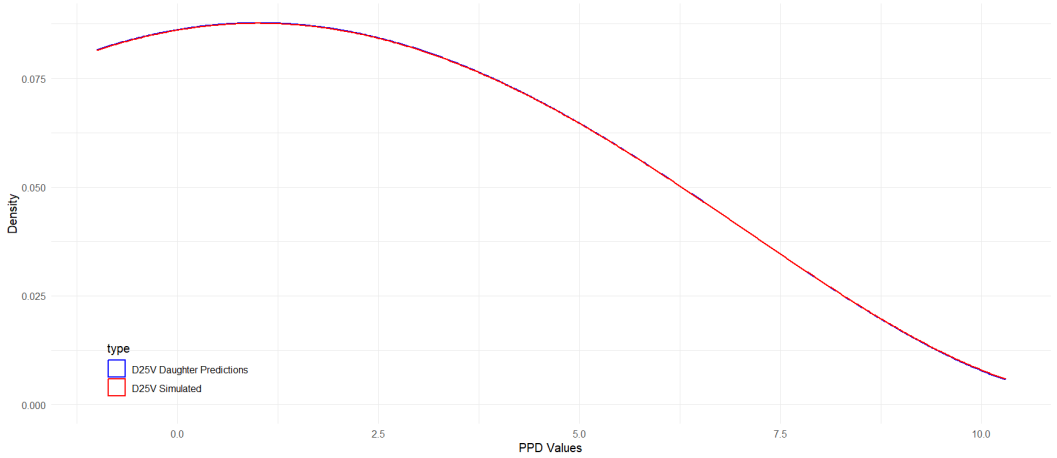
Site V

Figure IV.51



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

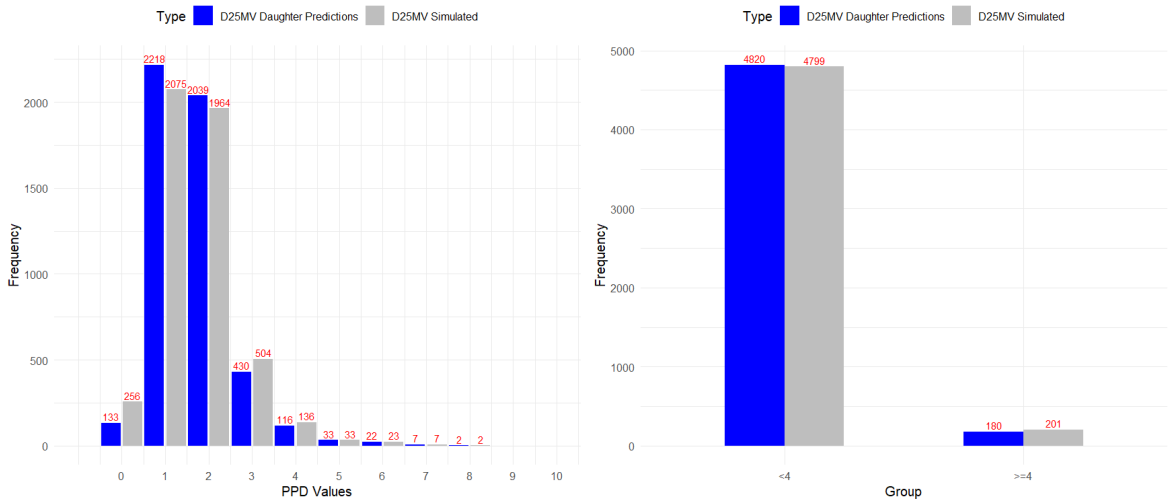
Figure IV.52



Optimal Kernel Density Plots of 25V for Simulated and Daughter Predicted

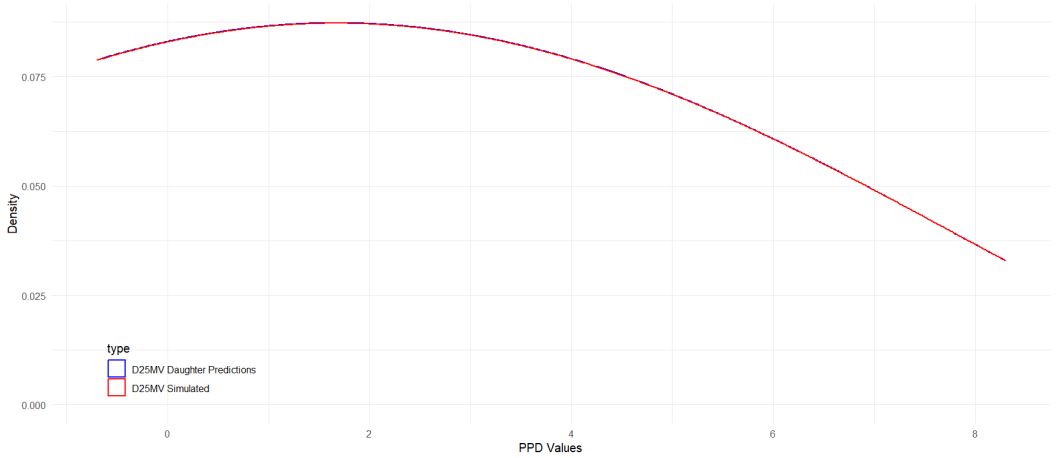
Site MV

Figure IV.53



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

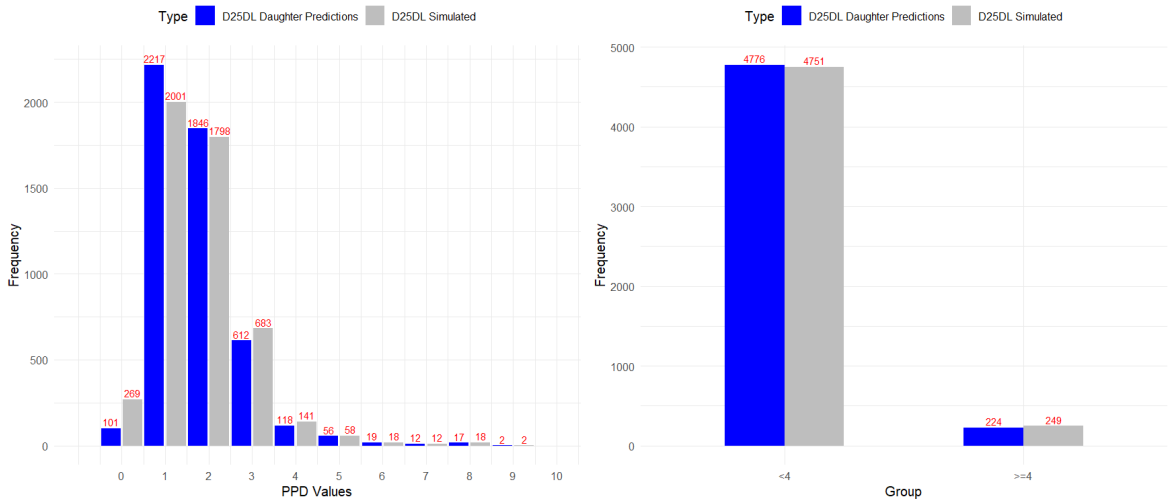
Figure IV.54



Optimal Kernel Density Plots of 25MV for Simulated and Daughter Predicted

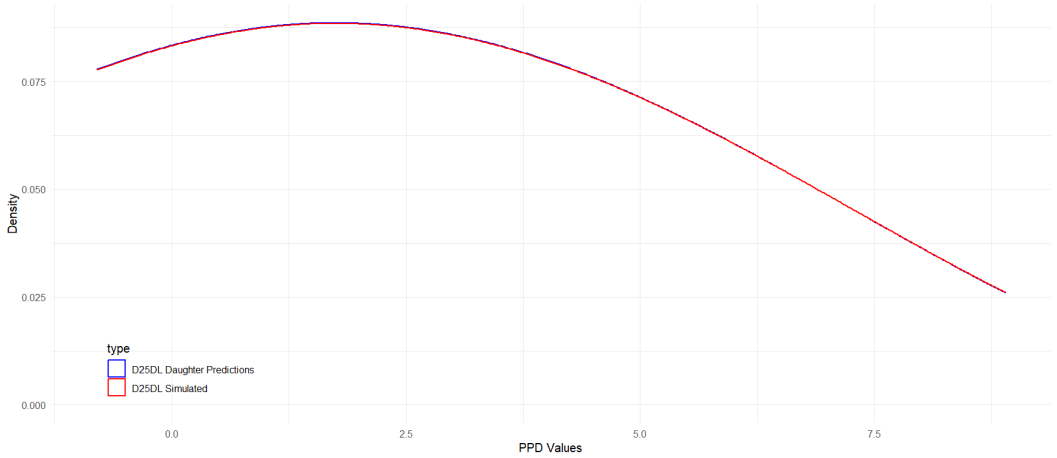
Site DL

Figure IV.55



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

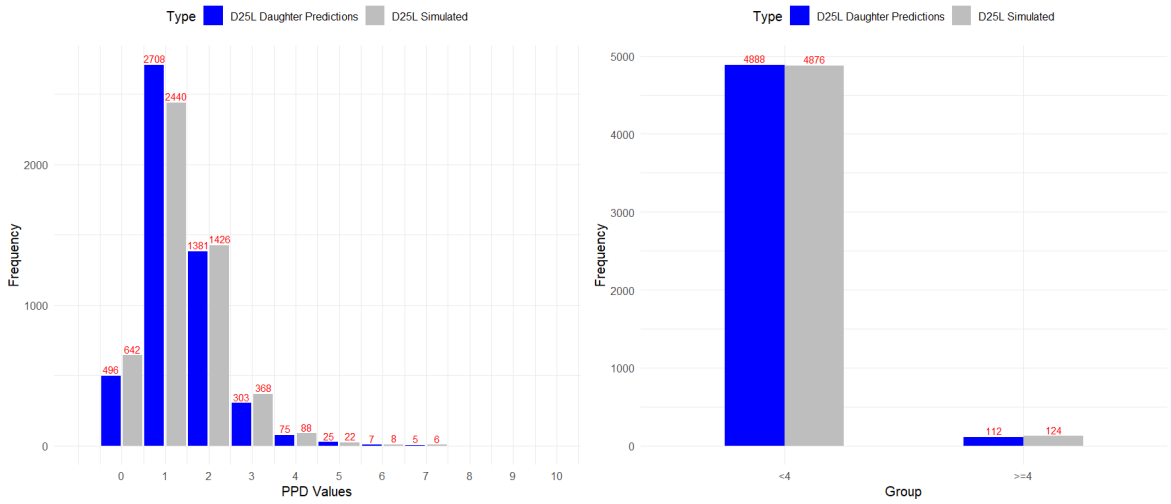
Figure IV.56



Optimal Kernel Density Plots of 25DL for Simulated and Daughter Predicted

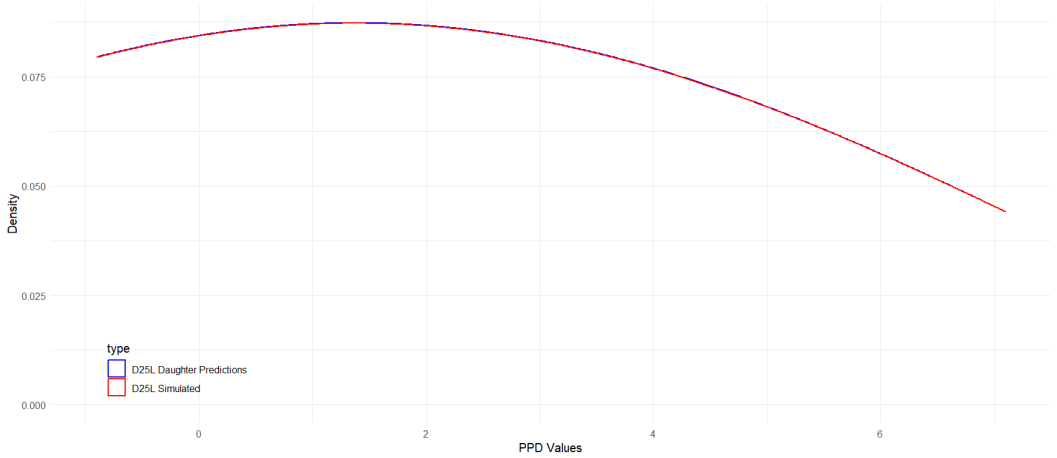
Site L

Figure IV.57



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

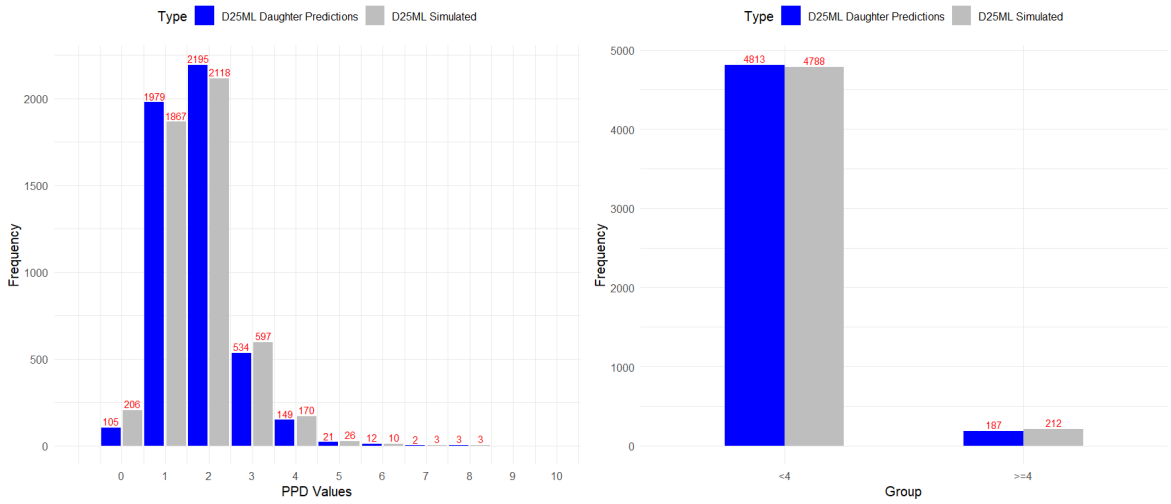
Figure IV.58



Optimal Kernel Density Plots of 25L for Simulated and Daughter Predicted

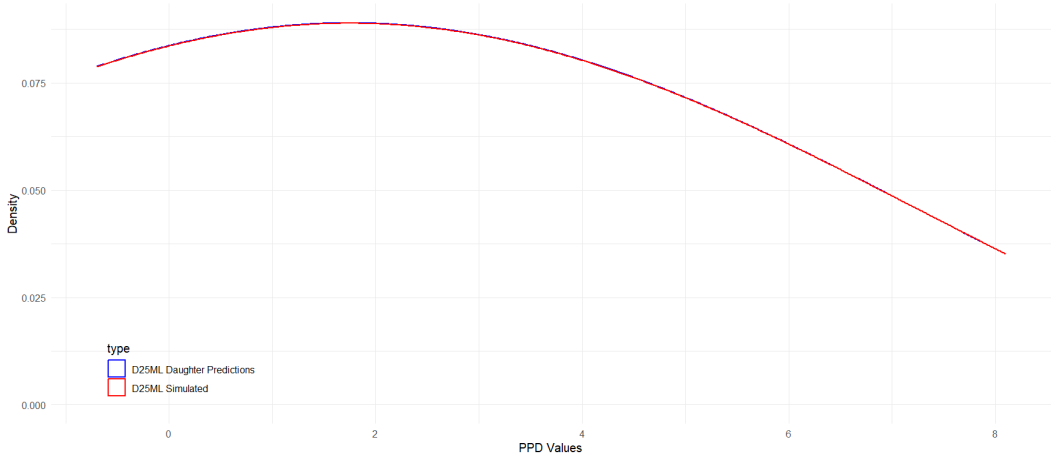
Site ML

Figure IV.59



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Figure IV.60



Optimal Kernel Density Plots of 25ML for Simulated and Daughter Predicted

IV.6 Upper First Molar

Analysis of M26 Mother Models: Characteristics, Performance, and Feature Importance

Table IV.50: Distinctive Characteristics of the M26 Mother Models by Site

Metric	M26DV	M26V	M26MV	M26DL	M26L	M26ML
Mother Model Size (Kb)	337.40	201.70	488.30	415.30	401.70	266.90
N.Iter.	306	191	472	412	383	252
Init. Train. RMSE	1.643	0.967	1.396	1.652	1.068	1.439
Final Train. RMSE	2.83e-02	3.17e-02	2.42e-02	2.07e-02	1.15e-02	2.40e-02
Features	γSM.DV 16DV	γSM.V 16V	γSM.MV 16MV	γSM.DL 16DL	γSM.L 16L	γSM.ML 16ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table IV.51: Performance Metrics for the M26 Mother Models by Site

Metric	M26DV	M26V	M26MV	M26DL	M26L	M26ML
RMSE	0.028	0.032	0.024	0.021	0.011	0.024
MAE	0.005	0.006	0.002	0.004	0.001	0.006
MSE	8.03e-04	1.01e-03	5.79e-04	4.22e-04	1.31e-04	5.75e-04
R^2	99.94%	99.84%	99.93%	99.96%	99.98%	99.93%

Table IV.52: Mother Models features Importance Metrics - M26 by Site

Model	Feature	Gain	Cover	Frequency
M26DV	γSM.DV	0.680	0.476	0.653
	Original 16DV	0.320	0.524	0.347
M26V	γSM.V	0.662	0.461	0.592
	Original 16V	0.338	0.539	0.408
M26MV	γSM.MV	0.709	0.541	0.650
	Original 16MV	0.291	0.459	0.350
M26DL	γSM.DL	0.668	0.492	0.567
	Original 16DL	0.332	0.508	0.433
M26L	γSM.L	0.700	0.486	0.635
	Original 16L	0.300	0.514	0.365
M26ML	γSM.ML	0.640	0.532	0.641
	Original 16ML	0.360	0.468	0.359

Analysis of D26 Daughter Models: Characteristics, Performance, and Feature Importance

Table IV.53: Distinctive Characteristics of the D26 Daughter Models by Site

Metric	D26DV	D26V	D26MV	D26DL	D26L	D26ML
Model Size (Mb)	2.3	32.0	32.4	31.6	32.3	25.9
N.Iter.	2135	30000	30000	30000	30000	23978
Init. Train. RMSE	1.798	1.060	1.487	1.778	1.177	1.542
Final Train. RMSE	0.397	0.141	0.150	0.201	0.139	0.174

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Table IV.54: Performance Metrics for the Daughter Models - D26 Data

Metric	**D26DV**	**D26V**	**D26MV**	**D26DL**	**D26L**	**D26ML**
RMSE	0.397	0.141	0.150	0.201	0.139	0.174
MAE	0.315	0.111	0.119	0.158	0.109	0.137
MSE	0.158	0.020	0.023	0.040	0.019	0.030
R^2	89.68%	97.40%	97.60%	96.91%	97.42%	96.92%

Table IV.55: Feature Importance Metrics Of Daughter Models D26 by Site

Model	**Feature**	**Gain**	**Cover**	**Frequency**
D26DV	Mother Predictions	0.570	0.329	0.340
	New Mother Predictions	0.364	0.129	0.111
	16DV	0.066	0.542	0.549
D26V	Mother Predictions	0.443	0.299	0.305
	New Mother Predictions	0.279	0.096	0.074
	16V	0.278	0.605	0.620
D26MV	Mother Predictions	0.448	0.337	0.338
	New Mother Predictions	0.277	0.072	0.064
	16MV	0.275	0.591	0.598
D26DL	Mother Predictions	0.447	0.291	0.336
	New Mother Predictions	0.286	0.129	0.104
	16DL	0.267	0.580	0.560
D26L	Mother Predictions	0.446	0.321	0.323
	New Mother Predictions	0.275	0.084	0.064
	16L	0.280	0.595	0.613
D26ML	Mother Predictions	0.464	0.343	0.344
	New Mother Predictions	0.282	0.073	0.065
	16ML	0.254	0.583	0.591

Daughter Model D26 Predictions

Table IV.56: Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted for Site

Groups	Data	**26DV**	**26V**	**26MV**	**26DL**	**26L**	**26ML**
$\geq$ 4 mm	Simulated	496	68	249	515	114	247
	Predicted	458	64	226	477	108	223
χ^2 statistic		1.5863	0.0691	1.0698	1.5320	0.1152	1.1810
$p - value$		0.208	0.793	0.301	0.216	0.734	0.277

Table IV.57: Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site

Site	Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
26DV	0	152	400	111.420	**4.79e-26**
	1	1843	1577	20.689	**5.40e-06**
	2	1971	1788	8.909	**2.80e-03**
	3	576	739	20.205	**6.96e-06**
	4	304	315	0.195	0.658
	5	89	106	1.482	0.223
	6	34	45	1.532	0.216
	7	19	17	0.111	0.739
	8	10	9	0.053	0.819
	10	2	2	0	1
26V	0	684	794	8.187	**4.20e-03**
	1	2963	2754	7.641	**5.70e-03**
	2	1128	1188	1.554	0.212
	3	161	196	3.431	0.064
	4	34	36	0.057	0.811
	5	18	20	0.105	0.746
	6	1	1	0	1
	7	9	9	0	1
	8	2	2	0	1
26MV	0	159	268	27.824	**1.33e-07**
	1	2111	1976	4.459	**0.035**
	2	1942	1877	1.106	0.293
	3	562	630	3.879	**0.049**
	4	169	188	1.011	0.315
	5	50	53	0.087	0.768
	6	3	4	0.143	0.705
	7	4	4	0	1
26DL	0	118	256	50.920	**9.62e-13**
	1	1673	1571	3.207	0.073
	2	2045	1887	6.349	**0.012**
	3	687	771	4.840	**0.028**
	4	313	336	0.815	0.367
	5	123	138	0.862	0.353
	6	34	33	0.015	0.903
	7	6	7	0.077	0.782
	8	1	1	0	1
26L	0	365	480	15.651	**7.62e-05**
	1	2803	2625	5.837	**0.016**
	2	1474	1479	0.008	0.927
	3	250	302	4.899	**0.027**
	4	84	87	0.053	0.819
	5	19	22	0.220	0.639
	6	2	2	0	1
	7	3	3	0	1
26ML	0	117	240	42.378	**7.52e-11**
	1	2027	1884	5.229	**0.022**
	2	1976	1898	1.570	0.210
	3	657	731	3.945	**0.047**
	4	160	181	1.293	0.255
	5	41	43	0.048	0.827
	6	15	16	0.032	0.857
	7	5	5	0	1
	8	2	2	0	1

Comparation of Distributions Trough Kernels

Table IV.58: Kernel Density Estimates Differences and Kolmogorov-Smirnov Test

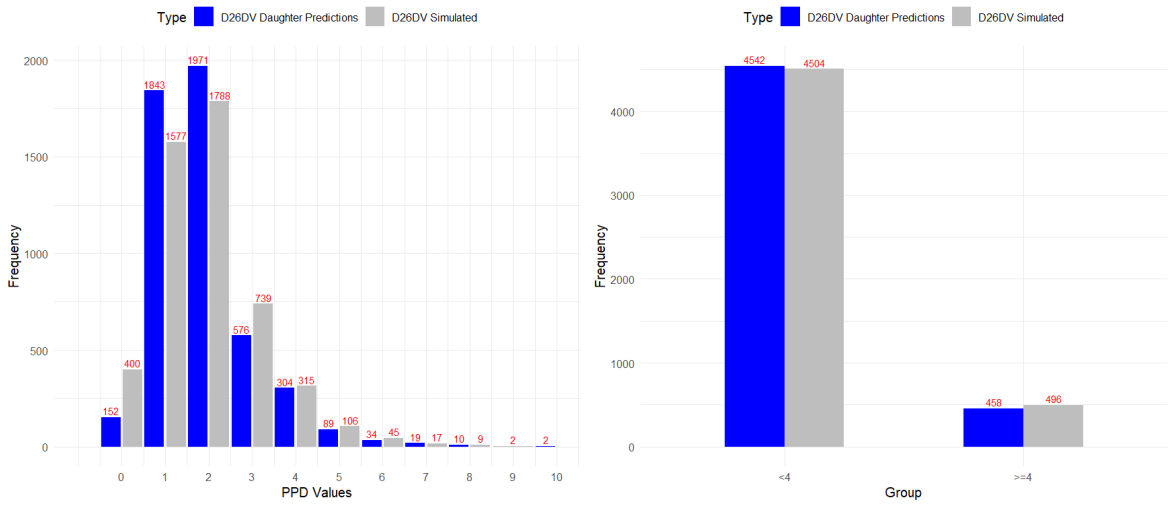
26	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
M26DV	Cosine	2	2	2.81e-03	[8.97e-04, 5.27e-03]	0.112	0
M26V	Biweight	2	2	1.53e-03	[4.64e-04, 3.12e-03]	0.068	2.09e-10
M26MV	Biweight	2	2	4.95e-03	[2.93e-03, 8.41e-03]	0.042	2.72e-04
M26DL	Cosine	2	2	3.67e-03	[1.49e-03, 4.25e-03]	0.047	3.19e-05
M26L	Biweight	2	2	3.44e-03	[3.56e-05, 5.49e-03]	0.059	5.52e-08
M26ML	Cosine	2	2	1.91e-03	[1.29e-03, 3.99e-03]	0.046	5.57e-05

Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

Visual Comparisons of Proportions and Distributions

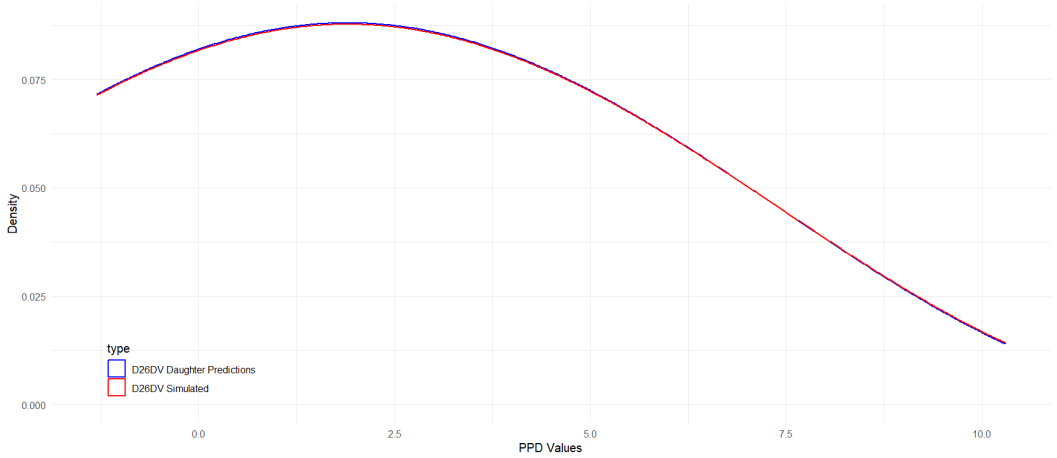
Site DV

Figure IV.61



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

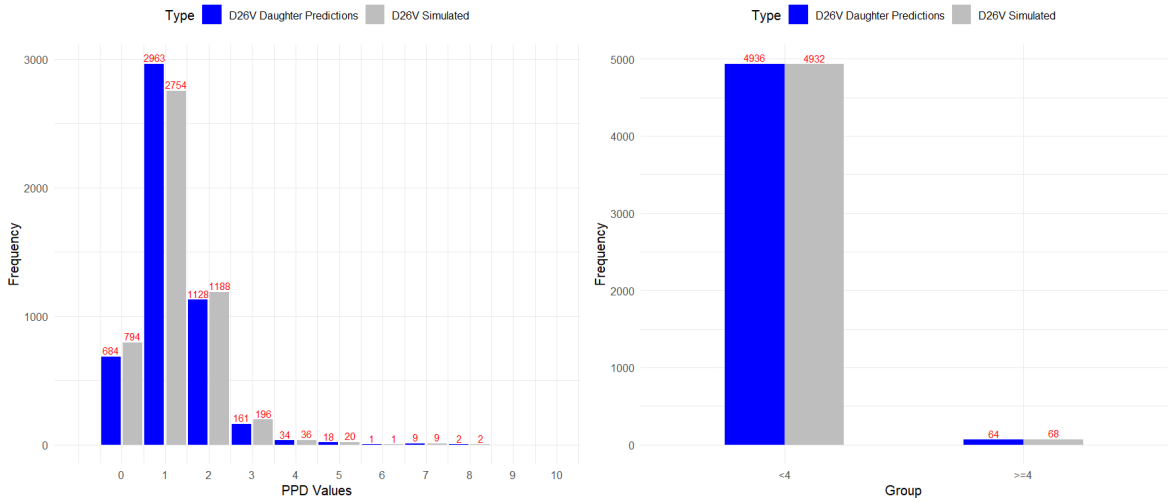
Figure IV.62



Optimal Kernel Density Plots of 26DV for Simulated and Daughter Predicted

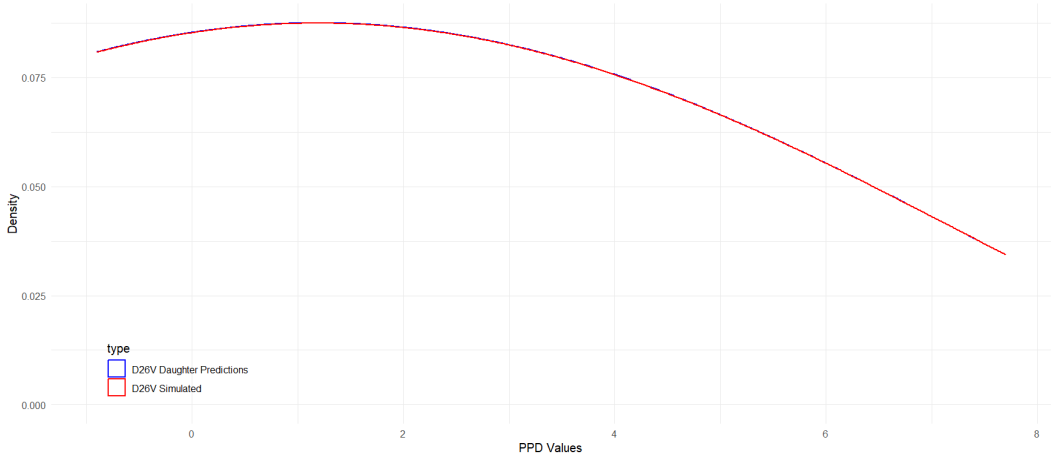
Site V

Figure IV.63



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

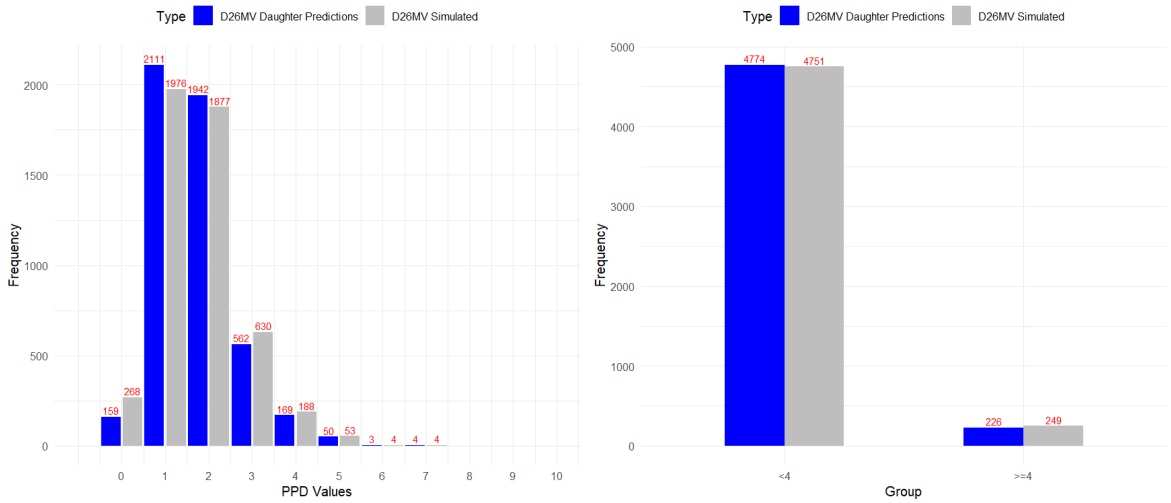
Figure IV.64



Optimal Kernel Density Plots of 26V for Simulated and Daughter Predicted

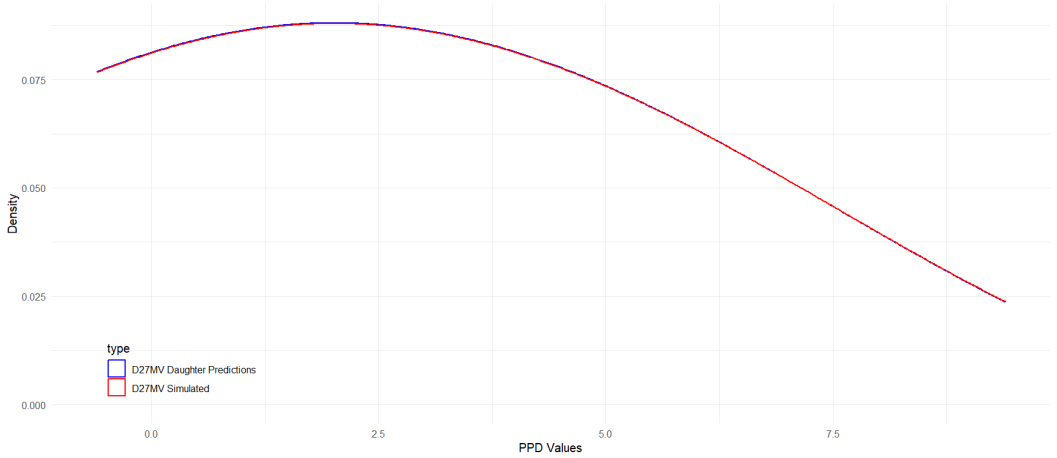
Site MV

Figure IV.65



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

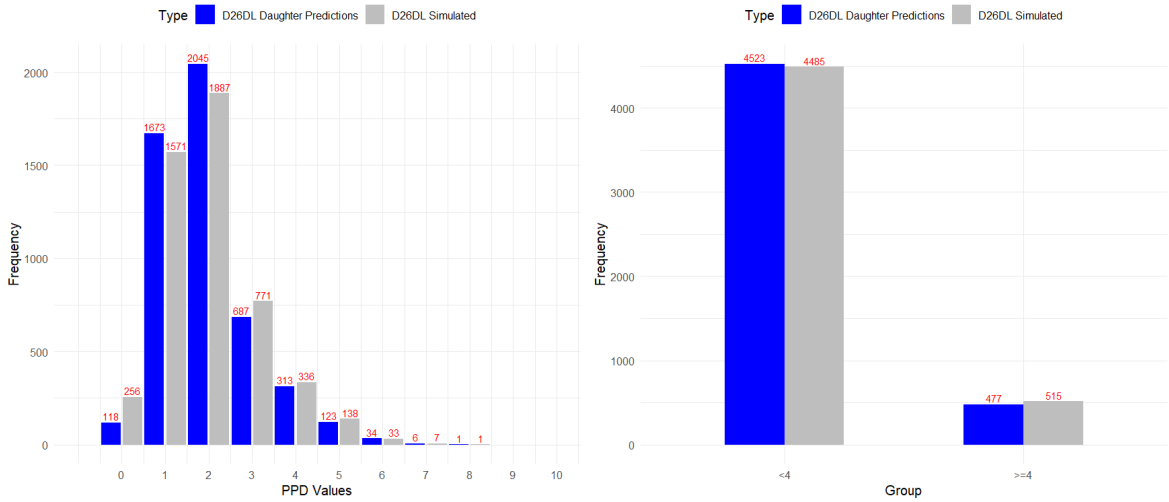
Figure IV.66



Optimal Kernel Density Plots of 26MV for Simulated and Daughter Predicted

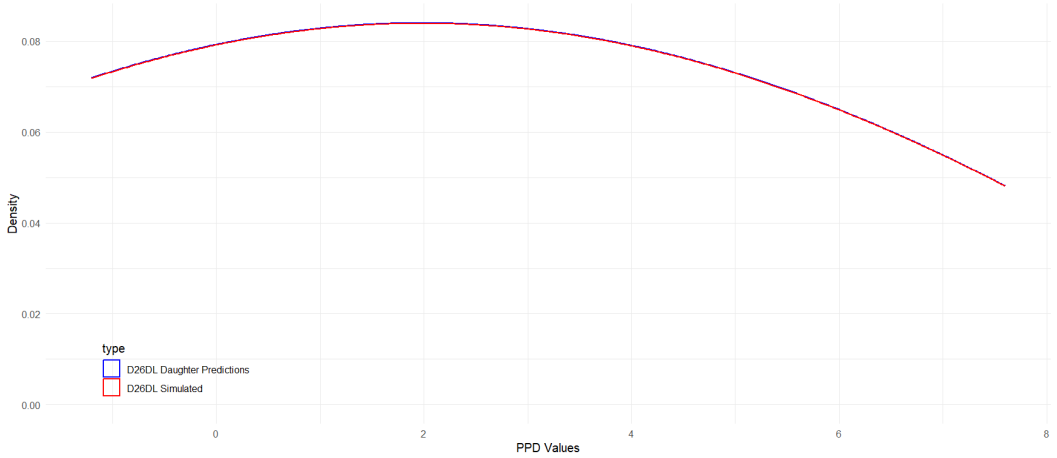
Site DL

Figure IV.67



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

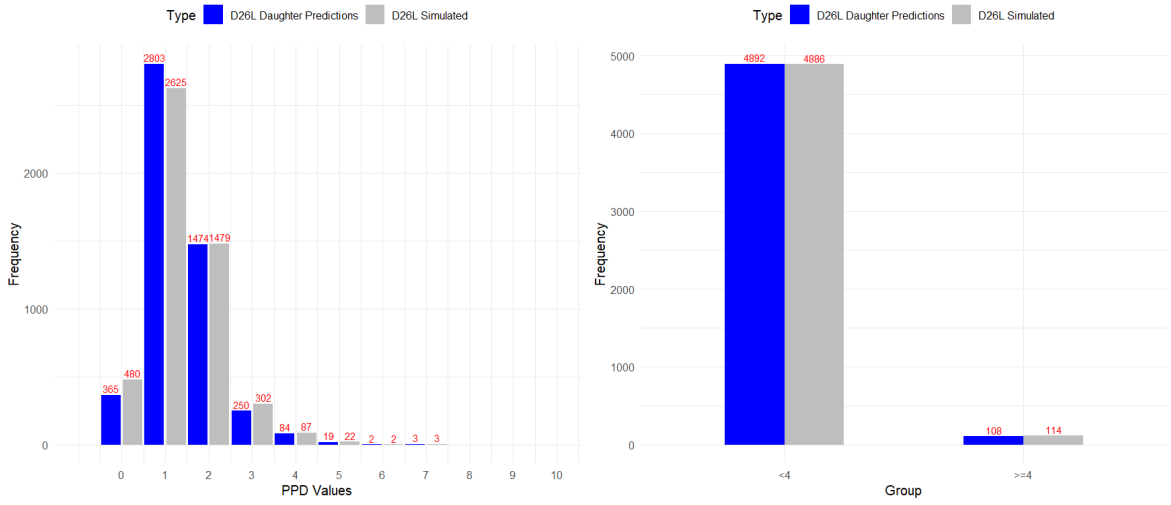
Figure IV.68



Optimal Kernel Density Plots of 26DL for Simulated and Daughter Predicted

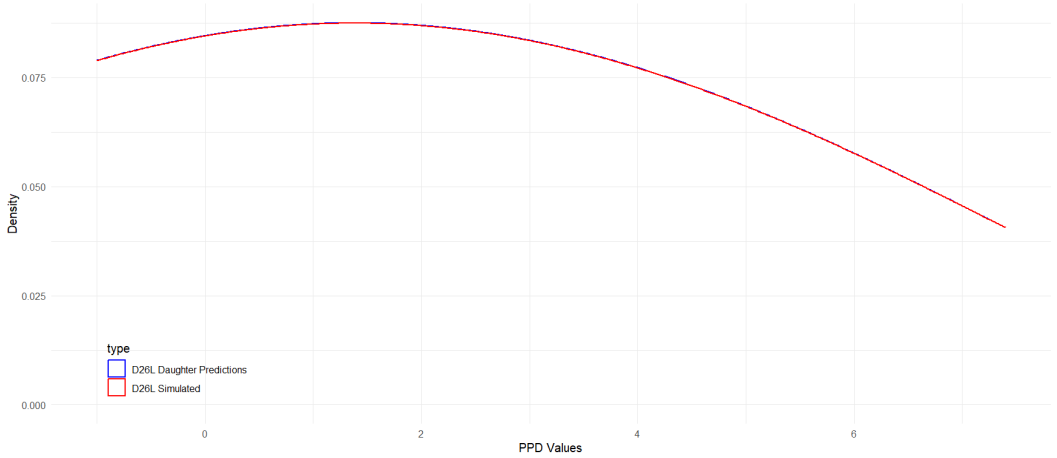
Site L

Figure IV.69



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

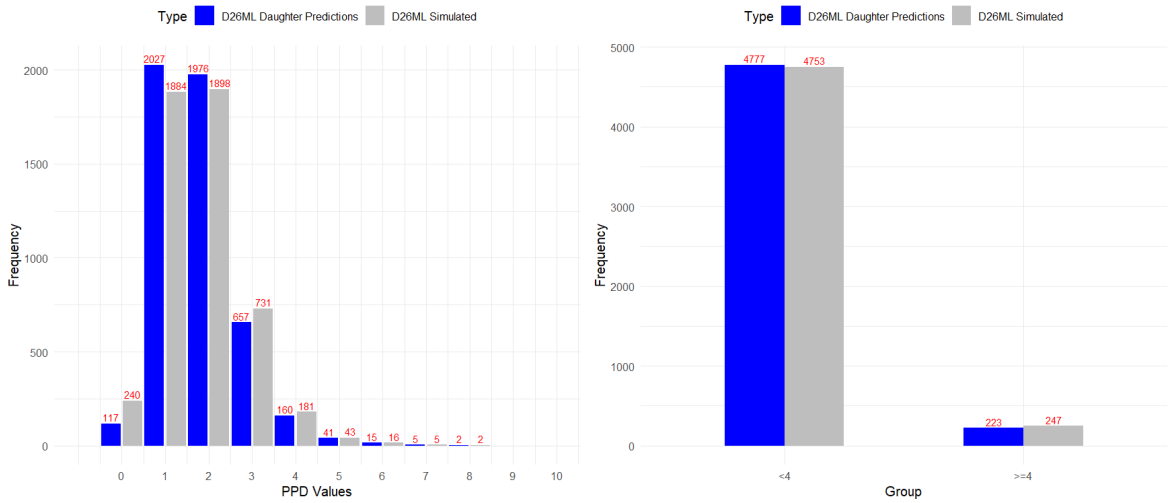
Figure IV.70



Optimal Kernel Density Plots of 26L for Simulated and Daughter Predicted

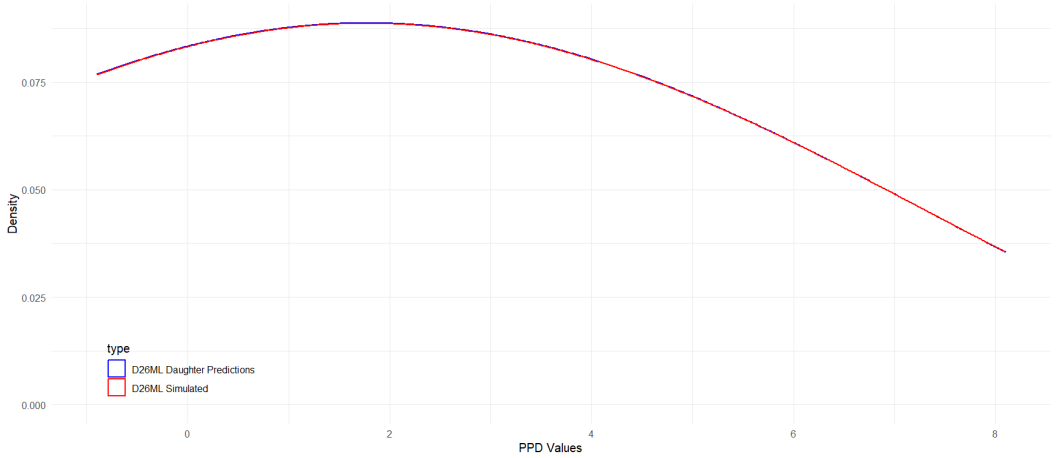
Site ML

Figure IV.71



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Figure IV.72



Optimal Kernel Density Plots of 26ML for Simulated and Daughter Predicted

IV.7 Upper Second Molar

Analysis of M27 Mother Models: Characteristics, Performance, and Feature Importance

Table IV.59: Distinctive Characteristics of the M27 Mother Models by Site

Metric	M27DV	M27V	M27MV	M27DL	M27L	M27ML
Mother Model Size (Kb)	342.90	233.80	281.90	198.30	198.40	709.40
N.Iter.	351	220	271	185	181	672
Init. Train. RMSE	1.503	1.172	1.729	1.699	1.292	1.706
Final Train. RMSE	4.45e-02	4.67e-02	4.37e-02	4.47e-02	4.08e-02	9.68e-03
Features	γSM.DV 17DV	γSM.V 17V	γSM.MV 17MV	γSM.DL 17DL	γSM.L 17L	γSM.ML 17ML

Abbreviations: NIter – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE; γSM.Site – Directional Symmetry Measure computed from original data; 11Site – Original NHANES 2011/2012 11 sites

Table IV.60: Performance Metrics for the M27 Mother Models by Site

Metric	M27DV	M27V	M27MV	M27DL	M27L	M27ML
RMSE	0.044	0.047	0.044	0.045	0.041	0.010
MAE	0.007	0.005	0.007	0.010	0.005	0.002
MSE	1.98e-03	2.18e-03	1.91e-03	1.99e-03	1.66e-03	9.36e-05
R^2	99.81%	99.72%	99.85%	99.83%	99.77%	99.99%

Table IV.61: Mother Models features Importance Metrics - M27 by Site

Model	Feature	Gain	Cover	Frequency
D27DV	γSM.DV	0.700	0.519	0.599
	Original 17DV	0.300	0.481	0.401
D27V	γSM.V	0.681	0.572	0.672
	Original 17V	0.319	0.428	0.328
D27MV	γSM.MV	0.693	0.514	0.636
	Original 17MV	0.307	0.486	0.364
D27DL	γSM.DL	0.679	0.580	0.650
	Original 17DL	0.321	0.420	0.350
D27L	γSM.L	0.675	0.526	0.681
	Original 17L	0.325	0.474	0.319
D27ML	γSM.ML	0.672	0.650	0.713
	Original 17ML	0.328	0.350	0.287

Analysis of D27 Daughter Models: Characteristics, Performance, and Feature Importance

Table IV.62: Distinctive Characteristics of the D27 Daughter Models by Site

Metric	D27DV	D27V	D27MV	D27DL	D27L	D27ML
Model Size (Mb)	**6.6**	32.3	32.3	22.8	32.4	22.8
N.Iter.	6287	30000	30000	21060	30000	21060
Init. Train. RMSE	1.599	1.263	1.851	1.845	1.401	1.845
Final Train. RMSE	0.306	0.147	0.185	0.221	0.148	0.221

Abbreviations: N.Iter. – Number of Iterations; Init. Train. RMSE – Initial Training RMSE; Final Train. RMSE – Final Training RMSE

Table IV.63: Performance Metrics for the Daughter Models - D27 Data

Metric	D27DV	D27V	D27MV	D27DL	D27L	D27ML
RMSE	0.306	0.147	0.185	0.221	0.148	0.172
MAE	0.242	0.116	0.146	0.174	0.117	0.136
MSE	0.094	0.022	0.034	0.049	0.022	0.030
R^2	91.79%	97.53%	97.63%	96.64%	97.47%	97.65%

Table IV.64: Feature Importance Metrics Of Daughter Models D27 by Site

Model	Feature	Gain	Cover	Frequency
D27DV	Mother Predictions	0.534	0.357	0.375
	New Mother Predictions	0.332	0.068	0.062
	17DV	0.133	0.574	0.563
D27V	Mother Predictions	0.433	0.253	0.290
	New Mother Predictions	0.293	0.152	0.105
	17V	0.274	0.595	0.604
D27MV	Mother Predictions	0.452	0.319	0.327
	New Mother Predictions	0.278	0.103	0.084
	17MV	0.270	0.578	0.590
D27DL	Mother Predictions	0.467	0.275	0.308
	New Mother Predictions	0.296	0.151	0.112
	17DL	0.237	0.574	0.579
D27L	Mother Predictions	0.443	0.289	0.309
	17L	0.279	0.590	0.602
	New Mother Predictions	0.278	0.121	0.088
D27ML	Mother Predictions	0.452	0.323	0.331
	New Mother Predictions	0.281	0.103	0.083
	17ML	0.268	0.574	0.585

Daughter Model D27 Predictions

Table IV.65: Chi-squared Test Results for proportions of PPD values $\geq$ 4mm comparisons between Original and predicted by Site 27

Groups	Data	**27DV**	**27V**	**27MV**	**27DL**	**27L**	**27ML**
$\geq$ 4 mm	Simulated	339	152	560	529	167	497
	Predicted	307	142	521	501	148	468
χ^2 statistic		1.5904	0.2839	1.4977	0.7890	1.0620	0.8992
$p - value$		0.207	0.594	0.221	0.374	0.303	0.343

Comparison of Proportions between Simulated and Predicted by Unique Value of PPD

Table IV.66: Chi-squared Test Results for Comparison between Simulated and Predicted PPD Counts by Each Value by Site

Site	Value	Counts PPD (Predicted)	Counts PPD (Simulated)	χ^2 Statistic	p-value
27DV	0	73	272	114.786	**8.77e-27**
	1	2184	1946	13.715	**2.10e-04**
	2	1918	1823	2.412	0.120
	3	518	620	9.142	**2.50e-03**
	4	203	217	0.467	0.495
	5	66	87	2.882	0.090
	6	32	22	1.852	0.174
	7	2	9	4.455	**0.035**
	8	2	2	0	1
	9	2	2	0	1
27V	0	396	531	19.660	**9.25e-06**
	1	2608	2397	8.895	**2.90e-03**
	2	1553	1567	0.063	0.802
	3	301	353	4.135	**0.042**
	4	96	103	0.246	0.620
	5	30	34	0.250	0.617
	6	16	15	0.032	0.857
27MV	0	140	261	36.511	**1.52e-09**
	1	1597	1519	1.953	0.162
	2	1975	1820	6.331	**0.012**
	3	767	840	3.316	0.069
	4	351	375	0.793	0.373
	5	112	124	0.610	0.435
	6	32	34	0.061	0.806
	7	19	20	0.026	0.873
	8	4	4	0	1
	9	3	3	0	1
27DL	0	125	270	53.228	**2.97e-13**
	1	1628	1528	3.169	0.075
	2	2006	1812	9.858	**1.70e-03**
	3	740	861	9.145	**2.50e-03**
	4	324	335	0.184	0.668
	5	114	126	0.600	0.439
	6	37	44	0.605	0.437
	7	18	16	0.118	0.732
	8	7	7	0	1
	9	1	1	0	1
27L	0	164	286	33.076	**8.86e-09**
	1	2299	2147	5.197	**0.023**
	2	1891	1828	1.067	0.302
	3	498	572	5.118	**0.024**
	4	105	125	1.739	0.187
	5	31	30	0.016	0.898
	6	3	3	0	1
	7	6	6	0	1
	8	2	2	0	1
	10	1	1	0	1
27ML	0	91	187	33.151	**8.53e-09**
	1	1403	1369	0.417	0.518
	2	2245	2070	7.097	**7.70e-03**
	3	793	877	4.225	**0.040**
	4	318	334	0.393	0.531
	5	111	120	0.351	0.554
	6	26	31	0.439	0.508
	7	4	3	0.143	0.705
	8	4	3	0.143	0.705
	9	1	1	0	1
	10	1	NA	NA	NA

Comparation of Distributions Trough Kernels

Table IV.67: Kernel Density Estimates Differences and Kolmogorov-Smirnov Test

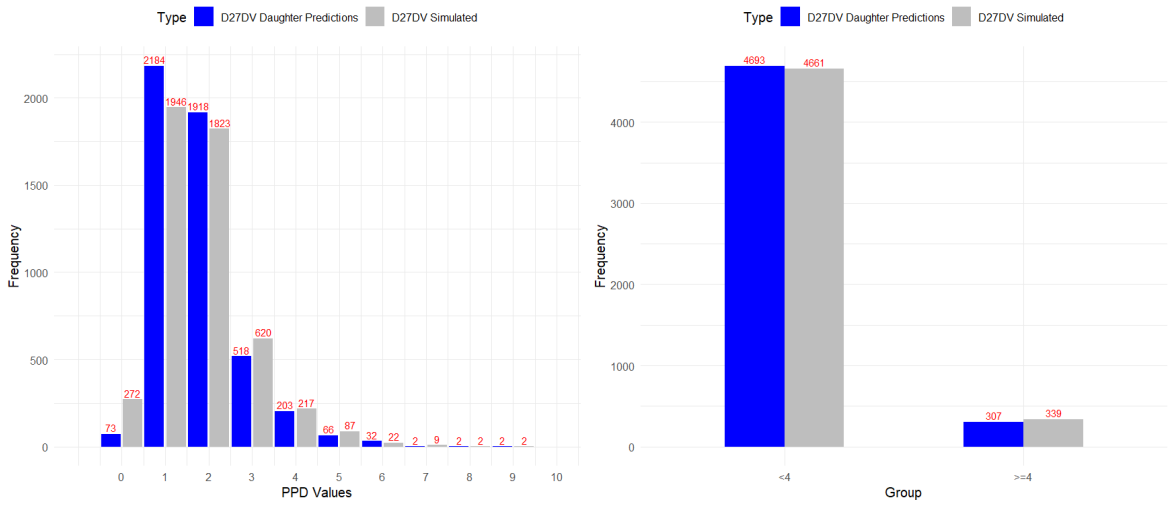
27	Kern.	Band.	Adj.	KDE Dif.	KDE Dif. CI	KS Stat	KS p-val.
27DV	Biweight	2	2	5.92e-03	[3.18e-03, 6.56e-03]	0.0920	0
27V	Biweight	2	2	4.32e-03	[2.55e-03, 5.49e-03]	0.0526	1.96e-06
27MV	Cosine	2	2	1.86e-03	[8.92e-04, 3.96e-03]	0.0342	5.77e-03
27DL	Cosine	2	2	2.80e-03	[8.70e-04, 6.37e-03]	0.0418	3.21e-04
27L	Biweight	2	2	8.94e-03	[2.96e-03, 9.92e-03]	0.0482	1.80e-05
27ML	Cosine	2	2	2.80e-03	[8.70e-04, 6.37e-03]	0.0418	3.21e-04

Abbreviations: Kern. – Kernel Type Function; Band. – Bandwidth; Adj. – Adjustment; KDE Dif. – Mean Kernel Density Difference; KDE Dif. CI – Confidence Interval for Mean Kernel Density Difference; KS Stat – Kolmogorov Smirnov Statistic; KS p-val. – Kolmogorov Smirnov p-value

Visual Comparisons of Proportions and Distributions

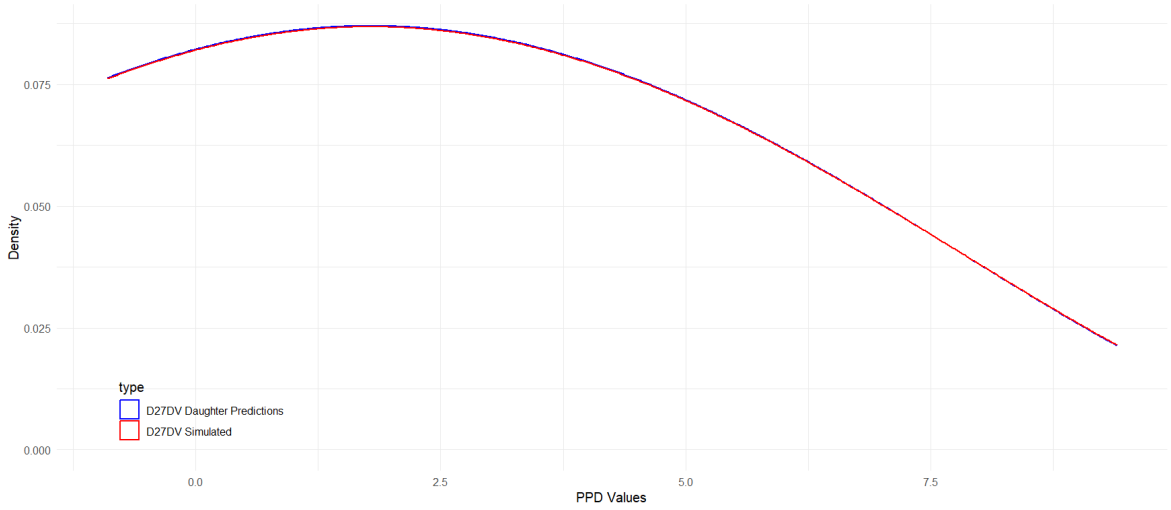
Site DV

Figure IV.73



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

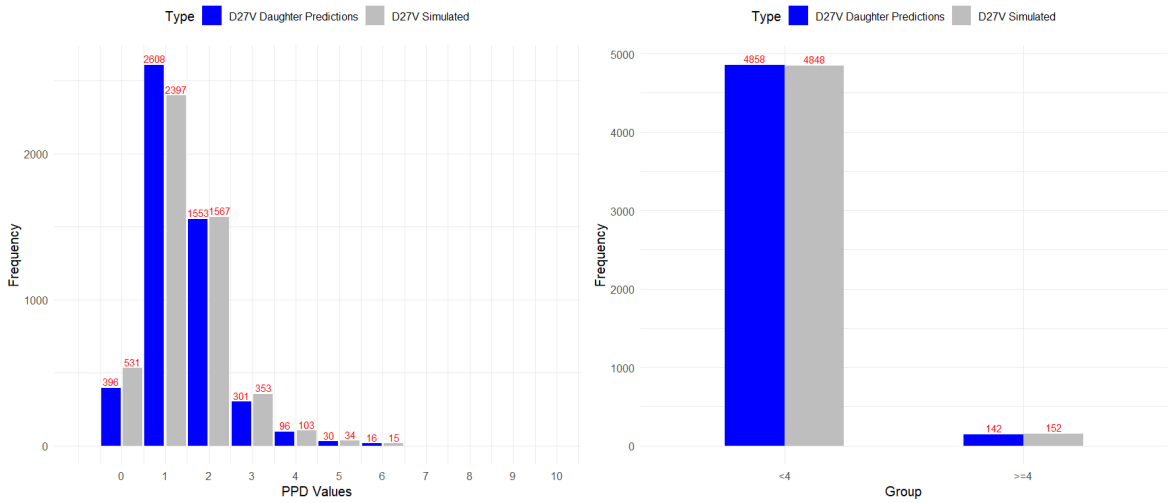
Figure IV.74



Optimal Kernel Density Plots of 27DV for Simulated and Daughter Predicted

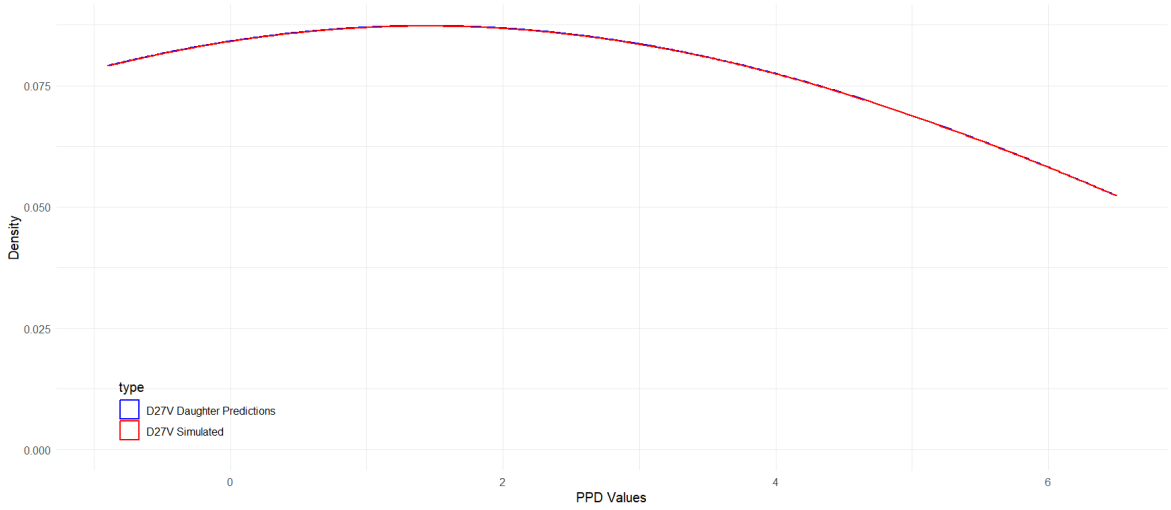
Site V

Figure IV.75



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

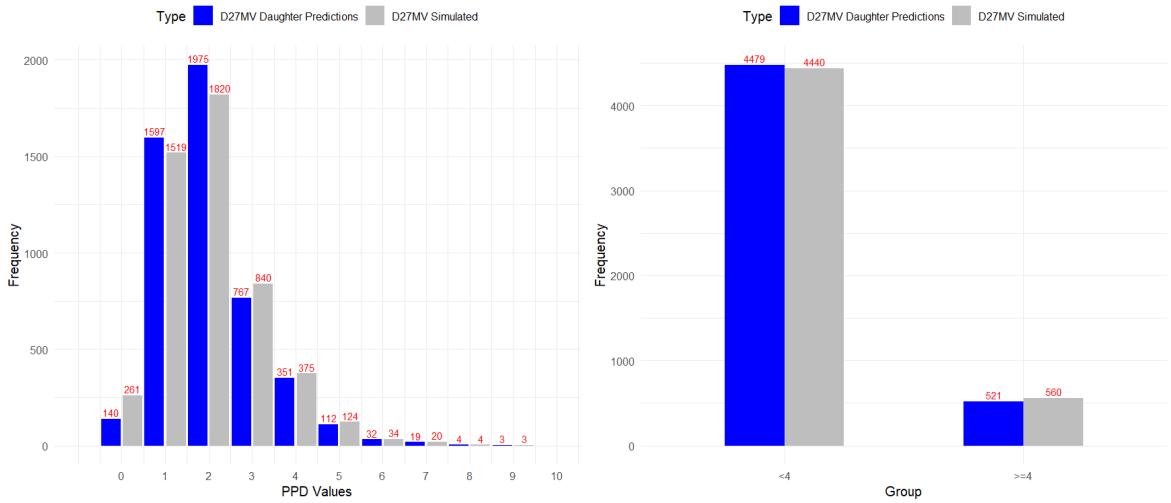
Figure IV.76



Optimal Kernel Density Plots of 27V for Simulated and Daughter Predicted

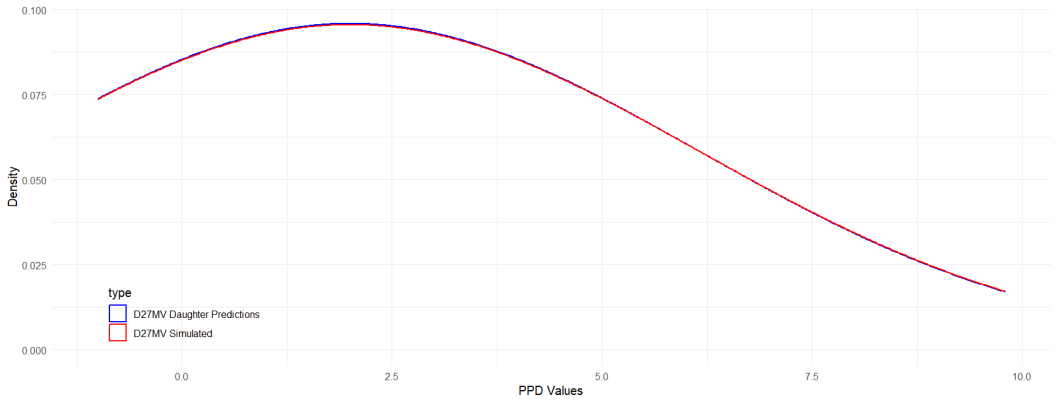
Site MV

Figure IV.77



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

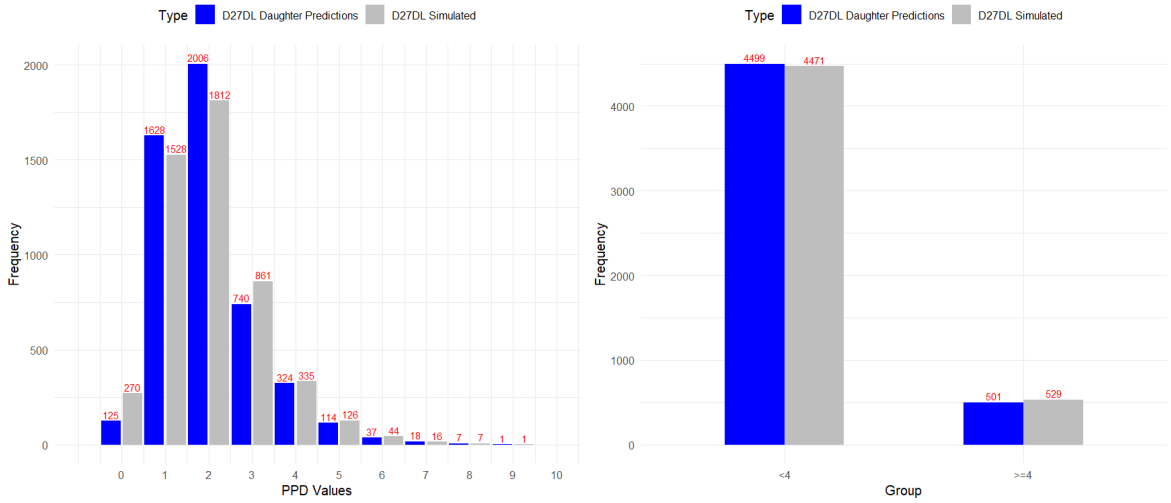
Figure IV.78



Optimal Kernel Density Plots of 27MV for Simulated and Daughter Predicted

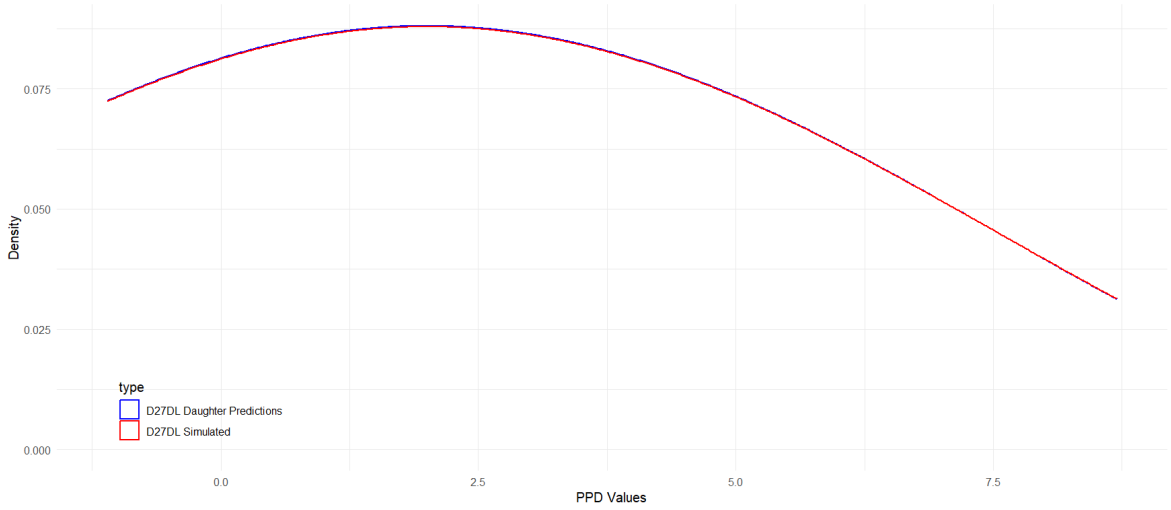
Site DL

Figure IV.79



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

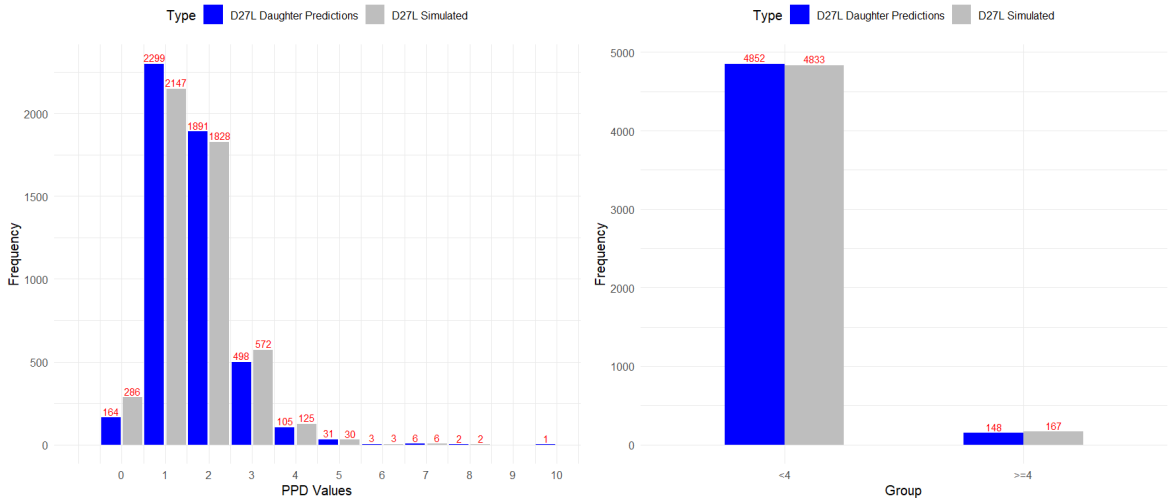
Figure IV.80



Optimal Kernel Density Plots of 27DL for Simulated and Daughter Predicted

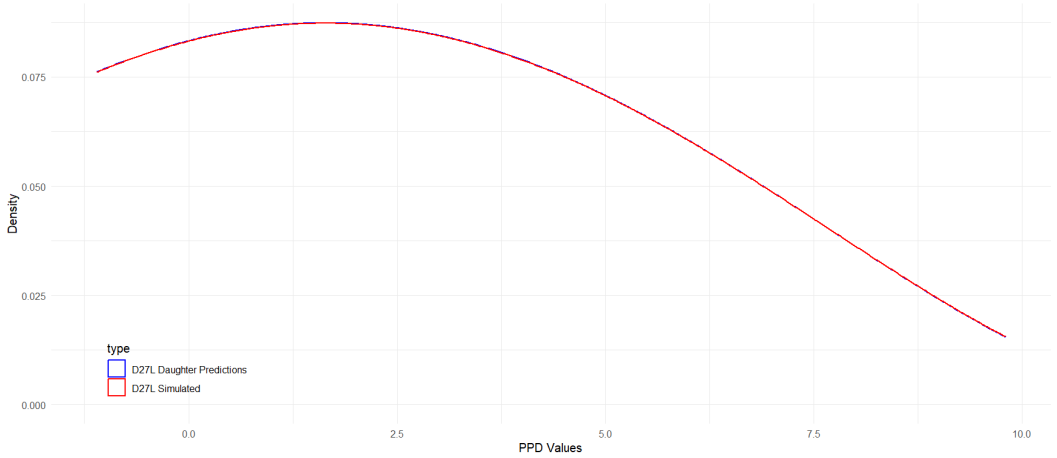
Site L

Figure IV.81



Histograms of simulated vs Predicted: by Unique Value and Values $\geq$ 4 mm

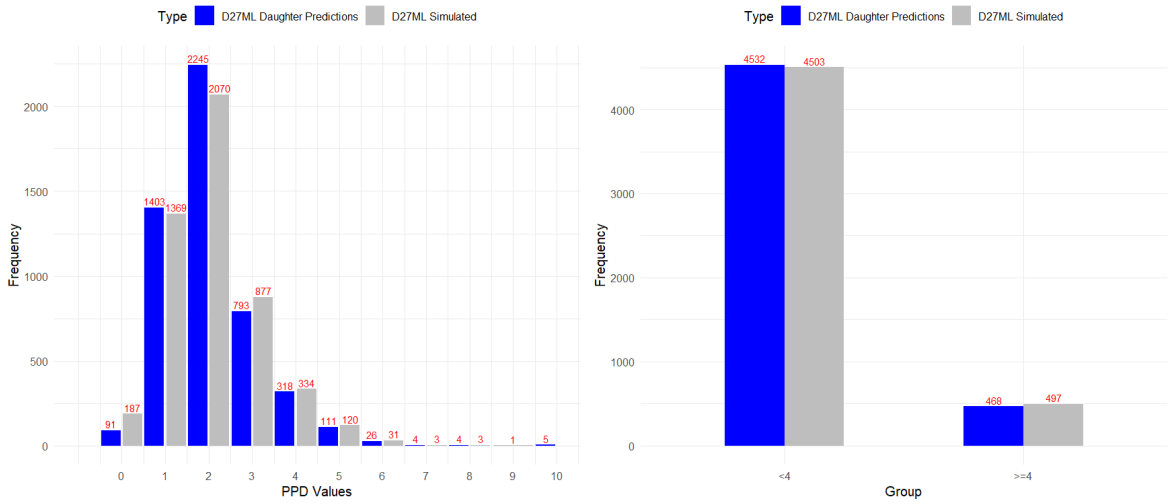
Figure IV.82



Optimal Kernel Density Plots of 27L for Simulated and Daughter Predicted

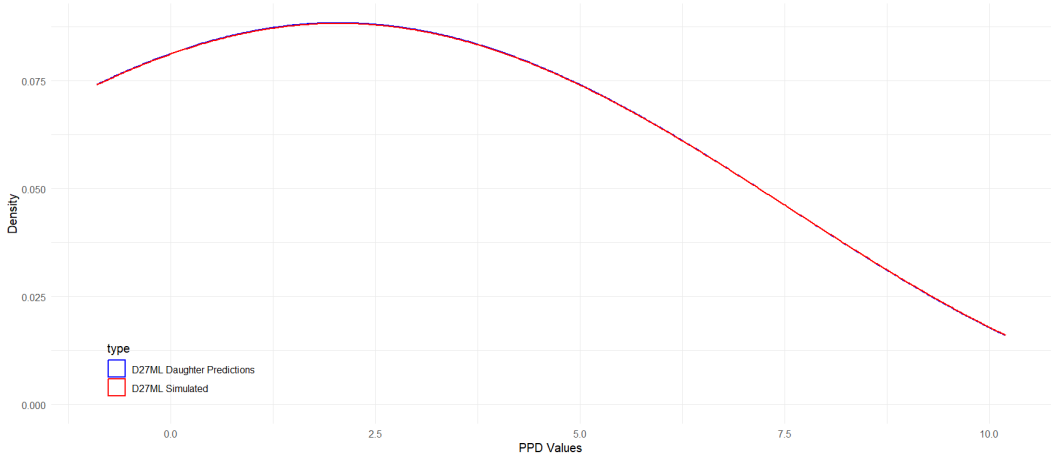
Site ML

Figure IV.83



Histograms of simulated vs Predicted: by Unique Value and Values ≥ 4 mm

Figure IV.84



Optimal Kernel Density Plots of 27ML for Simulated and Daughter Predicted

Appendix V

Appendix V: R Scripts

Data Preparation

```
1|
2|NHANES_11.12_Demog <- read.table("D:/XXXOneDrive/
3|                                  Ambiente de Trabalho/
4|                                  R - Tese UaB/
5|                                  R - NHANES 2011-2012 Demografic.csv",
6|                                  header=T, sep=";")
7|NHANES_11.12_Perio <- read.table("D:/XXXOneDrive/
8|                                  Ambiente de Trabalho/
9|                                  R - Tese UaB/
10|                                 R - NHANES 2011-2012 Periodontal.csv",
11|DEMO <- NHANES_11.12_Demog %>%
12|         dplyr::select(SEQN, RIAGENDR, RIDAGEYR, RIDRETH1,
13|                        DMDEDUC2, INDFMPIR, DMDHREDU)
14|Perio.comp <- dplyr::filter(NHANES_11.12_Perio, OHDPDSTS == 1)
15|QQQ <- filter(Perio.comp,
16|              c(  D17BoDV==99 & D16BoDV==99 & D15BoDV==99 &
17|                   D14BoDV==99 & D13BoML==99 & D12BoDV==99 &
18|                   D11BoDV==99 & D27BoDV==99 & D26BoDV==99 &
19|                   D25BoDV==99 & D24BoDV==99 & D23BoML==99 &
20|                   D22BoDV==99 & D21BoDV==99 & D37BoDV==99 &
21|                   D36BoDV==99 & D35BoDV==99 & D34BoDV==99 &
22|                   D33BoML==99 & D32BoDV==99 & D31BoDV==99 &
23|                   D47BoDV==99 & D46BoDV==99 & D45BoDV==99 &
24|                   D44BoDV==99 & D43BoML==99 & D42BoDV==99 &
25|                   D41BoDV==99))
26|Perio.comp <- anti_join(Perio.comp, QQQ, by = "SEQN")
27|           dplyr::select(SEQN, contains("Bo"))
28|PPD.na <- PPD.ds %>%
29|          mutate_at(vars(-SEQN), na_if, 99)
30|PPD.na_IO <- PPD.na %>%
31|             dplyr::mutate_at(vars(-SEQN), ifelse(is.na(.), "0", "1"))
32|column_pairs <- list(
33|  c("SEQN","D17BoDV", "D27BoDV"), c("SEQN","D17BoV", "D27BoV"),
34|  c("SEQN","D17BoMV", "D27BoMV"), c("SEQN","D17BoDL", "D27BoDL"),
35|  c("SEQN","D17BoL", "D27BoL"),   c("SEQN","D17BoML", "D27BoML"),
36|  c("SEQN","D16BoDV", "D26BoDV"), c("SEQN","D16BoV", "D26BoV"),
37|  c("SEQN","D16BoMV", "D26BoMV"), c("SEQN","D16BoDL", "D26BoDL"),
38|  c("SEQN","D16BoL", "D26BoL"),   c("SEQN","D16BoML", "D26BoML"),
39|  c("SEQN","D15BoDV", "D25BoDV"), c("SEQN","D15BoV", "D25BoV"),
40|  c("SEQN","D15BoMV", "D25BoMV"), c("SEQN","D15BoDL", "D25BoDL"),
41|  c("SEQN","D15BoL", "D25BoL"),   c("SEQN","D15BoML", "D25BoML"),
42|  c("SEQN","D14BoDV", "D24BoDV"), c("SEQN","D14BoV", "D24BoV"),
43|  c("SEQN","D14BoMV", "D24BoMV"), c("SEQN","D14BoDL", "D24BoDL"),
44|  c("SEQN","D14BoL", "D24BoL"),   c("SEQN","D14BoML", "D24BoML"),
45|  c("SEQN","D13BoDV", "D23BoDV"), c("SEQN","D13BoV", "D23BoV"),
46|  c("SEQN","D13BoMV", "D23BoMV"), c("SEQN","D13BoDL", "D23BoDL"),
47|  c("SEQN","D13BoL", "D23BoL"),   c("SEQN","D13BoML", "D23BoML"),
48|  c("SEQN","D12BoDV", "D22BoDV"), c("SEQN","D12BoV", "D22BoV"),
49|  c("SEQN","D12BoMV", "D22BoMV"), c("SEQN","D12BoDL", "D22BoDL"),
50|  c("SEQN","D12BoL", "D22BoL"),   c("SEQN","D12BoML", "D22BoML"),
51|  c("SEQN","D11BoDV", "D21BoDV"), c("SEQN","D11BoV", "D21BoV"),
52|  c("SEQN","D11BoMV", "D21BoMV"), c("SEQN","D11BoDL", "D21BoDL"),
53|  c("SEQN","D11BoL", "D21BoL"),   c("SEQN","D11BoML", "D21BoML"),
54|  c("SEQN","D47BoDV", "D37BoDV"), c("SEQN","D47BoV", "D37BoV"),
55|  c("SEQN","D47BoMV", "D37BoMV"), c("SEQN","D47BoDL", "D37BoDL"),
56|  c("SEQN","D47BoL", "D37BoL"),   c("SEQN","D47BoML", "D37BoML"),
57|  c("SEQN","D46BoDV", "D36BoDV"), c("SEQN","D46BoV", "D36BoV"),
58|  c("SEQN","D46BoMV", "D36BoMV"), c("SEQN","D46BoDL", "D36BoDL"),
59|  c("SEQN","D46BoL", "D36BoL"),   c("SEQN","D46BoML", "D36BoML"),
```

```
  c("SEQN","D45BoDV", "D35BoDV"), c("SEQN","D45BoV", "D35BoV"),
  c("SEQN","D45BoMV", "D35BoMV"), c("SEQN","D45BoDL", "D35BoDL"),
  c("SEQN","D45BoL", "D35BoL"),   c("SEQN","D45BoML", "D35BoML"),
  c("SEQN","D44BoDV", "D34BoDV"), c("SEQN","D44BoV", "D34BoV"),
  c("SEQN","D44BoMV", "D34BoMV"), c("SEQN","D44BoDL", "D34BoDL"),
  c("SEQN","D44BoL", "D34BoL"),   c("SEQN","D44BoML", "D34BoML"),
  c("SEQN","D43BoDV", "D33BoDV"), c("SEQN","D43BoV", "D33BoV"),
  c("SEQN","D43BoMV", "D33BoMV"), c("SEQN","D43BoDL", "D33BoDL"),
  c("SEQN","D43BoL", "D33BoL"),   c("SEQN","D43BoML", "D33BoML"),
  c("SEQN","D42BoDV", "D32BoDV"), c("SEQN","D42BoV", "D32BoV"),
  c("SEQN","D42BoMV", "D32BoMV"), c("SEQN","D42BoDL", "D32BoDL"),
  c("SEQN","D42BoL", "D32BoL"),   c("SEQN","D42BoML", "D32BoML"),
  c("SEQN","D41BoDV", "D31BoDV"), c("SEQN","D41BoV", "D31BoV"),
  c("SEQN","D41BoMV", "D31BoMV"), c("SEQN","D41BoDL", "D31BoDL"),
  c("SEQN","D41BoL", "D31BoL"),   c("SEQN","D41BoML", "D31BoML"))
df_list <- list()
for (i in seq_along(column_pairs)) {
  pair <- column_pairs[[i]]
  current_columns <- pair
  # Simplify names for pair identification
  simplified_pair_name <- gsub("Bo", "", paste(pair[-1],
                               collapse = "_"))
  simplified_pair_name <- gsub("D(\\d)(\\d)(DV|V|MV|DL|L|ML)",
                               "\\1.\\2\\3",
                               simplified_pair_name)
  df_list[[simplified_pair_name]] <- PPD.na[, current_columns,
                                            drop = FALSE]}
df_list <- lapply(names(df_list), function(pair_name) {
  df <- df_list[[pair_name]]
  if (ncol(df) >= 3) {
    beta <- cor(df[[2]], df[[3]], use = "complete.obs")
    alpha <- 1  # Alpha is given as 1
    df[[paste0("SM.", pair_name)]] <- round(exp(-abs(df[[2]] - df[[3]]) / (alpha + beta * (df[[2]] +
        df[[3]]) / 2)), 1)
    df[[paste0("gamma1.", pair_name)]] <- ifelse(df[[2]] == df[[3]], 1, ifelse(df[[2]] > df[[3]], 1,
        -1))
    df[[paste0("gamma0.", pair_name)]] <- ifelse(df[[2]] == df[[3]], 0, ifelse(df[[2]] > df[[3]], 1,
        -1))
    df[[paste0("SM.d.", pair_name)]] <- df[[paste0("SM.", pair_name)]] * df[[paste0("gamma1.", pair_
        name)]]
    df[[paste0("SM.d0.", pair_name)]] <- df[[paste0("SM.", pair_name)]] * df[[paste0("gamma0.", pair
        _name)]]  }
  return(df)})
specific_df <- df_list[["12.6DL"]]
print(specific_df)
all_dfs <- df_list[1:length(df_list)]
ALL.all.Bo <- reduce(all_dfs, left_join, by = "SEQN")
UABo_BI <- ALL.all.Bo %>%
  dplyr::select(
    SEQN,  # Always include SEQN
    matches("^(D1|D2|SM\\.1|gamma1\\.1\\.|gamma0\\.1|
    SM\\.d\\.1\\.|SM\\.d0\\.1\\.)")      )
LABo_BI <- ALL.all.Bo %>%
  dplyr::select(
    SEQN,  # Always include SEQN
    matches("^(D3|D4|SM\\.4|gamma1\\.4\\.|gamma0\\.4|
    SM\\.d\\.4\\.|SM\\.d0\\.4\\.)"     )
UABo_BI <- UABo_BI %>%
            rename_with(~ gsub("^D(.+)Bo", "D\\1", .x),
            starts_with("D"))
write.csv(UABo_BI, "C:/R Works/Tese/UABo_BI.csv",
```

```
116|           row.names = FALSE)
```

Assessing Symmetry

```
1|
2| packages <- c("ggplot2","dplyr", "tidyr", "gamlss", "caret", "Metrics","reshape2", "car")
3| invisible(lapply(packages, library, character.only = TRUE))
4|UABo_BI <-read.table("C:/R Works/Tese/UABo_BI.csv", header=T, sep=",")
5| ctrl <- gamlss.control(n.cyc = 500, trace = FALSE)
6| library(dplyr)
7| library(tidyr)
8| process_teeth_data <- function(data, tooth_suffixes, main_col = "SEQN") {
9|   patterns <- paste0("(", paste(tooth_suffixes, collapse = "|"), ")")
10|   columns_to_select <- grep(patterns, names(data), value = TRUE)
11|   columns_to_select <- c(main_col, columns_to_select)
12|   selected_data <- data[, columns_to_select]
13|   na_count <- rowSums(is.na(selected_data[-1]))
14|   clean_data <- selected_data[na_count < (ncol(selected_data) - 1), ]
15|   return(clean_data)}
16| reshape_for_modeling <- function(data, sites) {
17|   long_data_list <- lapply(names(sites), function(site) {
18|     pivot_longer(data, cols = sites[[site]], names_to = "Side", values_to = paste0("PPD.", site, ".
        U1"))  })
19|   names(long_data_list) <- names(sites)
20|   return(long_data_list)}
21| sites_to_model <- list(
22|   DV = c("D11DV", "D21DV"),
23|   V  = c("D11V", "D21V"),
24|   MV = c("D11MV", "D21MV"),
25|   DL = c("D11DL", "D21DL"),
26|   L  = c("D11L", "D21L"),
27|   ML = c("D11ML", "D21ML"))
28| UABo_BI_clean <- process_teeth_data(UABo_BI, c("11", "21"))
29| PPD_models_data <- reshape_for_modeling(UABo_BI_clean, sites_to_model)
30| UABo_BI_cleanU1 <- process_teeth_data(UABo_BI, c("11", "21"))
31| PPD_models_dataU1 <- reshape_for_modeling(UABo_BI_cleanU1, sites_to_model)
32| PPD.DV.U1.L <- PPD_models_dataU1$DV
33| PPD.V.U1.L  <- PPD_models_dataU1$V
34| PPD.MV.U1.L <- PPD_models_dataU1$MV
35| PPD.DL.U1.L <- PPD_models_dataU1$DL
36| PPD.L.U1.L  <- PPD_models_dataU1$L
37| PPD.ML.U1.L <- PPD_models_dataU1$ML
38| PPD.DV.U1.L <- na.omit(PPD.DV.U1.L)
39| PPD.V.U1.L  <- na.omit(PPD.V.U1.L)
40| PPD.MV.U1.L <- na.omit(PPD.MV.U1.L)
41| PPD.DL.U1.L <- na.omit(PPD.DL.U1.L)
42| PPD.L.U1.L  <- na.omit(PPD.L.U1.L)
43| PPD.ML.U1.L <- na.omit(PPD.ML.U1.L)
44| PPDDV_U1 <- fitDist(PPD.DV.U1, data = PPD.DV.U1.L, type = "realplus", control = ctrl)
45| PPDDV_U1$fits
46| #####     exGAUS       EXP      PARETO2o    PARETO2
47| #####   10954.65    13863.83   13865.84   13865.85
48| modelU1.DV.A <- gamlss( PPD.DV.U1 ~ 1,
49|                         sigma.formula = ~1,
50|                         nu.formula = ~1,
51|                         tau.formula = ~1,
```

```
                          data = PPD.DV.U1.L,
                          family = exGAUS,
                          method = mixed(400, 400),
                          control = ctrl)
modelU1.DV.B <- gamlss( PPD.DV.U1 ~ Side,
                          sigma.formula = ~Side,
                          nu.formula = ~Side,
                          tau.formula = ~Side,
                          data = PPD.DV.U1.L,
                          family = exGAUS,
                          method = mixed(200, 200),
                          control = ctrl)
LR.test_modelU1.DV.AB <- lrtest(modelU1.DV.A, modelU1.DV.B)
summary(modelU1.DV.A)
summary(modelU1.DV.B)
print(LR.test_modelU1.DV.AB)
set.seed(123)
folds <- createFolds(PPD.DV.U1.L$PPD.DV.U1, k = 5)
results_modelPPD.DV.U1A <- list()
results_modelPPD.DV.U1B <- list()
for (i in seq_along(folds)) {
  train_data <- PPD.DV.U1.L[-folds[[i]], ]
  test_data <- PPD.DV.U1.L[folds[[i]], ]
  modelPPD.DV.U1A <- gamlss(PPD.DV.U1 ~ 1,
                            data = train_data,
                            family = exGAUS,
                            method = mixed(200, 200),
                            control = ctrl)
  modelPPD.DV.U1B <- gamlss(PPD.DV.U1 ~ Side,
                            sigma.formula = ~Side,
                            nu.formula = ~Side,
                            tau.formula = ~Side,
                            data = train_data,
                            family = exGAUS,
                            method = mixed(200, 2050),
                            control = ctrl)
  predictions_modelPPD.DV.U1A <- predict(modelPPD.DV.U1A,
                                         newdata = test_data,
                                         type = "response")
  predictions_modelPPD.DV.U1B <- predict(modelPPD.DV.U1B,
                                         newdata = test_data,
                                         type = "response")
  results_modelPPD.DV.U1A[[i]] <- list(
                    RMSE = rmse(test_data$PPD.DV.U1,
                                predictions_modelPPD.DV.U1A),
                    MAE = mae(test_data$PPD.DV.U1,
                                predictions_modelPPD.DV.U1A))
  results_modelPPD.DV.U1B[[i]] <- list(
                    RMSE = rmse(test_data$PPD.DV.U1,
                                predictions_modelPPD.DV.U1B),
                    MAE = mae(test_data$PPD.DV.U1,
                                predictions_modelPPD.DV.U1B))}
mean_rmse_modelPPD.DV.U1A <- mean(sapply(results_modelPPD.DV.U1A,
                                         function(x) x$RMSE))
mean_rmse_modelPPD.DV.U1B <- mean(sapply(results_modelPPD.DV.U1B,
                                         function(x) x$RMSE))
mean_mae_modelPPD.DV.U1A <- mean(sapply(results_modelPPD.DV.U1A,
                                        function(x) x$MAE))
mean_mae_modelPPD.DV.U1B <- mean(sapply(results_modelPPD.DV.U1B,
                                        function(x) x$MAE))
t_test_rmse <- t.test(sapply(results_modelPPD.DV.U1A,
                             function(x) x$RMSE),
                      sapply(results_modelPPD.DV.U1B,
```

```
                                function(x) x$RMSE),
                                paired = TRUE)
t_test_mae <- t.test(sapply(results_modelPPD.DV.U1A,
                                function(x) x$MAE),
                                sapply(results_modelPPD.DV.U1B,
                                function(x) x$MAE),
                                paired = TRUE)
print(paste("Mean RMSE A =", round(mean_rmse_modelPPD.DV.U1A, 3)))
print(paste("Mean RMSE B =", round(mean_rmse_modelPPD.DV.U1B, 3)))
print(t_test_rmse)
print(paste("Mean MAE A =", round(mean_mae_modelPPD.DV.U1A, 3)))
print(paste("Mean MAE B =", round(mean_mae_modelPPD.DV.U1B, 4)))
print(t_test_mae)
```

V.0.1 MoDau IMputation

```
UABo_AI <- read.table("C:/R Works/Tese/UABo_AI.csv", header = TRUE, sep = ",")
library(dplyr)
library(xgboost)
library(ggplot2)
library(reshape2)
library(boot)
library(caret)
library(gridExtra)
NRrounds=300000
Imp.U2 <- UABo_AI %>%
  dplyr::select(SEQN, contains("12"), contains("22")) %>%
  dplyr::select(-contains("_imp"))
create_new_vars <- function(df, col1, col2, tooth_type, site) {
  df %>%
    mutate(
      !!paste0("U2Dif_", tooth_type, "_", site) := !!sym(col1) - !!sym(col2),
      !!paste0("U2rat_", tooth_type, "_", site) := ifelse((!!sym(col2) == 0), 0, !!sym(col1) / !!sym
          (col2)),
      !!paste0("U2mean_", tooth_type, "_", site) := rowMeans(cbind(!!sym(col1), !!sym(col2)), na.rm
          = TRUE)     )}
sites <- c("DV", "V", "MV", "DL", "L", "ML")
for (site in sites) {
  col1 <- paste0("D12", site)
  col2 <- paste0("D22", site)
  Imp.U2 <- create_new_vars(Imp.U2, col1, col2, "U2", site)}
corrD22V <- cor(Imp.U2$D12V, Imp.U2$D22V, method = "spearman")
SM <- function(A, B, alpha = 1, beta = corrD22V) {
  sapply(seq_along(A), function(i) {
    exp(-abs(A[i] - B[i]) / (alpha + beta * (A[i] + B[i]) / 2))  })}
Imp.U2$SM.U2.V <- SM(Imp.U2$D12V, Imp.U2$D22V)
Imp.U2$gammaU2V <- with(Imp.U2, ifelse(D12V != D22V, abs(D12V - D22V) / (D12V - D22V), 1))
Imp.U2$SMId.U12.V <- with(Imp.U2, SM.U2.V * gammaU2V)
set.seed(123)
training_indices <- sample(seq_len(nrow(Imp.U2)), size = 3320)
trainingU2.V <- Imp.U2[training_indices, ]
Mother_features <- c("D12V", "SMId.U12.V")
train_matrix <- xgb.DMatrix(data = as.matrix(trainingU2.V[, Mother_features]),
                            label = trainingU2.V$D22Dv)
Mother_params <- list(objective = "reg:squarederror",
  max_depth = 3,
  eta = 0.1,
```

```r
  gamma = 0,
  colsample_bytree = 1.0,
  min_child_weight = 3,
  subsample = 0.7)
Mother_cv <- xgb.cv(
  params = Mother_params,
  data = train_matrix,
  nrounds = 1000,
  nfold = 5,
  verbose = 0,
  early_stopping_rounds = 10,
  maximize = FALSE)
best_nrounds <- Mother_cv$best_iteration
Mother_model <- xgb.train(
  params = Mother_params,
  data = train_matrix,
  nrounds = best_nrounds,
  watchlist = list(train = train_matrix),
  verbose = 0)
Mother_predictions <- predict(Mother_model, train_matrix)
D22v <- Mother_predictions
Mother_model_rmse <- sqrt(mean((trainingU2.V$D22V - Mother_predictions)^2))
Mother_model_mae <- mean(abs(trainingU2.V$D22V - Mother_predictions))
Mother_model_rsq <- cor(trainingU2.V$D22V, Mother_predictions)^2
Mother_model_mse <- mean((trainingU2.V$D22V - Mother_predictions)^2)
NnS <- 5000          # Length of new data set
noise_level <- 0.4   # Noise Level
firstRows <- 5      # Number of retraining rows
add_noise <- function(data, noise_level = noise_level) {
  noise <- rnorm(n = length(data), mean = 0, sd = noise_level * sd(data))
  data + noise}
sample_indices <- sample(1:nrow(trainingU2.V), NnS, replace = TRUE)
new_D12V_values <- add_noise(trainingU2.V$D12V[sample_indices], noise_level = noise_level)
new_D22V_values <- add_noise(trainingU2.V$D22V[sample_indices], noise_level = noise_level)
Mother_predictions_expanded <- Mother_predictions[sample_indices]
new_data <- data.frame(D12V = new_D12V_values, D22V = new_D22V_values)
new_data <- na.omit(new_data)
sampled_original_D22V <- trainingU2.V$D22V[sample(1:nrow(trainingU2.V), NnS, replace = TRUE)]
corrU2 <- cor(new_data$D12V[1:firstRows], new_data$D22V[1:firstRows], method = "spearman")
SM <- function(A, B, alpha = 1, beta = corrU2) {
  sapply(seq_along(A), function(i) {
    exp(-abs(A[i] - B[i]) / (alpha + beta * (A[i] + B[i]) / 2))  })}
new_data$new_SM.U2.V <- SM(new_data$D12V, new_data$D22V)
new_data$gammaU2V <- with(new_data, ifelse(D12V != D22V, abs(D12V - D22V) / (D12V - D22V), 1))
new_data$new_SMId.U12.V <- with(new_data, new_SM.U2.V * gammaU2V)
for (col in names(new_data)) {
  if (any(is.infinite(new_data[[col]]) | is.na(new_data[[col]]))) {
    median_value <- median(new_data[[col]], na.rm = TRUE)
    new_data[[col]][is.infinite(new_data[[col]]) | is.na(new_data[[col]])] <- median_value }}
Mother_matrix_200 <- xgb.DMatrix(data = as.matrix(new_data[1:firstRows, c("D12V", "new_SMId.U12.V")
    ]),
                                 label = new_data$D22V[1:firstRows])
Mother_model_200 <- xgb.train(
  params = Mother_params,
  data = Mother_matrix_200,
  nrounds = 1000,
  watchlist = list(train = Mother_matrix_200),
  verbose = 0)
new_Mother_predictions <- predict(Mother_model_200, Mother_matrix_200)
new_Mother_predictions <- predict(Mother_model_200, xgb.DMatrix(data = as.matrix(new_data[1:
    firstRows, c("D12V", "new_SMId.U12.V")])))
new_data_200 <- new_data[1:firstRows, ]
new_data_200$new_Mother_predictions <- new_Mother_predictions
```

```r
remaining_data <- new_data[(firstRows+1):nrow(new_data), ]
remaining_data$new_Mother_predictions <- Mother_predictions_expanded[(firstRows+1):nrow(new_data)]
new_data_combined <- rbind(new_data_200, remaining_data)
new_data_combined$Mother_predictions <- c(Mother_predictions_expanded[1:firstRows], Mother_predictions_expanded[(firstRows+1):nrow(new_data)])
new_data_combined <- na.omit(new_data_combined)
set.seed(42)  # For reproducibility
new_data_combined <- new_data_combined[sample(nrow(new_data_combined)), ]
new_data_combined <- na.omit(new_data_combined)
Daughter_matrix <- xgb.DMatrix(data = as.matrix(new_data_combined[, c("D12V", "Mother_predictions", "new_Mother_predictions")]),
                               label = new_data_combined$D22V)
tune_grid <- expand.grid(
  max_depth = c(3),
  eta = c(0.05),
  gamma = c(0),
  colsample_bytree = c(0.8),
  min_child_weight = c(2),
  subsample = c(0.7),
  nrounds = NRrounds)
train_control <- trainControl(
  method = "cv",
  number = 5,
  verboseIter = FALSE,
  returnData = FALSE,
  returnResamp = "none",
  classProbs = FALSE,
  summaryFunction = defaultSummary,
  allowParallel = TRUE)
Daughter_caret_model <- train(
  x = as.matrix(new_data_combined[, c("D12V", "Mother_predictions", "new_Mother_predictions")]),
  y = new_data_combined$D22V,
  trControl = train_control,
  tuneGrid = tune_grid,
  method = "xgbTree")
best_params <- list(
  objective = "reg:squarederror",
  max_depth = Daughter_caret_model$bestTune$max_depth,
  eta = Daughter_caret_model$bestTune$eta,
  gamma = Daughter_caret_model$bestTune$gamma,
  colsample_bytree = Daughter_caret_model$bestTune$colsample_bytree,
  min_child_weight = Daughter_caret_model$bestTune$min_child_weight,
  subsample = Daughter_caret_model$bestTune$subsample,
  nrounds = Daughter_caret_model$bestTune$nrounds)
watchlist <- list(train = Daughter_matrix)
Daughter_model <- xgb.train(
  params = best_params,
  data = Daughter_matrix,
  nrounds = best_params$nrounds,
  watchlist = watchlist,
  verbose = 0,
  early_stopping_rounds = 10)
Daughter_predictions <- predict(Daughter_model, Daughter_matrix)
new_data_combined$predicted_D22V <- Daughter_predictions
new_Daughter_rmse <- sqrt(mean((new_data_combined$D22V - Daughter_predictions)^2))
new_Daughter_mae <- mean(abs(new_data_combined$D22V - Daughter_predictions))
new_Daughter_rsq <- cor(new_data_combined$D22V, Daughter_predictions)^2
new_Daughter_mse <- mean((new_data_combined$D22V - Daughter_predictions)^2)
density_diff <- function(data, indices, kernel = "gaussian", bw = "nrd0") {
  d1 <- data[indices, 1]
  d2 <- data[indices, 2]
  d1 <- na.omit(as.numeric(d1))
  d2 <- na.omit(as.numeric(d2))
```

```
  if (length(d1) < 2 || length(d2) < 2) {
    return(NA)  }
  density1 <- density(d1, kernel = kernel, bw = bw)
  density2 <- density(d2, kernel = kernel, bw = bw)
  return(mean(abs(density1$y - density2$y)))}
min_length <- min(length(new_data_combined$D12V), length(new_data_combined$D22V),
                  length(trainingU2.V$D12V), length(trainingU2.V$D22V))
rounded_new_D12V <- round(as.numeric(new_data_combined$D12V), 0)
rounded_new_D22V <- round(as.numeric(new_data_combined$D22V), 0)
rounded_training_D12V <- round(as.numeric(trainingU2.V$D12V), 0)
rounded_training_D22V <- round(as.numeric(trainingU2.V$D22V), 0)
set.seed(42)
rounded_new_D12V <- sample(rounded_new_D12V, min_length, replace = TRUE)
rounded_new_D22V <- sample(rounded_new_D22V, min_length, replace = TRUE)
rounded_training_D12V <- sample(rounded_training_D12V, min_length, replace = TRUE)
rounded_training_D22V <- sample(rounded_training_D22V, min_length, replace = TRUE)
plot_data_density <- data.frame(
  value = c(rounded_new_D12V, rounded_new_D22V, rounded_training_D12V, rounded_training_D22V),
  type = c(rep("Simulated D12V", min_length),
           rep("Simulated D22V", min_length),
           rep("Original D12V", min_length),
           rep("Original D22V", min_length)))
plot_data_long_density <- melt(plot_data_density, id.vars = "type")
Kernel_Density_Plot_of_Simulated_U2V <- ggplot(plot_data_long_density, aes(x = value, color = type))
    +
  geom_density(alpha = 0.9, size = 1) +  # Increase the line thickness with size parameter
  scale_fill_manual(values = c("Simulated D12V" = "blue", "Simulated D22V" = "red",
                               "Original D12V" = "lightblue", "Original D22V" = "pink")) +
  scale_color_manual(values = c("Simulated D12V" = "blue", "Simulated D22V" = "red",
                                "Original D12V" = "lightblue", "Original D22V" = "pink")) +
  scale_x_continuous(breaks = 0:8, limits = c(0, 10)) +
  labs(title = paste("Kernel Density Plot of Simulated D12V, D22V (noise level =", noise_level, ")
      and Original D12V, D22V"),
       x = "Value",
       y = "Density",
       fill = "Type",
       color = "Type") +
  theme_minimal() +
  theme(legend.position = "top")
Kernel_Density_Plot_of_Simulated_U2V
data_for_comparison <- data.frame(
  simulated_D12V = rounded_new_D12V,
  simulated_D22V = rounded_new_D22V,
  original_D12V = rounded_training_D12V,
  original_D22V = rounded_training_D22V)
data_for_comparison <- na.omit(data_for_comparison)
rounded_new_D22V <- round(new_data_combined$D22V, 1)
rounded_Daughter_predictions <- round(new_data_combined$predicted_D22V, 1)
plot_data_Daughter <- data.frame(
  value = c(rounded_new_D22V, rounded_Daughter_predictions),
  type = rep(c("D22V Simulated", "D22V Daughter Predictions"), each = length(rounded_new_D22V)))
plot_data_long_Daughter <- melt(plot_data_Daughter, id.vars = "type")
bandwidths <- seq(0.1, 2, by = 0.1)
kernels <- c("gaussian", "epanechnikov", "rectangular", "triangular", "biweight", "cosine", "
    optcosine")
adjustments <- seq(0.5, 2, by = 0.1)
parameter_grid <- expand.grid(bandwidth = bandwidths, kernel = kernels, adjust = adjustments)
compute_kde_difference <- function(data, bw, kern, adj) {
  density1 <- density(data[,1], bw = bw, kernel = kern, adjust = adj)
  density2 <- density(data[,2], bw = bw, kernel = kern, adjust = adj)
  diff <- mean((density1$y - density2$y)^2)
  return(diff)}
results <- apply(parameter_grid, 1, function(params) {
```

```r
  diff <- compute_kde_difference(
    data = as.matrix(data_for_comparison),
    bw = as.numeric(params["bandwidth"]),
    kern = as.character(params["kernel"]),
    adj = as.numeric(params["adjust"])  )
  return(c(bw = params["bandwidth"], kernel = params["kernel"], adjust = params["adjust"], score =
      diff))})
results_df <- as.data.frame(t(results), stringsAsFactors = FALSE)
colnames(results_df) <- c("bandwidth", "kernel", "adjust", "score")
results_df$score <- as.numeric(results_df$score)  # Ensure 'score' is numeric for further analysis
best_params <- results_df[which.min(results_df$score),]
best_bw <- as.numeric(best_params$bandwidth)
best_kernel <- as.character(best_params$kernel)
best_adjust <- as.numeric(best_params$adjust)
Optimal_Kernel_Density_Plot_of_D22V <- ggplot(plot_data_long_Daughter, aes(x = value, color = type))
    +
  geom_density(bw = best_bw, kernel = best_kernel, adjust = best_adjust, size = .7) +
  scale_color_manual(values = c("D22V Simulated" = "red", "D22V Daughter Predictions" = "blue")) +
  labs(title = "Optimal Kernel Density Plot of D22V PPD for Simulated and Daughter Predicted",
       x = "PPD Values",
       y = "Density") +
  theme_minimal() +
  theme(legend.position = c(0.05, 0.05), legend.justification = c(0, 0))
Optimal_Kernel_Density_Plot_of_D22V
density_diff <- function(data, indices) {
  d1 <- data[indices, 1]
  d2 <- data[indices, 2]
  density1 <- density(d1, bw = best_bw, kernel = best_kernel, adjust = best_adjust)
  density2 <- density(d2, bw = best_bw, kernel = best_kernel, adjust = best_adjust)
  diff <- mean(abs(density1$y - density2$y))
  return(diff)}
sim_vs_orig_Daughter <- as.matrix(data_for_comparison)
boot_result_Daughter <- boot(data = sim_vs_orig_Daughter, statistic = density_diff, R = 1000)
print(paste("kernel density estimates - Best Kernel:", best_kernel))
print(paste("kernel density estimates - Best Bandwidth:", best_bw))
print(paste("kernel density estimates -  Best Adjust:", best_adjust))
cat("Bootstrap test for kernel density estimates difference between D22V and Daughter Predictions:\n
    ")
cat("Mean difference in densities:", mean(boot_result_Daughter$t), "\n")
ci_Daughter <- boot.ci(boot_result_Daughter, type = "perc")
cat("95% CI for the difference in densities:", ci_Daughter$percent[4:5], "\n")
Daughter_predictions_rounded <- round(Daughter_predictions, 0)
Daughter_predictions_rounded[Daughter_predictions_rounded < 0] <- 0
new_data_combined$D22V_rounded <- round(new_data_combined$D22V, 0)
new_data_combined$D22V_rounded[new_data_combined$D22V_rounded < 0] <- 0
plot_data <- data.frame(
  Value = c(Daughter_predictions_rounded, new_data_combined$D22V_rounded),
  Type = c(rep("D22V Daughter Predictions", length(Daughter_predictions_rounded)),
           rep("D22V Simulated", length(new_data_combined$D22V_rounded))))
plot_data_counts <- plot_data %>%
  group_by(Value, Type) %>%
  summarise(Frequency = n(), .groups = 'drop')
Hist_Unique_Val_D22DV <- ggplot(plot_data_counts, aes(x = Value, y = Frequency, fill = Type)) +
  geom_bar(stat = "identity", position = position_dodge(width = 1), width = 0.9) +
  scale_fill_manual(values = c("D22V Daughter Predictions" = "blue", "D22V Simulated" = "grey")) +
  scale_x_continuous(breaks = 0:10, limits = c(-1, 10)) +
  labs(title = "A - D22V Prevalences of unique values of PPD",
       x = "PPD Values",
       y = "Frequency",
       fill = "Type") +
  theme_minimal() +
  theme(legend.position = "top",
    axis.text.x = element_text(angle = 0, hjust = 0.5)  ) +
```

```
  geom_text(aes(label = Frequency), vjust = -0.3, position = position_dodge(1), size = 3, color = "red")
Daughter_predictions_grouped <- ifelse(Daughter_predictions_rounded < 4, "<4", ">=4")
new_data_combined_grouped <- ifelse(new_data_combined$D22V_rounded < 4, "<4", ">=4")
plot_data_grouped <- data.frame(
  Group = c(Daughter_predictions_grouped, new_data_combined_grouped),
  Type = c(rep("D22V Daughter Predictions", length(Daughter_predictions_grouped)),
           rep("D22V Simulated", length(new_data_combined_grouped))))
plot_data_counts_grouped <- plot_data_grouped %>%
  group_by(Group, Type) %>%
  summarise(Frequency = n(), .groups = 'drop')
Hist_Group_Val_D22DV <- ggplot(plot_data_counts_grouped, aes(x = Group, y = Frequency, fill = Type)) +
  geom_bar(stat = "identity", position = position_dodge(width = .5), width = 0.5) +
  scale_fill_manual(values = c("D22V Daughter Predictions" = "blue", "D22V Simulated" = "grey")) +
  labs(title = "B - D22V Prevalence of PPD < 4mm, PPD >= 4mm",
       x = "Group",
       y = "Frequency",
       fill = "Type") +
  theme_minimal() +
  theme(legend.position = "top",
    axis.text.x = element_text(angle = 0, hjust = 0.5)  ) +
  geom_text(aes(label = Frequency), vjust = -0.3, position = position_dodge(.5), size = 3, color = "red")
grid.arrange(Hist_Unique_Val_D22DV, Hist_Group_Val_D22DV, ncol = 2, widths = c(1, 1))
unique_values <- unique(plot_data$Value)
chi_squared_results <- lapply(unique_values, function(val) {
  subset_data <- plot_data %>%
    filter(Value == val) %>%
    count(Type)
  if (nrow(subset_data) == 2) {
    chisq_test <- chisq.test(subset_data$n)
    list(value = val, counts = subset_data$n, p.value = chisq_test$p.value, statistic = chisq_test$statistic)
  } else {list(value = val, counts = subset_data$n, p.value = NA, statistic = NA)  }})
table_data_grouped <- table(plot_data_grouped$Group, plot_data_grouped$Type)
chisq_test_result_grouped <- chisq.test(table_data_grouped)
cat("Length of new data set:", NnS, "\n")
cat("Noise Level:", noise_level, "\n")
cat("Number of retraining rows:", firstRows, "\n")
Mother_model
cat("Mother Model RMSE:", Mother_model_rmse, "\n")
cat("Mother Model MAE:", Mother_model_mae, "\n")
cat("Mother Model MSE:", Mother_model_mse, "\n")
cat("Mother Model R-squared:", Mother_model_rsq, "\n")
imp.Mother_model.2V <- xgb.importance(model = Mother_model)
print(imp.Mother_model.2V)
Daughter_model
cat("New Data Daughter Model RMSE:", new_Daughter_rmse, "\n")
cat("New Data Daughter Model MAE:", new_Daughter_mae, "\n")
cat("New Data Daughter Model MSE:", new_Daughter_mse, "\n")
cat("New Data Daughter Model R-squared:", new_Daughter_rsq, "\n")
imp.Daughter_model.2V <- xgb.importance(model = Daughter_model)
print(imp.Daughter_model.2V)
cat("Best Kernel:", best_kernel, "\n")
cat("Best Bandwidth:", best_bw, "\n")
cat("Best Adjust:", best_adjust, "\n")
cat("Bootstrap test for kernel density estimates difference between D22V and Daughter Predictions:\n")
cat("Mean difference in densities:", mean(boot_result_Daughter$t), "\n")
ci_Daughter <- boot.ci(boot_result_Daughter, type = "perc")
cat("95% CI for the difference in densities:", ci_Daughter$percent[4:5], "\n")
ks_test_result <- ks.test(new_data_combined$D22V, new_data_combined$predicted_D22V)
```

```
cat("KS test statistic:", ks_test_result$statistic, "\n")
cat("KS test p-value:", ks_test_result$p.value, "\n")
cat("Chi-squared test results for each single value:\n")
for (result in chi_squared_results) {
  cat("Value:", result$value,
      " - Counts:", result$counts,
      " - Chi-squared statistic:", result$statistic,
      " - p-value:", result$p.value, "\n")}
cat("Chi-squared test result for grouped data:\n")
print(chisq_test_result_grouped)
```