UNIVERSIDADE ABERTA

Aplicação de Métodos de Estatística Espacial e Multivariada na Análise da Qualidade da Água no Sul do Mar do Norte

Christopher Ricardo Ody

Mestrado em Estatística, Matemática e Computação

Ramo em Estatística Computacional

2024

UNIVERSIDADE ABERTA

Aplicação de Métodos de Estatística Espacial e Multivariada na Análise da Qualidade da Água no Sul do Mar do Norte

Christopher Ricardo Ody

Mestrado em Estatística, Matemática e Computação

Ramo em Estatística Computacional

Dissertação orientada por

Professora Doutora Elisabete Teresa da Mata Almeida Carolino e

por Professora Doutora Maria do Rosário Olaia Duarte Ramos

2024

Aplicação de Métodos de Estatística Espacial e Multivariada na Análise da Qualidade da Água no Sul do Mar do Norte

Resumo

A região do Sul do Mar do Norte desempenha um papel vital tanto para a economia quanto para a sociedade dos países circunvizinhos. A análise da qualidade da sua água é um processo crítico que envolve a avaliação de parâmetros físicos, químicos e biológicos, essencial para garantir a sustentabilidade ambiental e a saúde das comunidades locais e ecossistemas marinhos. Utilizando métodos de Estatística Multivariada e Espacial, esta investigação busca identificar padrões e autocorrelações espaciais para avaliar a qualidade da água naquela região. Os dados utilizados foram coletados em cruzeiro científico realizado em dezembro de 2020 a bordo da embarcação RV Meteor, liderado por uma equipe de pesquisadores alemães. Os dados brutos passaram por pré-tratamento orientado pelo protocolo de Controle de Qualidade de Dados da SeaDataNet, um projeto internacional de oceanografia destinado a disponibilizar dados marítimos europeus. Foram realizados testes de pico e gradiente, além da padronização dos dados e imputação através de interpolação de ponderação pela distância inversa. Para um melhor entendimento da área estudada, os dados foram agregados por zonas para determinadas análises e, por vezes, foram considerados globalmente. Foi realizada uma análise exploratória de dados espaciais (AEDE) de modo a resumir suas principais características. Também realizou-se uma redução da dimensionalidade dos dados originais através da análise de componentes principais como ferramenta auxiliar à análise espacial. A autocorrelação espacial foi analisada através do cálculo da Estatística I de Moran global e local. As conclusões indicaram uma autocorrelação espacial significativa para todas as variáveis consideradas nas zonas de águas doce e um expressivo achatamento da amplitude das variáveis nas zonas de mar aberto, o que possivelmente ocasionou a inexistência de autocorrelação espacial significativa naquelas zonas.

Palavras-chave: Mar do Norte, Qualidade da Água, Análise Exploratória de Dados Espaciais, Análise de Componentes Principais e Autocorrelação Espacial

Application of Spatial and Multivariate Statistical Methods in the Analysis of Water Quality in the Southern North Sea

Summary

The Southern North Sea region plays a vital role for both the economy and society of the surrounding countries. Analyzing the quality of your water is a critical process that involves an assessment of physical, chemical, and biological parameters, essential to guarantee environmental sustainability and the health of local communities and marine ecosystems. Using Multivariate and Spatial Statistical methods, this investigation seeks to identify spatial patterns and autocorrelations to assess water quality in that region. The data used was taken on a scientific cruise carried out in December 2020 aboard the RV Meteor vessel, led by a team of German researchers. The raw data went through pre-treatment guided by the Data Quality Control protocol of *SeaDataNet*, an international oceanography project aimed at making European maritime data available. Spike and gradient tests were performed, in addition to data standardization and imputation through inverse distance weighting interpolation. For a better understanding of the scientific area, the data were aggregated by zones for certain analyses, and were sometimes considered globally. An exploratory spatial data analysis (AEDE) was carried out in order to summarize its main characteristics. A reduction in the dimensionality of the original data was carried out through principal component analysis as an auxiliary tool for spatial analysis. The Spatial autocorrelation was analyzed by calculating global and local Moran's I Statistics. The outcomes indicated a significant spatial autocorrelation for all variables considered in the freshwater areas and a notable range flattening of the variables in the open sea areas, which possibly caused the lack of significant spatial autocorrelation in those areas.

Keywords: North Sea, Water Quality, Exploratory Spatial Data Analysis, Principal Component Analysis and Spatial Autocorrelation

Dedicatória

À minha querida esposa Silvia Ody e
aos meus amados filhos Cristal e Ezekiel

AGRADECIMENTOS

Agradeço à Silvia Eula Ody, minha companheira de todas as horas, pelo apoio e paciência necessários na caminhada deste curso e no curso de nossas vidas.

Aos meus filhos, pelo constante aprendizado em conduzi-los nos caminhos do Senhor. Ter vocês em minha vida é motivo de seguir em frente sempre.

Agradeço à minha zelosa mãe, Maria Celia Fauth, pelo carinho e esforço em ter me proporcionado ensino de qualidade desde a minha juventude e por sempre ter acreditado em meu potencial.

Também agradeço a Erlindi e Eda Basha, por me encorajar a realizar este mestrado e pelo apoio financeiro provido.

Às minhas orientadoras Professora Doutora Elisabete Carolino e Professora Doutora Maria do Rosário Ramos, pelo encorajamento, dedicação, tempo e pelo olhar atento na orientação deste trabalho.

Aos professores e colaboradores do DCeT da Universidade Aberta em Portugal.

Aos colegas e amigos que conheci através deste mestrado, pelas contribuições acadêmicas, pelas discussões construtivas e por tornar esse curso mais prazeroso.

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.

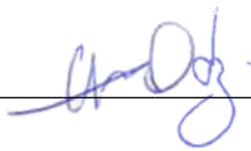
I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

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, 13 de março de 2024

Nome completo/Full name: Christopher Ricardo Ody

Assinatura/Signature:



ÍNDICE

INDICE DE TABELAS

INDICE DE FIGURAS

LISTA DE ABREVIATURAS, SIGLAS E ACRÔNIMOS

AEDE	Análise Exploratória de Dados Espaciais
AI	Zona "Águas Internas"
AU	Unidade de Absorção (*Absorbance Unit*)
PCs	Componentes Principais (*Principal Components*)
IQR	Amplitude Interquartil (*Interquartile Range*)
KNN	Matriz de k-vizinhos próximos (*k-nearest neighbours*)
LISA	Indicadores Locais de Associação Espacial (Local Indicators of Spatial Association)
MDS	Escalonamento Multidimensional (*Multidimensional Scaling*)
MN	Zona "Mar do Norte"
MT	Zona "Mar Territorial"
PCA	Análise de Componentes Principais (*Principal Component Analysis*)
TSG	Termossalinógrafo (*thermosalinograph*)

INTRODUÇÃO

O Mar do Norte é uma região vital para muitos países, incluindo a Alemanha, os Países Baixos, a Dinamarca e o Reino Unido. É um centro crucial para o comércio, a pesca e a produção de energia, envolvido significativamente na economia alemã e nas economias de outros países vizinhos do Mar do Norte. A importância da região levou a uma extensa exploração científica, com investigadores a estudar o ecossistema único da área e o impacto das atividades humanas no ambiente.

A análise da água é um componente crítico no entendimento dessas complexas questões, fornecendo informações sobre a qualidade da água, os níveis de nutrientes e a possível presença de poluentes. Estudos recentes mostram que a qualidade da água do Mar do Norte melhorou ao longo das últimas décadas, graças aos esforços para reduzir a poluição e melhorar o tratamento de águas residuais. John Huthnance e colaboradores mostraram que as alterações do dióxido de carbono (CO_2) e do pH são influenciadas pelos padrões de circulação, fluxo de entrada do Atlântico, condições climáticas locais e propriedades das massas de água componentes (Huthnance J, 2016). Também demostraram mudanças de nutrientes influenciadas pelos padrões de circulação, e as mudanças na concentração de oxigênio, especialmente perto do fundo do mar, influenciadas pela respiração da matéria orgânica, decomposição da matéria orgânica, falta de suprimento de oxigênio e temperatura. No entanto, ainda há muito a aprender sobre o ecossistema da região e a investigação contínua é essencial para garantir a sua sustentabilidade a longo prazo.

Estudos sobre salinidade e temperatura, dentre outras variáveis, são constantemente realizados na área do Mar do Norte. Um importante estudo feito nessa área foi realizado sobre o evento oceanográfico conhecido como "*Great Salinity Anomaly*" (GSA). Durante os anos 1960 e 1970, houve uma mudança significativa nos padrões de salinidade na superfície do oceano Atlântico Norte, particularmente na região do Mar do Norte e do Mar da Noruega. A anomalia de salinidade foi caracterizada por uma grande acumulação de água doce na superfície do oceano, resultando em uma diminuição acentuada na salinidade da água. Essa água doce foi principalmente proveniente do derretimento acelerado do gelo marinho e dos glaciares que ocorreu devido a condições

climáticas específicas, como padrões de vento e temperatura (Belkin, Levitus, Antonov, & Malmberg, 1998). Este evento foi ainda revisitado por outros autores mais recentemente (Kim, Yeager, & Danabasoglu, 2022), sugerindo outras dinâmicas para as causas do evento, enfatizando não propriamente as variáveis como salinidade, temperatura e pH, mas para os eventos de fluxo de calor superficial mostrando que esses possuem um papel dominante na supressão da convecção e no frescor no interior do Mar. Estes indícios indicam a amplidão da complexidade do ecossistema local e dos respectivos fatores que determinam a composição da água.

A análise espacial possui algumas nuances que algumas vezes não podemos perceber ao analisar um mapa. Almeida (2012) cita que é necessário ter cautela, uma vez que o olho humano é treinado para buscar padrões e estruturas em todos os aspectos da realidade. Então, uma análise visual de um mapa acaba sendo um instrumento viesado. É preciso aplicar um teste estatístico que examine de maneira abrangente se a distribuição espacial da variável em análise é aleatória (Almeida, 2012). Veremos a seguir que na análise de processos espaciais, a interação entre heterogeneidade e dependência espacial apresenta desafios significativos na especificação de modelos espaciais, tornando o processo de especificação demorado e propenso a erros, levando potencialmente a modelos inadequados. Uma ferramenta eficaz para definir tais modelos é a Análise Exploratória de Dados Espaciais (AEDE), sendo essencial para entender melhor os dados antes da modelagem. Essa abordagem preliminar ajuda a garantir uma modelagem econométrico-espacial mais precisa, seguindo a prática recomendada de explorar os dados antes de proceder com análises mais sofisticadas (Anselin, 1988).

Ainda há estudos que destacam a importância da análise espacial e das técnicas estatísticas na compreensão dos padrões de salinidade e temperatura e seus efeitos sobre o ambiente marinho e a vida marinha, ainda que estes não se concentrem especificamente no Mar do Norte. Zhiyi Fu e colaboradores sugerem a utilização de um método para estimação da salinidade da superfície do mar e na análise de sua heterogeneidade espaço-temporal no Golfo do México utilizando uma abordagem espaço-temporal denominada Cubist (Fu, et al., 2021). O modelo foi utilizado para estimar a salinidade da superfície do mar e explora relações não lineares nos dados observados, ajusta variáveis preditas por

meio de equações lineares compostas sob regras geradas pelo modelo. Este método, segundo os autores, conseguiu refletir a distribuição gradiente de salinidade e analisar a heterogeneidade espacial-temporal na região. A técnica provou ser eficaz para estimar a salinidade com alta precisão e pode ser ideal para análises semelhantes em áreas costeiras com ambientes geográficos similares. Ainda, voltado a atividades pesqueiras no Mar do Norte, destaca-se o estudo de Anna Akimova e colaboradores (2016) que investiga a relação entre a hidrografia do Mar do Norte (temperatura e salinidade) e a variabilidade de estoques de peixes comerciais (Akimova et al., 2016). O estudo utiliza a técnica de análise espacial chamada de correlação cruzada espacialmente resolvida e revela regiões onde variáveis ambientais influenciam significativamente os estoques de peixes. As principais conclusões indicam correlações negativas entre a temperatura e o recrutamento de bacalhau e solha, e correlações positivas entre a biomassa de desova de arenque e a temperatura. Também foi encontrada uma forte correlação positiva entre as variáveis de estoque de espadilha e a salinidade no centro do Mar do Norte. Esses resultados fornecem informações valiosas para a gestão pesqueira, incorporando variáveis ambientais em modelos de estoque-recrutamento.

Os estudos mencionados destacam o papel importante da análise espacial e das metodologias estatísticas para entender os padrões de salinidade e temperatura, bem como seus impactos no ambiente marinho e nos organismos que nele vivem. Estes estudos exemplificam como métodos estatísticos espaciais podem ser empregados para investigar a dinâmica oceânica e as respostas dos ecossistemas marinhos a variações ambientais.

O monitoramento das condições geoambientais do Mar do Norte, realizado por organizações como a Agência Europeia do Ambiente (EEA), o Conselho Internacional para a Exploração do Mar (ICES), a Agência Ambiental do Reino Unido, e a Direção Norueguesa de Assuntos Marítimos, é essencial para a proteção da rica biodiversidade marinha, a sustentabilidade das atividades comerciais como navegação, pesca e exploração de recursos naturais, e a mitigação de impactos na agricultura costeira. Este monitoramento ajuda a identificar e mitigar ameaças ambientais, garantindo a preservação dos recursos marinhos, a prevenção de desastres ambientais e o suporte a práticas econômicas sustentáveis, essenciais para o bem-estar das comunidades costeiras e para a conservação

dos ecossistemas marinhos (Dronen & Gjengedal, 2020). O conhecimento geoespacial estatístico de variáveis de análise de águas no Mar do Norte é ainda motivado por uma série de razões críticas, incluindo as citadas anteriormente. Além de monitorar a qualidade da água para a saúde pública e a sustentabilidade da vida marinha, o constante monitoramento dessas variáveis avalia os impactos das mudanças climáticas e da atividade humana na região, como poluição e exploração de recursos naturais. Além disso, fornecem dados essenciais para a tomada de decisões políticas e econômicas informadas, visando a sustentabilidade ambiental e o desenvolvimento econômico da região (Walday & Kroglund, 2008; Huthnance J, 2016).

Navios de cruzeiro científicos desempenham um papel crucial nesta investigação, fornecendo uma plataforma para os cientistas realizarem experiências e recolherem dados. Estas embarcações estão equipadas com laboratórios e equipamentos de amostragem de última geração, permitindo aos investigadores analisar amostras de água e estudar o ecossistema único do Mar do Norte.

O projeto TRAM (sigla em inglês para *Tracing origin and distribution of geogenic and anthropogenic dissolved and particulate critical high-technology metals in the southern North Sea*), liderado pela Dra. Andrea Koschinsky da faculdade Jacobs na Alemanha, foi um cruzeiro de pesquisa que ocorreu em 2020 que teve por objetivo de estudar insumos antropogênicos de contaminantes metálicos críticos emergentes, como elementos de terras raras, Sc, Ga, Ge, Pt, Zr, Ti, Mo e V, dos rios alemães Elba, Weser e Ems para o oceano (TRAM, 2024). Em sua viagem, a área de trabalho compreendeu os estuários desses rios, desde o membro final de água doce até o membro final de água do mar ao longo do gradiente de salinidade, bem como a dispersão da pluma ao longo das correntes predominantes, majoritariamente em direção ao leste.

Este trabalho tem por objetivo utilizar técnicas estatísticas multivariadas e geoespaciais para obter indicadores que permitam qualificar as variáveis obtidas a bordo do *RV Meteor* e procurar padrões de semelhança entre elas na área estudada. Para tanto, foram analisados conjuntos de dados espaciais de variáveis distintas na tentativa de localizar padrões específicos e caracterizar o perfil oceanográfico na região. As técnicas referidas nesta introdução são utilizadas em dois grandes campos de estudo estatístico

espacial, a Econometria e a Oceanografia Física. As análises espaciais de características das águas, como salinidade e temperatura, geralmente não são categorizadas sob econometria, mesmo que possam empregar métodos estatísticos similares aos usados em econometria espacial para análise de dados. Estas análises pertencem ao campo da oceanografia física, que é uma disciplina na ciência oceânica focada no estudo das propriedades físicas e processos dinâmicos dos oceanos. A oceanografia física utiliza métodos de análise espacial e temporal para entender a circulação oceânica, as interações entre o oceano e a atmosfera, a distribuição de temperatura e salinidade, entre outros aspectos físicos do sistema oceânico (Knauss, 2005). Embora técnicas de modelagem e análise estatística sejam comuns em ambas as disciplinas, a aplicação e o contexto dessas técnicas podem ser distintas, refletindo os objetivos específicos de cada campo de estudo, o que será tomado em consideração nesse trabalho.

Estruturalmente essa dissertação está organizada em duas partes. A primeira parte, dividida em três capítulos, apresenta o referencial teórico e metodológico da Econometria Espacial utilizado no trabalho. No primeiro capítulo definem-se os métodos da Análise Exploratória de Dados Espaciais (AEDE), a Autocorrelação Espacial e a Heterogeneidade Espacial. O segundo capítulo introduz as matrizes de ponderação espacial seguida da definição da Estatística I de Moran Global e Local. No terceiro capítulo é apresentada a Análise de Componentes Principais (PCA).

Na segunda parte apresenta-se o estudo prático realizado sobre os dados do Mar do Norte e são discutidos os resultados do estudo. Assim, o quarto capítulo descreve o pré-tratamento dos dados, variáveis do estudo e método de coleta dos dados. No quinto capítulo se apresentam os resultados iniciais da AEDE e algumas interpretações iniciais dos resultados. No sexto capítulo é utilizada a PCA como ferramenta auxiliar para entendimento da variabilidade dados globais e redução de variáveis. No sétimo capítulo são apresentados os resultados e interpretações da Estatística I de Moran global e local (LISA) para as diferentes variáveis do conjunto de dados, bem como para os dois componentes principais calculados. Os resultados são então discutidos à luz do enquadramento teórico e metodológico utilizado. Na conclusão, são revisitados e discutidos os resultados análises, além das limitações do estudo e perspectivas futuras.

Os dados brutos utilizados nesse estudo passaram por pré-tratamento orientado em sua preparação para o uso na aplicação das técnicas selecionadas para este estudo, a PCA e a análise da autocorrelação espacial. As observações do conjunto de dados foram em parte agregadas por zonas para determinadas análises ou por vezes consideradas globalmente. O pré-tratamento e as análises foram feitos através da linguagem R versão 4.3.2, sendo os principais pacotes utilizados e suas versões:

- *dplyr* (1.1.4): utilizado para manipulação de *dataframes*.
- *ggplot2* (3.4.4), *tmap* (3.3-4), *plotly* (4.10.3) *e leaflet* (2.1.1): utilizados para plotagem de mapas.
- *rgdal* (1.6-7) *e spdep* (1.3-1): ferramentas para manipulação e análise de dados espaciais.
- *gstat* (2.1-1): utilizado no pré-tratamento de dados para interpolação dos dados faltantes.
- *Corrplot* (0.92) *e FactoMineR* (2.9): utilizados nos correlogramas e na Análise de Componentes Principais

PARTE 1 – REFERENCIAL TEÓRICO E METODOLÓGICO

Este capítulo fornece um enquadramento teórico essencial para a compreensão dos métodos estatísticos utilizados neste estudo, a Análise Exploratória de Dados Espaciais (AEDE) a Análise de Componentes Principais (PCA) e o Índice de Moran I, e como estes foram aplicados para buscar identificar padrões nas principais componentes do conjunto de dados estudado.

A Análise Exploratória de Dados Espaciais (AEDE) é definida como o conjunto de técnicas que descrevem e visualizam distribuições espaciais, identificam locais atípicos, descobrem esquemas de associação (autocorrelação espacial) e sugerem estruturas no espaço geográfico (heterogeneidade espacial) (Anselin, 1988). Logo, a AEDE é mais uma técnica descritiva (estatística) do que confirmatória (Sicilia, Rivera, & Navarro, 2017). Nesse sentido, reafirma-se que a análise exploratória dos dados é o estudo prévio útil à análise confirmatória dos dados espaciais, no qual são formulados modelos de regressão e realizada a estimação dos parâmetros amostrais.

A Análise de Componentes Principais (PCA) é uma técnica para reduzir a complexidade dos dados, combinando variáveis correlacionadas em um conjunto reduzido de variáveis não correlacionadas, conhecidas como componentes principais, através de uma combinação linear, preservando o máximo possível da variância dos dados originais. Este método é crucial para simplificar a complexidade dos dados, permitindo uma análise mais eficiente e a identificação de estruturas subjacentes de menor dimensão.

Em adição, o outro método empregado nesse estudo, o Índice I de Moran, é uma medida de autocorrelação espacial utilizada para avaliar se um padrão é agrupado, disperso ou aleatório em relação a uma localização espacial. Este índice é fundamental para entender a distribuição espacial de variáveis e identificar possíveis áreas de interesse para investigações mais detalhadas (Almeida, 2012).

A combinação desses métodos neste estudo permite uma análise robusta e detalhada dos componentes principais selecionados, facilitando a identificação de possíveis padrões espaciais significativos que poderiam permanecer ocultos. As aplicações potenciais desses resultados são vastas, abrangendo desde a otimização de recursos ambientais e a gestão sustentável de ecossistemas até o planejamento pesqueiro e

regional. Por exemplo, em estudos ambientais, essa abordagem pode ser usada para identificar áreas de alta poluição ou degradação ambiental, orientando políticas de conservação.

1. A ANÁLISE EXPLORATÓRIA DE DADOS ESPACIAIS (AEDE)

1.1. Definição e Exemplos da AEDE

A Análise Exploratória de Dados Espaciais (AEDE) engloba uma série de técnicas destinadas à descrição e análise de distribuições espaciais, bem como à identificação de *outliers* e padrões de associação espaciais (*clusters*), respectivamente (Anselin, 1988). A AEDE é uma ferramenta técnica que conduz o analista no auxílio na etapa posterior de especificação dos modelos espaciais, semelhante a uma análise exploratória convencional, que também serve como ponto de partida antes de seguir para uma análise confirmatória ou modelagem econométrica (Almeida, 2012).

Para fins deste trabalho, a AEDE será apresentada numa perspectiva de dados espaciais na forma de pontos. Porém, a AEDE pode ser também estudada em polígonos em uma perspectiva geoestatística, com um conjunto de ferramentas para descrever os dados espaciais deste tipo (Almeida, 2012).

Um método inicial de análise descritiva pode envolver a criação de mapas que categorizam as unidades observadas em intervalos de valores para uma variável específica. Uma proposta para o atual estudo foi a separação da área navegada em 3 zonas: Águas Internas (verde), Mar Territorial (vermelho) e Mar do Norte (azul). As zonas, bem como os pontos tomados em cada uma, são exibidos na Figura 1. Já a Figura 2 abaixo exibe um gráfico de *boxplot* para as unidades de observação consideradas em cada zona, seguido pela Figura 3 de estatísticas descritivas para as diversas variáveis tomadas na área das Águas Internas.

Figura 1: Zonas Demarcadas Para o Estudo e seus respectivos pontos

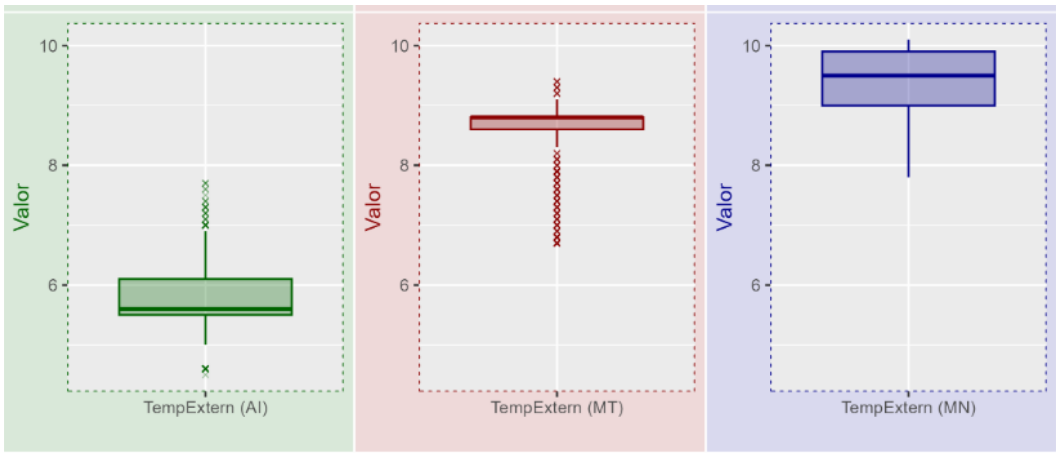


Figura 2: Exemplo de gráfico de boxplot por zona

Salinity	TempExtern	N.NO3
Min. : 0.509	Min. :4.500	Min. :0.0006
1st Qu.: 7.207	1st Qu.:5.500	1st Qu.:1.3959
Median :11.362	Median :5.600	Median :1.7859
Mean :12.402	Mean :5.787	Mean :1.7135
3rd Qu.:16.209	3rd Qu.:6.100	3rd Qu.:2.1411
Max. :32.646	Max. :7.700	Max. :6.9791

Abs210	Abs254	Abs360
Min. :0.9493	Min. :0.1717	Min. :0.0986
1st Qu.:3.5562	1st Qu.:0.7850	1st Qu.:0.4563
Median :3.9920	Median :0.9726	Median :0.5895
Mean :3.9462	Mean :1.1817	Mean :0.7490
3rd Qu.:4.4436	3rd Qu.:1.6167	3rd Qu.:1.0563
Max. :5.8984	Max. :3.5527	Max. :2.4934

Figura 3: Resumo estatístico de algumas variáveis na zona Águas Internas

Ao analisarmos as distribuições da variável *TempExt* (temperatura da água) é possível notar o aumento gradual da temperatura na direção do alto Mar do Norte.

Um dos componentes mais relevantes na AEDE é a análise gráfica, onde podemos visualmente começar a entender os dados no espaço. Entretanto, não temos ainda uma forma de medir os agrupamentos espaciais e como os diferentes pontos no espaço se correlacionam de maneira consistente. Esta análise inicial pode então ser aliada a técnicas de análise estatística para mensurar a dependência (autocorrelação) e a heterogeneidade espaciais (Sicilia, Rivera, & Navarro, 2017). Estas serão discutidas no próximo tópico.

1.2. A Autocorrelação Espacial e a Heterogeneidade Espacial

Por vezes é necessário modelar a distribuição espacial de algum fenômeno matematicamente, seja em municípios ou estados de um país, ou nos quadrantes de uma área de estudo, ou ainda como um mapa com localizações pontuais de captação de dados. Para Arselin (1988), duas questões básicas devem ser consideradas pelo analista ao realizar seu modelo. A primeira é se o padrão espacial apresentado pelo fenômeno é significativo em algum sentido, portanto, vale a pena ser interpretado. Se assim o for, a segunda pergunta é se é possível obter alguma informação sobre os processos que produziram o padrão observado a partir de uma análise da distribuição mapeada do fenômeno.

O autor ilustra sua técnica exemplificando um caso de cólera ocorrido em Londres em 1848-9 e apresentando uma motivação para a autocorrelação espacial através deste. Conhecendo que o abastecimento de água contaminada é a principal causa da transmissão da cólera e que o tempo de sobrevivência de seu micróbio na água aumenta à medida que a composição química da água aumenta em basicidade. Logo, levanta, portanto, a questão geral de saber se podemos relacionar áreas de alta/baixa incidência de cólera com a qualidade variável do abastecimento público de água. Conhecendo as áreas da metrópole atendidas pelas diversas companhias de água em 1849 e os locais de seus sistemas de abastecimento de água e reservatórios, os mapas exibiam claramente uma alta incidência de mortes nas áreas atendidas pela companhia de Água *Southwark & Vauxhall*. Uma consulta pública foi feita para análise química de tais águas e descobriu-se que a empresa possuía os mais altos níveis de matéria orgânica em suas águas, juntamente com os mais altos níveis de cal solúvel (base). Com essas informações e dados foi possível desenhar mapas e determinar a correlação espacial entre as casas atendidas pela empresa e os altos casos da doença. Ainda, quando levamos em consideração o fato de que outra companhia de água que acidificou sua água antes de ser distribuída na mesma época e que forneceu água em casas no mesmo distrito, que a cota de seus locais em relação ao rio é idêntica e que os habitantes abastecidos por ambas as empresas foram exatamente semelhantes no que diz respeito aos meios apresentou cerca de 2000 mortes a menos em sua rede que a Companhia Southwark. Esse exemplo ilustra de maneira clara algumas particularidades da correlação espacial, com uma instigante causalidade química a variável número de mortes.

Além das mencionadas particularidades, Anselin (1988) ainda destaca alguns efeitos espaciais relevantes. O autor sugere que estes efeitos são a razão essencial para a existência de um campo separado de econometria espacial. A autocorrelação (ou dependência) espacial e a heterogeneidade espacial são os dois aspectos de dados e modelos que merecem especial atenção do ponto de vista metodológico.

A autocorrelação espacial ocorre quando os dados coletados desafiam a ideia de que são independentes uns dos outros. Essencialmente, ela é entendida como a ausência de independência que está geralmente presente entre as observações de

variáveis em conjuntos de dados transversais, como a deste estudo. Isso acontece porque a dependência espacial leva em conta a proximidade ou a localização relativa das observações, destacando que a distância entre elas pode influenciar sua relação. Nas vezes em que a noção de espaço é estendida para além do sentido euclidiano estrito, pode ainda incluir os espaços como político, redes sociais, ou distância interpessoal, entre outros (Isard, 1969). A falta de autocorrelação espacial é conhecida como aleatoriedade espacial ou independência espacial. Em termos técnicos, diz-se que há ausência de autocorrelação espacial quando os valores observados em uma variável não exibem um padrão discernível de dependência espacial, ou seja, a distribuição dos valores é aleatória no espaço. Nestes casos, os valores observados numa determinada área não dependem dos valores observados nas áreas vizinhas, e o padrão espacial observado naquele contexto é tão provável quanto qualquer outro padrão espacial e a localização de valores de variáveis pode ser alterada sem perturbar o conjunto da informação definida naquele cenário (Kopczewska, 2021). Na presença de autocorrelação espacial, pode-se falar em autocorrelação positiva e negativa. Positiva significa que os valores em determinado espaço e suas áreas vizinhas são semelhantes. No caso de autocorrelação espacial negativa, áreas vizinhas diferem, mais do que pareceria em uma distribuição aleatória. Isso é chamado de “padrão xadrez”, pelo contraste em seu tabuleiro (Kopczewska, 2021).

Arselin (1988) discute que a dependência espacial pode ser causada por uma variedade de problemas de medição frequentemente encontrados em aplicações. O autor cita exemplos como a delimitação arbitrária de unidades espaciais de observação, problemas de agregação espacial e ressalta a presença de externalidades espaciais e efeitos colaterais (*spill-over*). Além disso, e bastante independente desta questão de medição, a organização espacial inerente e a estrutura espacial dos fenômenos tenderão a gerar padrões complexos de interação e dependências que são de interesse em si e ainda diferentes de padrões temporais. Elhorst cita como exemplo que duas unidades geográficas podem afetar-se mutuamente, ao passo que duas observações no tempo não o podem e que outro complicador é a grande variedade de unidades de medida utilizadas para modelar a dependência espacial (vizinhos, distância, links, etc.) em comparação com a medição da dependência temporal (Elhorst, 2014).

Assim, Anselin (1988) descreve a autocorrelação espacial como a relação entre valores de uma variável que estão localizados próximos no espaço geográfico, ou seja, é a medida de similaridade entre valores observados em diferentes locais e como essa similaridade diminui ou aumenta com a distância entre esses locais.

O segundo tipo de efeito espacial descrito por Anselin (1988), a heterogeneidade espacial, relaciona-se à falta de estabilidade das relações comportamentais no espaço ou outras relações em estudo. Isto pode ocorrer em modelos econométricos estimados com base num conjunto de dados de unidades espaciais diferentes como, por exemplo, regiões ricas em determinada área de um país e regiões pobres em outra. Em contraste com o caso da dependência espacial, os problemas causados pela heterogeneidade espacial podem, na sua maior parte, ser resolvidos através de técnicas econométricas padrão, segundo o autor.

Em suma, as relações estruturais que mudam com a localização do objeto são relacionadas à heterogeneidade espacial. Já a dependência espacial é um conceito que se refere à situação em que o valor de uma variável em um local é influenciado pelos valores dessa mesma variável em locais próximos. Ela implica que há uma correlação entre os dados de diferentes locais geográficos, e essa correlação tende a diminuir à medida que a distância entre os locais aumenta. Kopczewska discute ainda que evidências práticas mostram que a heterogeneidade e a dependência espaciais não são totalmente distinguíveis. (Kopczewska, 2021).

2. A ESTATÍSTICA *I* DE MORAN

A análise geoespacial enfatiza a importância fundamental da autocorrelação espacial, que mede a semelhança de objetos ou atividades em locais próximos na superfície terrestre. Refletindo sobre a conhecida primeira lei da Geografia de Tobler, que postula que coisas próximas são mais relacionadas do que aquelas distantes (Tobler, 1970), essa noção é incontornável na geografia, sugerindo que fenômenos espaciais são interdependentes, proporcionando padrão, previsibilidade e ordem (Miller, 2004). Este conceito serve tanto como um índice descritivo da distribuição espacial quanto um

processo causal, evidenciando a influência de uma localidade sobre suas vizinhanças. (Goodchild M. , 1986).

Os avanços tecnológicos e metodológicos recentes que ampliaram as capacidades de análise e interpretação de dados espaciais com o desenvolvimento de softwares de Sistemas de Informações Geográficas (SIG) mais sofisticados, o aumento da disponibilidade de dados geoespaciais de alta resolução e a integração de métodos estatísticos avançados destacam a importância de análises espaciais de correlação. Sua utilização empodera o analista com informações críticas na modelagem preditiva e na tomada de decisões baseada em evidências em diversas áreas de pesquisa e aplicações (Goodchild M. , 2007).

O I de Moran é uma estatística de autocorrelação espacial, projetada para medir a correlação de um fenômeno em relação ao espaço geográfico. Este teste é fundamental na análise geoespacial, uma vez que permite ao analista avaliar se o padrão espacial de determinada variável é agrupado, disperso ou aleatório em relação à sua localização geográfica (Li, Calder, & Cressie, 2007). A importância dessa análise reside na sua capacidade de revelar padrões ocultos em dados espaciais, o que é crucial para uma ampla gama de aplicações, desde a epidemiologia até a economia regional, passando pela ecologia e além (Getis, 2008).

Historicamente, o Teste de Moran I foi proposto por Patrick Moran em 1950, marcando um avanço significativo na análise espacial. O teste surgiu em um contexto em que os cientistas buscavam métodos mais robustos para entender a estrutura espacial dos fenômenos naturais e humanos. Desde então, ele se tornou uma ferramenta essencial em estudos geoespaciais, permitindo análises mais precisas da distribuição espacial e da dependência espacial entre as observações (Li, Calder, & Cressie, 2007).

No contexto da análise de águas, especialmente em áreas de mares e oceanos, o Índice I de Moran permite a identificação de padrões espaciais na qualidade da água, na distribuição de contaminantes e na biodiversidade marinha. Essa análise pode contribuir significativamente para a gestão sustentável dos recursos hídricos, a conservação dos ecossistemas marinhos e a prevenção de riscos ambientais. Estudos amplamente reconhecidos nessa área incluem análises de dispersão de poluentes, mapeamento de

habitats críticos para a conservação marinha e monitoramento da acidificação dos oceanos (Rigby, Barber, & Burt, 2009).

2.1. Definindo a Matriz de Pesos Espaciais

A definição da matriz de pesos espaciais ou Matriz de Ponderação Espacial (W) é um passo crucial no cálculo do índice de I de Moran, pois ela especifica as relações espaciais entre as unidades de análise no estudo. A matriz de pesos espaciais descreve como as diferentes localizações estão conectadas umas às outras, influenciando diretamente a análise da autocorrelação espacial que resume a estrutura de correlação espacial. Tais matrizes levam em consideração pesos que são construídos utilizando a lógica de que unidades geograficamente mais próximas possuem pesos maiores do que as se encontram mais distantes (Tyszler, 2006).

Citando Almeida (2012), "uma matriz de ponderação espacial é uma matriz quadrada de dimensão n x n. Os pesos espaciais w_{ij} representam o grau de conexão entre as regiões segundo algum critério de proximidade, mostrando a influência da região j sobre a região i. Assim, a matriz W é útil por realizar uma espécie de ponderação da influência que as regiões exercem entre si."

Existem alguns métodos para definir a matriz de pesos, cada um refletindo diferentes concepções de proximidade ou conexão espacial entre as observações. Anselin (1988) detalha cada método com suas vantagens e limitações, e enfatiza que a escolha do método adequado depende da natureza dos dados e dos objetivos específicos da análise.

No método de contiguidade, os pesos são atribuídos com base na contiguidade geográfica das unidades. Se duas unidades *i* e *j* compartilham uma fronteira comum, o peso é geralmente definido como 1; caso contrário, é 0. Este método é particularmente útil em análises de dados de área, como bairros, municípios ou países. Com base nesse conceito, temos o valor 1 atribuído a regiões vizinhas e 0 para os demais casos (Almeida, 2012). Logo:

$$w_{ij}(k) = \begin{cases} 1, \text{se } i \; e \; j \; são \; contíguos \\ 0, se \; i \; e \; j \; não \; são \; contíguos \end{cases} \quad (1)$$

e, por convenção, é presumido que $w_{ii} = 0$, ou seja, a região não é considerada vizinha de si própria, implicando que a ser construída a partir desse critério, a matriz de contiguidade,

possua zeros em sua diagonal principal. Almeida (2012) acrescenta que o modelo pode parecer simples, mas há espaços para formar distintas convenções. Alguns modelos conhecidos para definição de vizinhança são baseados nos movimentos das peças de xadrez, conforme Figura 4. Os quadrantes em negrito são considerados vizinhos de A, B e C, respectivamente.

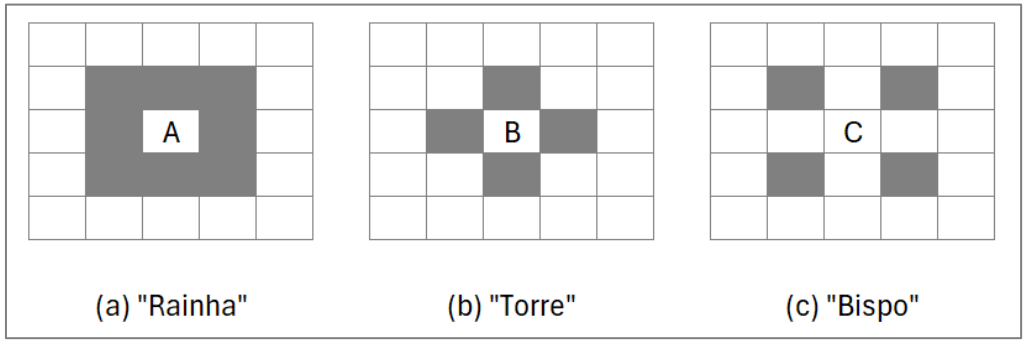


Figura 4: Convenções de definição de vizinhança na matriz de contiguidade

Um benefício compartilhado por todas as matrizes de pesos espaciais baseadas em contiguidade é a capacidade de estabelecer, de forma mais precisa, contiguidades de ordens superiores. Aqui nos referimos como 'ordens superiores' a consideração de vizinhos dos vizinhos. Por outro lado, a matriz binária de contiguidade apresenta uma desvantagem relacionada ao equilíbrio da conectividade. Isso ocorre porque algumas áreas podem ter uma grande quantidade de vizinhos, enquanto outras podem ter apenas alguns, resultando em uma distribuição desigual de conexões (Almeida, 2012).

Além da contiguidade, outro critério que é particularmente interessante é a definição dos pesos espaciais através da distância geográfica. Desta forma, os pesos podem ser baseados na distância entre as unidades, com pesos decrescendo à medida que a distância aumenta (Almeida, 2012).

Existem formas diferentes de especificar os pesos baseados em distância. Uma forma simples é a chamada matriz dos k-vizinhos mais próximos, $w_{ij}(k)$:

$$w_{ij}^{*}(k) = \begin{cases} 0, & \text{se } i = j \\ 1, & \text{se } d_{ij} \leq d_i(k) \\ 0, & se\ d_{ij} > d_i(k) \end{cases} \quad (2)$$

sendo $d_i(k)$ conhecida como a distância de corte dada pela distância do vizinho de ordem k (Tyszler, 2006), ou seja, a distância mínima para que sejam obtidos o número de vizinhos selecionado.

O benefício da abordagem de k-vizinhos próximos é corrigir o desbalanceamento na conectividade de uma matriz, garantindo que todas as unidades espaciais possuam um número igual de vizinhos. Além disso, essa matriz assegura a ausência de "ilhas", ou seja, regiões isoladas sem nenhum vizinho (Almeida, 2012).

Outra forma de especificar os pesos é baseá-los na distância inversa dos vizinhos mais próximos (KNN), de modo que unidades mais próximas têm pesos maiores:

$$w_{ij}^*(k) = \begin{cases} 0, & \text{se } i = j \\ \dfrac{1}{d_{ij}}, & \text{se } d_{ij} \leq d_i(k) \\ 0, & se\ d_{ij} > d_i(k) \end{cases} \tag{3}$$

onde d_{ij} é a distância entre o ponto e os vizinhos mais próximos ou iguais a $d_i(k)$, que continua sendo a distância de corte, dada pelo vizinho de ordem k (Tyszler, 2006). Utilizando-se este critério, os vizinhos influenciarão de forma diferente a região i, segundo o inverso de sua distância.

Para simplificar os cálculos, a matriz W é obtida após a normalização de $w_{ij}^*(k)$ nas linhas da matriz:

$$w_{ij}(k) = \frac{w_{ij}^*(k)}{\sum_j w_{ij}^*(k)} \tag{4}$$

Almeida (2012) discute a relevância da normalização da matriz W. Primeiramente, ter uma matriz normalizada por linhas significa que a soma dos elementos de cada linha e coluna é restrita a um valor finito n, correspondendo ao tamanho da amostra. Em segundo lugar, facilita a compreensão da média dos valores das variáveis dos vizinhos em relação à defasagem espacial. Essa média é importante para definir o conceito de defasagem espacial, tanto para a variável dependente (y) quanto para as variáveis explicativas (X) e os termos de erro teórico e residual (ε e e). Desse modo, a normalização contribui para suavizar os dados espacialmente. Tyszler (2006) acrescenta que normalizar

a matriz W ajuda a entender o peso espacial w_{ij} ao demonstrar a porcentagem do impacto total que a região j tem sobre a região i.

O analista precisa ser crítico em relação a escolha pela normalização de dados. Dependendo do modelo, Anselin (1988) ressalta considerar se os valores originais não normalizados são relevantes em termos comparativos e, se assim o forem, os dados devem ser utilizados em suas quantidades originais. O autor também adverte que a ideia de diminuição da influência com o aumento da distância geográfica entre regiões, um conceito intrínseco à matriz de distância inversa, pode se diluir ao normalizar a matriz. A normalização por linha pode ser inadequada nesse contexto, já que o efeito de atenuação da distância se perde. Nessa situação específica, é preferível utilizar a matriz sem a normalização por linha (Anselin, 1988).

A métrica a ser adotada para definição de distância também é relevante. Três métricas para distância são apresentadas a seguir (Almeida, 2012).

A distância euclidiana, onde u_i é a ordenada e v_i a abscissa:

$$d_{ij} = \sqrt{[(u_j - u_i)^2 + (v_j - v_i)^2]} \tag{5}$$

A distância Manhattan:

$$d_{ij} = |u_i - u_j| + |v_i - v_j| \tag{6}$$

E a distância do grande círculo, que representa a mínima distância entre dois pontos em um trajeto na superfície esférica (Almeida, 2012). Esta distância é calculada da seguinte forma:

$$d_{ij} = Rcos^{-1}[sen\phi_i sen\phi_j + cos\phi_i cos\phi_j cos(\mathrm{k_i} - \mathrm{k_j})] \tag{7}$$

Onde R o raio da Terra em torno do Equador (6378Km); ϕ e k são a latitude e a longitude.

No uso prático, a métrica euclidiana é empregada para calcular distâncias a partir das coordenadas. Em contrapartida, a métrica do arco é utilizada quando se dispõe de dados de latitude e longitude para estabelecer a distância entre regiões na superfície de uma esfera, como a Terra (Almeida, 2012).

Outros métodos são ainda apresentados na literatura como limiar de distância (binarização), onde pesos são definidos como 1 se a distância entre duas unidades é menor que um limiar especificado, e 0 caso contrário. Ainda podemos citar a função exponencial ou gaussiana, que atribui pesos que diminuem de forma mais suave à medida que a distância aumenta, baseando-se em uma função exponencial ou gaussiana da distância. Há ainda outro método de atribuição é conhecido por Kernel e utiliza funções kernel para ponderar as unidades com base na distância, permitindo que os pesos diminuam de forma suave conforme a distância aumenta, sem um corte abrupto. Isso permite uma modelagem mais flexível das relações espaciais (Anselin, 1988).

Almeida (2012) ainda lista um conjunto de propriedades desejáveis para as matrizes de peso:

- $0 \leq w_{ij} < \infty$: não existe distâncias negativas nem infinitas.
- $\sum_j w_{ij}^* \neq 0$: A matriz não deve possuir as chamadas "ilhas" (pontos ou conjunto de pontos isolados) especialmente quando se considera que a ilha em questão interage de alguma forma com as demais regiões.
- $w_{ij} = 0$: conforme mencionado anteriormente, uma região não é considerada sua própria vizinha. Logo, a diagonal principal da matriz de pesos é preenchida por zeros.
- $E(w_{ij}, \varepsilon) = 0$: A matriz é exógena ao fenômeno estudado. Em outras palavras, a distância pode afetar a salinidade em dois pontos, mas a salinidade não afetará em nenhuma hipótese a distância. Caso contrário, haverá inconsistência nos resultados.

Na área de econometria espacial, é reconhecido que a seleção de uma matriz de pesos espaciais específica pode levar a escolhas *ad-hoc* pelos pesquisadores. Contudo, existem abordagens mais comuns e predominantes para a construção e escolha dessas matrizes. Almeida (2012) recomenda que os pesquisadores utilizem procedimentos específicos para reduzir a arbitrariedade na escolha da matriz W e observem as propriedades desejadas mencionadas anteriormente. Entre esses procedimentos, destaca-se o método proposto por Baumont (2004), que visa identificar uma matriz W capaz de

capturar a maior parte da dependência espacial existente. Além disso, em certas situações, pode ser benéfico estimar modelos econométricos utilizando diferentes matrizes de ponderação espacial para comparar os resultados. Tal comparação permite identificar possíveis variações significativas nos resultados, evidenciando a sensibilidade destes em relação à escolha da matriz e, consequentemente, testando a robustez do modelo (Almeida, 2012).

A matriz dos k-vizinhos mais próximos é comumente utilizada para dados pontuais, pois examina diretamente os pontos individuais em vez de se referir a áreas. De maneira similar a dados pontuais, é possível criar uma matriz de k-vizinhos próximos para dados de área ao determinar primeiro os centróides das regiões, sendo os pontos que representam o centro de gravidade das geometrias espaciais. Logo, para dados pontuais, a matriz de k-vizinhos próximos é uma solução analítica natural, embora a escolha do número de vizinhos (k) seja geralmente baseada em modelagem ou experiência aleatória.

A definição da matriz de pesos é, portanto, uma etapa fundamental que reflete as hipóteses subjacentes sobre como os fenômenos espaciais interagem, influenciando significativamente os resultados da análise de autocorrelação espacial (Cliff & Ord, 1981).

2.2. A Estatística I de Moran Global

As estatísticas I de Moran indicam se existe um efeito espacial de aglomeração. A existência de autocorrelação espacial positiva é atribuída para valores positivos e significativos da estatística, ou seja, similaridade dos objetos examinados a uma determinada distância d, onde valores semelhantes entram em contato com mais frequência do que aleatoriamente (Tyszler, 2006). Valores negativos da estatística I significam exatamente o oposto, sendo autocorrelação negativa, quando valores semelhantes tocam ou aproximam-se com menos frequência do que aleatoriamente, chamado de dissimilaridade. Autocorrelação positiva significa a existência de *clusters* de valores semelhantes altos ou baixos, enquanto os valores negativos da estatística I são interpretados como os chamados *hot spots* ou *ilhas* de valores definitivamente diferentes (Goodchild M. , 1986). *Clusters* podem ser definidos como “agrupamentos distintos nos

dados, correspondendo frequentemente à multimodalidade na distribuição de probabilidades subjacente para os dados" (Almeida, 2012).

Almeida (2012) discute que um coeficiente de autocorrelação caracteriza um conjunto de dados organizado em uma determinada sequência. Especificamente, um coeficiente de autocorrelação espacial caracteriza dados organizados conforme uma sequência espacial. O autor adicionalmente aponta que qualquer coeficiente de autocorrelação é formulado a partir da relação entre uma medida de autocovariância e uma medida de variação total nos dados. Esta última pode ser calculada através de produto cruzado, quadrado da diferença ou módulo da diferença, variando de acordo com a escala de medida das variáveis, o que é um problema para o cálculo caso as variáveis tenham escalas de medidas diferentes. Para eliminar a influência da escala de medida, efetua-se a divisão pela variância dos dados. Também é preciso uma matriz de ponderação espacial W, conforme discutido na seção anterior, correspondendo a configuração da conectividade da interação espacial. Portanto, para elaborar uma estatística de autocorrelação espacial, é essencial contar com três componentes: uma medida de autocovariância, uma medida de variação dos dados e uma matriz de ponderação espacial W. (Almeida, 2012).

Cliff & Ord (1981) decorrem que normalmente as estatísticas I de Moran são interpretadas como coeficientes de correlação, embora o seu valor não seja limitado no intervalo [-1, 1]. A correlação acontece entre os valores de uma determinada variável em um local e os valores da mesma variável nas localidades vizinhas, analisada como defasagem espacial da variável testada w_x. O índice de Moran I é calculado pela fórmula:

$$I = \frac{N}{W} \times \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (8)$$

onde N é o número total de observações espaciais, w_{ij} é um elemento da matriz de pesos espaciais que define a relação espacial entre a localização i e j. Se i e j são vizinhos, w_{ij} é positivo; caso contrário, pode ser zero. x_i e x_j são os valores da variável de interesse nas localizações i e j, respectivamente. $\bar{x}$ é a média dos valores de x em todas as observações. W é a soma de todos os pesos espaciais, $\sum_i \sum_j w_{ij}$, em que $i \neq j$ (Cliff & Ord, 1981).

Concluímos então que o I de Moran pode ser utilizado para avaliar a autocorrelação espacial medindo a similaridade de cada região com suas áreas adjacentes e calculando a média dessas avaliações. Ao assumir a hipótese nula, que postula autocorrelação espacial nula, as observações x são consideradas distribuídas de forma independente e idêntica. Nesse cenário, o I de Moran segue uma distribuição normal assintótica, com valor esperado e variância iguais a

$$E[I] = -\frac{1}{n-1} \tag{9}$$

e

$$Var[I] = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2} \tag{10}$$

onde $S_0 = \sum_{i \neq j} w_{ij}$, $S_1 = \frac{1}{2}\sum_{i \neq j}\left(w_{ij} + w_{ji}\right)^2$ e $S_2 = \sum_k\left(\sum_j w_{kj} + w_{ik}\right)^2$ (Moraga, 2023).

O valor obtido de I deve coincidir com o seu valor esperado, considerando os limites da significância estatística, caso x_i seja independente dos valores das regiões adjacentes. Valores de I superiores ao valor esperado $E[I]$ sugerem uma autocorrelação espacial positiva, enquanto valores de I inferiores ao esperado indicam uma autocorrelação espacial negativa (Almeida, 2012).

Com um número suficientemente amplo de regiões, a estatística I adota uma distribuição normal, possibilitando a análise de desvios significativos de um padrão de um dado em relação a um padrão aleatório pela comparação do *z-score* (Z)

$$Z = \frac{I - E(I)}{Var(I)^{1/2}} \tag{11}$$

a uma distribuição normal padronizada. Um método alternativo para determinar a significância é a Randomização de Monte Carlo. Este procedimento gera padrões aleatórios ao redistribuir os valores observados entre as áreas e calcular o I de Moran para cada configuração, estabelecendo assim uma distribuição aleatória para a estatística I. Caso o I de Moran observado se localize nas extremidades dessa distribuição, a hipótese de independência entre as observações é descartada (Moraga, 2023). Desta forma, é possível verificar a presença de autocorrelação espacial seguindo esses passos:

1. Indicação das hipóteses nula e alternativa:

   $H_0: I = E[I]$ (não há autocorrelação espacial)

   $H_1: I \neq E[I]$ (há autocorrelação espacial)
2. Escolha do nível de significância α que se deseja tolerar. Usualmente considera-se como $\alpha = 0.05$ (Moraga, 2023).
3. Calcula-se o valor da estatística de teste sob as condições da hipótese nula dada pela Equação (11).
4. Determina-se o p-valor para os dados observados comparando o *z-score* com a distribuição normal padrão ou por meio da randomização de Monte Carlo. O p-valor representa a probabilidade de alcançar uma estatística de teste tão extrema ou mais extrema que a estatística de teste observada na direção da hipótese alternativa, sob a suposição de que a hipótese nula está correta, se traduz em $p = 2P(|Z| \geq z \mid H_0)$
5. Toma-se uma das duas decisões:

   Se p-valor $\leq \alpha$, rejeita-se a hipótese nula, então concluímos que os dados fornecem evidências para a hipótese alternativa.

   Se p-valor $> \alpha$, não se rejeita a hipótese nula, ou seja, os dados não fornecem evidências para a hipótese alternativa.

Outras combinações de hipóteses também podem ser feitas, como a verificação se há autocorrelação espacial positiva. Nesse caso, H_0: $I \leq E[I]$ e $H_1: I > E[I]$. Também é possível ajustar as hipóteses para verificar autocorrelação espacial negativa.

Uma metodologia alternativa para examinar a autocorrelação espacial é através do diagrama de dispersão de Moran, que ilustra a defasagem espacial da variável em análise no eixo vertical $(Wx)_i$ contra o valor dessa mesma variável no eixo horizontal (x_i). O peso espacial da unidade j em relação à unidade i é representado pelo elemento w_{ij}. O produto matricial da matriz de pesos W pelo vetor das variáveis originais x gera os elementos $(Wx)_i$ (Tyszler, 2006), conforme

$$(Wx)_i = \sum_{j=1}^{n} w_{ij} x_j \tag{12}$$

ou seja, a variável com defasagem espacial resulta da soma dos produtos entre as observações de todas as outras unidades e o peso que cada uma delas (unidade j) tem sobre a unidade i. Desta maneira, a variável defasada no espaço se converte em um vetor que, para cada unidade, incorpora a influência espacial dos valores de suas unidades vizinhas.

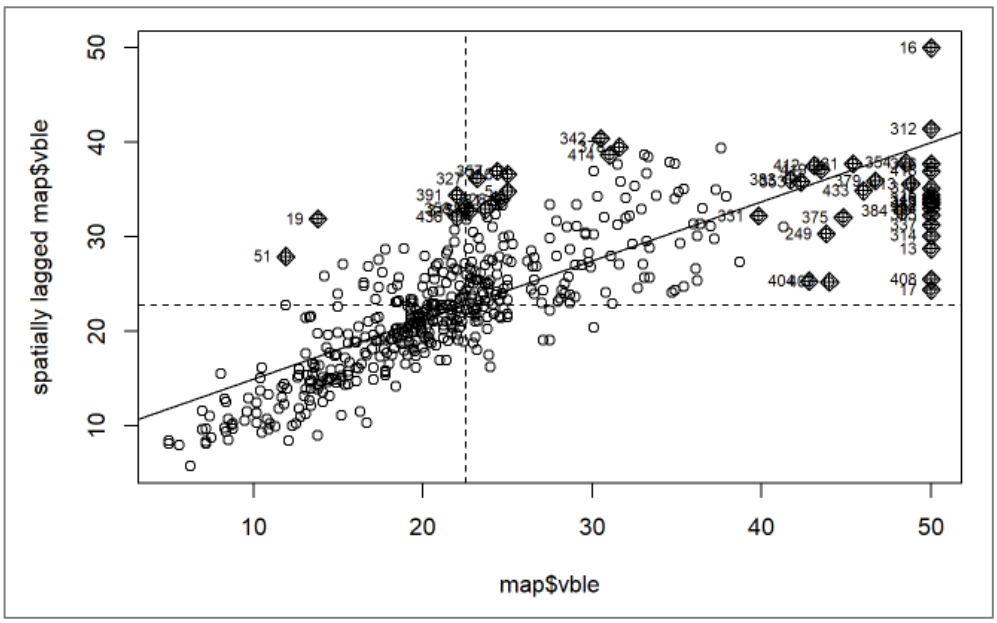


Figura 5: Exemplo do Diagrama de Dispersão de Moran (Fonte: Moraga, 2023)

Nos quadrantes do diagrama de dispersão, H representa valores altos e L, valores baixos. Assim, os quadrantes podem ser descritos da seguinte forma: HH inclui observações que estão acima da média e cuja defasagem espacial também é superior à média; HL abrange observações acima da média cuja defasagem espacial é inferior à média; LL engloba observações abaixo da média com defasagem espacial também abaixo da média; e LH reúne observações abaixo da média, mas com defasagem espacial acima da média (Tyszler, 2006).

O diagrama de dispersão de Moran efetivamente exibe a dispersão dos pontos que representam as regiões, marcando a inclinação da reta de regressão. A estatística I de

Moran corresponde então ao valor do coeficiente de inclinação de uma linha de regressão no diagrama apresentado. Dessa forma, esse traçado indica uma correlação espacial absoluta (Tyszler, 2006). No exemplo da Figura 5, uma correlação linear positiva é evidente entre as observações e as suas contrapartes espacialmente desfasadas. Com este gráfico, também podemos identificar pontos de dados que impactam significativamente a correlação linear entre o conjunto de dados e seus valores defasados (Moraga, 2023). A precisão com que a linha de regressão se ajusta indica o nível de correlação espacial presente. Ainda, concentrações de pontos nos quadrantes LL e HH sinalizam uma correlação espacial positiva, refletindo que unidades geográficas semelhantes possuem valores semelhantes. Por outro lado, concentrações nos quadrantes LH e HL sugerem o contrário. É importante notar que, em ambos os cenários, o foco está menos no sinal da correlação e mais na indicação de que o padrão espacial observado não é aleatório (Almeida, 2012).

Os outliers espaciais são observações que se destacam por não seguir o padrão de dependência espacial predominante nos dados. É crucial identificar esses outliers, pois eles podem influenciar indevidamente a análise global de autocorrelação espacial, comprometendo a precisão do teste. Eles também podem indicar problemas na especificação da matriz de pesos espaciais ou na escolha da escala espacial dos dados. Também é importante fazer a distinção entre um outlier espacial e um ponto de alavancagem. Um ponto de alavancagem segue a tendência de associação espacial dos dados, mas tem uma influência desproporcional na determinação do nível dessa associação (Almeida, 2012).

Utilizando o diagrama de dispersão de Moran, é possível identificar *outliers* espaciais e pontos de alavancagem. Esse diagrama, como discutido anteriormente, revela quatro tipos de associações espaciais (HH, LL, HL, LH) baseados em seu quadrante. A inclinação da linha de regressão, calculada de Wx contra x, oferece o valor do coeficiente de Moran I, auxiliando na detecção dessas influências atípicas (Almeida, 2012).

Em adição, outra medida global de autocorrelação espacial foi proposta por Geary em 1954. Ela é baseada em uma medida diferente de covariância: o quadrado da diferença entre os pares de valores do atributo em estudo. Novamente, a hipótese nula é

a de aleatoriedade espacial, ou seja, a ausência de dependência espacial nos dados. Algebricamente, a fórmula dessa estatística é dada por:

$$c = \frac{(n-1)\sum_i \sum_j w_{ij}\left(x_i - x_j\right)^2}{\sum_i \sum_j w_{ij} \ \sum_i (x_i - \bar{x})^2} \tag{13}$$

onde c é o índice Geary, n o número de observações, w_{ij} são os pesos espaciais entre as observações i e j, x_i e x_j são os valores do atributo em estudo nas observações i e j e $\bar{x}$ é a média dos valores do atributo. Se c é aproximadamente igual a 1, isso indica aleatoriedade espacial. Valores de c menores que 1 indicam autocorrelação espacial positiva, enquanto valores maiores que 1 indicam autocorrelação espacial negativa. Enquanto o I de Moran é mais sensível à autocorrelação de longo alcance, o c de Geary, baseado na soma dos quadrados das diferenças entre pares de valores de atributos, é mais sensível à autocorrelação de curto alcance. Ambos os índices são ferramentas importantes na análise espacial e podem ser usados complementarmente para obter uma compreensão mais abrangente dos padrões espaciais em um conjunto de dados (Almeida, 2012).

2.3. O I de Moran Local

Observamos que o I de Moran Global serve como uma métrica para avaliar a autocorrelação espacial em toda uma área de estudo. Por vezes, há também interesse em avaliar a semelhança local entre o valor de cada área e o das áreas vizinhas. Os Indicadores Locais de Associação Espacial (LISA), introduzidos por Anselin em 1995, visam identificar o grau de agrupamento espacial significativo de valores semelhantes em torno de cada observação. Uma característica fundamental dos LISAs é que a soma dos valores LISA para todas as áreas é igual a um múltiplo do indicador de associação espacial global. Consequentemente, as estatísticas globais podem ser divididas numa coleção de estatísticas locais, sendo a maioria dos LISAs adaptações locais de índices globais estabelecidos (Anselin, 1995; Almeida, 2012).

Assim, definimos a versão local do I de Moran para a i-nésima região

$$I_i = \frac{n(x_i - \bar{x})}{\sum_j \left(x_j - \bar{x}\right)^2} \sum_j w_{ij}\left(x_j - \bar{x}\right) \tag{14}$$

Podemos notar que o I de Moran global é proporcional a soma de todos os I_i, escrevendo a seguinte equação

$$I = \frac{1}{\sum_{i \neq j} w_{ij}} \sum_{i} I_i \qquad (15)$$

Assim sendo, para cada observação, calcula-se um I_i. Dessa forma, realizamos n cálculos da estatística I_i e determinamos seus respectivos níveis de significância a partir do valor esperado $E[I_i] = -\mathrm{w_i}/(\mathrm{n} = 1)$. Essa abundância de informações pode ser confusa para o pesquisador se apresentada em tabelas. Para compreender o I de Moran local para cada área, é essencial gerar um mapa de p-valores (Almeida, 2012). Os p-valores representam a probabilidade dos valores observados ocorrerem sob a suposição de que a hipótese nula de nenhuma associação espacial está correta. Estas probabilidades podem ser derivadas, independentemente da existência de associação espacial global, através de um método de simulação empregando uma estratégia de randomização condicional. Nesta estratégia, o valor observado x_i na região i é mantido constante, enquanto os demais valores são redistribuídos aleatoriamente entre as outras regiões (Moraga, 2023).

Um valor alto para I_i indica que a área está cercada por regiões com valores semelhantes, posicionando-a em um *cluster* de observações altas, baixas ou moderadas (Almeida, 2012). Por outro lado, um valor baixo para I_i sugere que a área está cercada por regiões com valores diferentes, marcando-a como um outlier e indicando que a observação para a área I diverge da maioria ou de todos os seus vizinhos (Anselin, 1995). O I de Moran Local permite a análise dos *clusters* que são significativos para a contribuição na autocorrelação espacial, sendo possível gerar mapas indicando esses *clusters*.

3. A ANÁLISE DE COMPONENTES PRINCIPAIS

Um grupo de métodos conhecidos como Estatística Multivariada analisa diversas variáveis que são usadas para descrever os objetos ou indivíduos em uma amostra simultaneamente. Estes métodos são frequentemente categorizados técnicas de dependência ou interdependência. Uma variável definida como variável dependente é explicada por outras variáveis, ditas independentes, nas técnicas de dependência (Hair,

Black, Babin, & Anderson, 2019). Modelos de regressão múltipla e análise discriminante são exemplos de métodos de dependência. Em contraste, nenhuma variável é considerada dependente ou independente quando se usam métodos de interdependência; em vez disso, todas as variáveis são examinadas simultaneamente para identificar uma estrutura para o conjunto completo de variáveis. A Análise Fatorial, Análise de C*lusters* e o Escalonamento Multidimensional (MDS) são exemplos de métodos de interdependência (Jolliffe, 2002).

A PCA enquadra-se na categoria de análise de interdependência porque não distingue entre variáveis dependentes e independentes. Em vez disso, ela procura identificar um conjunto de combinações lineares das variáveis originais. Estas combinações são selecionadas de forma a maximizar a variância explicada pelos dados, proporcionando uma visão simplificada da estrutura subjacente que captura as relações interdependentes entre todas as variáveis. A PCA é usada para descobrir ou reduzir a complexidade dos dados, revelando as direções de maior variabilidade e, por consequência, as relações interdependentes entre as variáveis originais (Hair, Black, Babin, & Anderson, 2019).

Portanto, a ideia central da PCA é a redução da dimensionalidade de um conjunto de dados que consiste em um número considerável de variáveis interrelacionadas, enquanto busca reter o quanto é possível a variabilidade desses dados. Para tanto, o conjunto original é transformado em um conjunto de novas variáveis, os Componentes Principais (PCs), como são tipicamente chamados (Jolliffe, 2002).

Cada PC é a combinação linear das variáveis originais mencionada acima ponderada pelos elementos do autovetor correspondente (Jolliffe, 2002). Para um conjunto de dados X, a transformação para o primeiro componente principal Y_1 é dada por

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \tag{16}$$

onde a_{1j} são os elementos do primeiro autovetor. Os dados originais são projetados nos espaços definidos pelos autovetores selecionados para formar os novos conjuntos de dados baseados nos PCs. Os chamados scores (ou pontuações) dos indivíduos nos PCs indicam a localização dos indivíduos no espaço multidimensional reduzido formado pelos PCs. Isso permite avaliar como os indivíduos contribuem ou estão relacionados com os PCs.

3.1. A Transformação das Variáveis e a Matriz de Covariância

Conforme mencionado, a diminuição da complexidade das variáveis originais ocorre por meio da conversão para um novo grupo de variáveis

Antes da transformação, geralmente é necessário padronizar as variáveis, especialmente se elas possuem escalas diferentes. Para tanto, basta subtrair-se a média e dividir pelo desvio padrão de cada variável, garantindo que todas tenham média 0 e variância 1. A padronização evita que variáveis com maior variabilidade dominem os primeiros PCs simplesmente devido à sua escala. Assim, dado um conjunto de dados X de dimensão $\mathrm{n} \times \mathrm{p}$, onde n é o número de observações e p é o número de variáveis. Então, para cada variável X_j, a variável padronizada Z_j é obtida por

$$Z_j = \frac{X_j - \mu_j}{\sigma_j} \tag{17}$$

onde μ_j é a média e σ_j é o desvio padrão da variável X_j.

A construção da matriz de covariância é essencial para entender as relações lineares entre as variáveis originais. Esta matriz desempenha um papel central na PCA, servindo como a base para identificar as direções (ou PCs) ao longo das quais os dados variam mais. A obtenção e interpretação desta matriz são fundamentais para entender a aplicabilidade e eficácia da PCA para um conjunto de dados específico (Jolliffe, 2002).

Assim a matriz de covariância C pode ser calculada a partir do mesmo conjunto de dados X apresentado anteriormente. Cada elemento c_{ij} da matriz C representa a covariância entre as variáveis i e j, calculada como

$$c_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) \tag{18}$$

onde x_{ki} é o valor da variável i na observação k, e $\bar{x}_i$ é a média da variável i. Essa matriz é simétrica, com variâncias das variáveis na diagonal principal e covariâncias entre pares de variáveis nas posições fora da diagonal, sendo

$$\mathrm{C} = \begin{bmatrix} Var(x_i) & c_{ij} & \dots & c_{in} \\ c_{ji} & Var(x_j) & \dots & c_{jn} \\ \vdots & \vdots & \ddots & \vdots \\ c_{ni} & c_{nj} & \dots & Var(x_n) \end{bmatrix} \tag{19}$$

É possível derivar a matriz de correlações a partir da matriz de covariâncias das variáveis. Esse processo envolve a conversão das covariâncias em coeficientes de correlação, que são adimensionais e variam de -1 a 1, permitindo uma comparação direta da força das relações lineares entre as variáveis, representada pelo coeficiente de Pearson. A matriz de correlações ρ que apresenta os valores da correlação linear de Pearson entre cada par de variáveis é apresentada conforme mostra a expressão (Fávero & Belfiore, 2017)

$$\rho = \begin{bmatrix} 1 & \rho_{ij} & \cdots & \rho_{in} \\ \rho_{ji} & 1 & \cdots & \rho_{jn} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{ni} & \rho_{nj} & \cdots & 1 \end{bmatrix} \tag{20}$$

O coeficiente de correlação linear de Pearson é calculado segundo a expressão

$$\rho_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}\left(x_{ki} - \bar{x}_i\right)^2}\sqrt{\sum_{k=1}^{n}\left(x_{kj} - \bar{x}_j\right)^2}} \tag{21}$$

Utilizando-se das equações (18) e (21), podemos então derivar a matriz de correlações da matriz de covariâncias. A matriz de correlações pode ser calculada normalizando os elementos da matriz de covariâncias pelas variâncias das variáveis envolvidas, através da seguinte expressão

$$\rho_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j} \tag{22}$$

onde σ_i e σ_j representam os desvios padrão das variáveis.

Retomando, a PCA busca direções (as PCs) que maximizam a variância total capturada através de transformações lineares, método que é diretamente influenciado pelas matrizes acima. Os autovetores da matriz de covariância apontam para as direções de máxima variância nos dados, e os autovalores associados quantificam essa variância. Portanto, a estrutura da matriz de covariância determina a orientação e importância relativa dos PCs (Jolliffe, 2002). A presença de covariâncias significativas (positivas ou negativas) sugere que as variáveis têm relações lineares entre si, o que é um indicativo de

que a PCA pode ser útil na simplificação da estrutura dos dados ao reduzir variáveis correlacionadas a PCs menos numerosos.

A interpretação da matriz de covariância pode informar sobre a validade e eficácia da PCA para um conjunto de dados. Se a matriz de covariância mostra alta variância para várias variáveis e covariâncias significativas entre elas, isso indica que os dados têm uma estrutura rica que a PCA pode explorar, potencialmente revelando PCs significativos. Ainda, se alguns autovalores são significativamente maiores que os outros, isso sugere que poucos PCs capturam a maior parte da variância dos dados. Por outro lado, se os autovalores são relativamente homogêneos (ou seja, têm valores semelhantes), isso pode indicar que os dados são isotrópicos, ou seja, têm variância semelhante em todas as direções, e que a PCA pode não ser tão eficaz na construção de PCs que carreguem consistentemente a variabilidade dos dados das variáveis originais. A matriz de covariância, portanto, não só fundamenta o processo de PCA, mas sua análise também orienta a interpretação dos resultados e a decisão sobre a aplicabilidade da PCA para revelar estruturas significativas em um conjunto de dados (Hair, Black, Babin, & Anderson, 2019).

Entretanto, Jollife (2002) discute que a matriz de correlações é preferida em PCAs por oferecer vantagens como invariância à escala das variáveis e comparabilidade entre diferentes análises, devido à padronização que torna os PCs combináveis de maneira significativa. Esta abordagem facilita a interpretação dos PCs, e é especialmente útil quando as variáveis têm tipos ou unidades de medida diferentes. Embora a PCA baseada em matrizes de covariância possa ter vantagens em inferência estatística, essa abordagem é frequentemente usada mais como ferramenta descritiva, tornando a invariância de escala e a otimização de critérios relevantes na PCA baseada em matrizes de correlações. O autor conclui que a escolha entre as duas metodologias depende das necessidades específicas da análise e das características do conjunto de dados (Jolliffe, 2002).

3.2. Autovalores e Autovetores

O próximo passo na transformação para PCs usa álgebra de matrizes na decomposição de Eigen da matriz de covariância $\mathbf{C}$ previamente descrita. Esta envolve encontrar os autovalores λ_i e os autovetores v_i tais que

$$\boldsymbol{C}v_i = \lambda_i v_i \tag{23}$$

Os autovalores λ_i representam a variância capturada por cada PC, e os autovetores v_i são os direcionadores dos PCs no espaço original das variáveis. Os autovalores são ordenados em ordem decrescente $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, e os autovetores são ordenados de acordo com esses autovalores. Logo, os PCs que capturam a maior variância são priorizados (Jolliffe, 2002).

A transformação das variáveis originais X em PCs Y é realizada usando os autovetores v_i. Para um dado autovetor v_i, o componente principal correspondente Y_i é calculado como:

$$Y_i = Xv_i \tag{24}$$

Isso resulta em uma nova matriz de dados Y onde cada coluna é um PC, a combinação linear das variáveis originais conforme mencionado anteriormente.

Após a obtenção dos autovalores e autovetores da matriz de covariância (ou correlação) na PCA, os próximos passos envolvem a seleção de PCs, a transformação dos dados originais e a interpretação dos resultados. Em seguida detalhamos cada uma dessas etapas finais da análise.

3.3. Seleção dos Componentes Principais e Interpretação dos Resultados

A ordenação dos autovalores e autovetores conforme seção anterior prioriza os PCs que capturam a maior variância. É possível fazer o cálculo da variância explicada por cada PC como a proporção do autovalor correspondente em relação à soma de todos os autovalores por

$$Variância\ explicada = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{25}$$

onde λ_i é o i-ésimo autovalor.

A quantidade ideal mínima de variância explicada para considerar uma análise válida varia conforme o contexto da análise e a natureza das variáveis. Hair (2019) cita que algumas aplicações em Ciências Naturais podem exigir 95% da variância, enquanto outras aplicações em Ciências Sociais 60% da variância total pode ser considerada satisfatória

(Hair, Black, Babin, & Anderson, 2019). Este critério é amplamente discutido na literatura estatística (Johnson & Wichern, 2007). Logo, analistas podem ajustar o número de PCs a reter considerando o contexto do estudo em questão.

O Teste de Esfericidade de Bartlett é uma análise estatística utilizada para avaliar a adequação dos dados para a PCA. O objetivo principal do teste é verificar se a matriz de correlação dos dados é significativamente diferente de uma matriz de identidade, que indica que as variáveis são suficientemente correlacionadas para justificar a realização da análise (Hair, Black, Babin, & Anderson, 2019). A estatística de Bartlett é calculada usando a seguinte fórmula:

$$\chi^2 = -\left(n - 1 - \frac{2p + 5}{6}\right) \cdot \ln\left(\det(\rho)\right) \tag{26}$$

onde n é o número de observações, p é o número de variáveis e ρ é a matriz de correlação das variáveis.

O índice Kaiser-Meyer-Olkin (KMO) é utilizado na PCA para avaliar a adequação da amostra e verificar se os dados são apropriados para realizar essas análises. O teste KMO mede a proporção da variância em suas variáveis explicada pelos componentes principais, garantindo que as variáveis estejam suficientemente correlacionadas para justificar a decomposição em componentes principais, sendo calculado utilizando a seguinte fórmula:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j}\left(r_{ij}^2 + h_{ij}^2\right)} \tag{27}$$

onde r_{ij}^2 é o quadrado das correlações entre as variáveis i e j e h_{ij}^2 é a parte não explicada pela correlação entre as variáveis i e j. Valores baixos de KMO sugerem que a análise pode não ser apropriada, pois as variáveis não têm correlações suficientemente fortes. Valores superiores a 0,7 indicam uma adequação da amostra boa, sendo valores acima de 0,8 considerados como adequação muito boa. Já valores abaixo de 0,6 indicam adequação de dados questionável ou inadequada (Jolliffe, 2002).

Um recurso visual útil para determinar o número apropriado de PCs é o gráfico chamado de *screeplot*. Este gráfico apresenta os autovalores ordenados de maneira decrescente, exibindo s magnitude do autovalor versus o número do PC. Para auxiliar na

determinação do número apropriado de PCs a se considerar, usualmente procura-se um cotovelo (curva mais acentuada) no gráfico. O número de PCs geralmente considerado é definido até o ponto em que os autovalores restantes são relativamente pequenos e aproximadamente do mesmo tamanho, indicando baixa representação da variabilidade dos dados originais naqueles componentes.

Após a transformação dos dados originais para identificar as direções de maior variabilidade, pode-se fazer interpretações significativas sobre os padrões subjacentes nos dados a partir de algumas propriedades das PCs. Os carregamentos (*loadings*) indicam a contribuição de cada variável original para o PC. Logo, variáveis com carregamentos altos (positivos ou negativos) em um PC têm forte influência sobre aquele componente e também indicam uma correlação forte entre eles, ajudando a interpretar os PCs em termos das variáveis originais. Esta interpretação não apenas facilita a compreensão das dinâmicas complexas que podem existir entre as variáveis iniciais, mas também fornece uma base sólida para decisões informadas e maior entendimento do espaço multivariado original para análises subsequentes (Hair, Black, Babin, & Anderson, 2019).

A qualidade da representação da variabilidade dos dados na PCA pode ser avaliada por diferentes métodos, cabendo ao analista fazer uso daqueles que entender serem de maior relevância no contexto de sua análise. Alguns recursos para avaliar a qualidade da representação na PCA já foram mencionados, por exemplo, o *screeplot* e a contribuição das variáveis aos PCs (*loadings*).

Outro recurso amplamente utilizado é conhecido pelo nome de quadrado do cosseno ($\cos^2$) das variáveis. Este último foca em quão bem cada variável original é representada nos PCs, fornecendo informações importantes sobre a eficácia da redução de dimensionalidade. Seu nome provém de cálculo o cosseno quadrado indica a contribuição de um componente para a distância quadrada da observação à origem. A contribuição corresponde ao quadrado do cosseno do ângulo do triângulo retângulo feito com a origem, a observação e sua projeção na componente (Abdi & Williams, 2010).

Para o cálculo do $\cos^2$ de cada variável em um determinado PC são utilizados os scores das variáveis naquele componente. Para uma dada variável j em um PC i, o cos^2 é calculado como o quadrado do score da observação (a) da variável j no componente i,

dividido pela soma dos quadrados de todos os scores da variável j em todos os PCs, que corresponde ao quadrado da distância da observação j com a origem. O denominador assegura que a soma dos cos² de uma observação em todos os componentes seja 1, refletindo a ideia de que 100% da informação (variância) da observação é distribuída entre os componentes. Então expressamos como

$$cos_{ij}^2 = \frac{a_{ij}^2}{\sum_{k=1}^{p} a_{jk}^2} = \frac{a_{ij}^2}{d_{ij}^2} \tag{28}$$

onde p é o número total de PCs. A distância quadrada, d_{ij}^2, é calculada como a soma dos valores quadrados de todas as pontuações dos fatores desta observação, dado Teorema de Pitágoras (Abdi & Williams, 2010).

Em sua interpretação, componentes com um valor alto de cos^2 contribuem com uma parcela significativa para a distância total, portanto, esses componentes são importantes para aquela observação. O valor de cos^2 varia de 0 a 1, onde valores próximos a 1 indicam que a variável é muito bem representada pelo PC em questão, ou seja, a maior parte da sua variância é capturada por esse componente.

Para auxílio na interpretação, recursos computacionais permitem projetar gráficos que mostram o cos^2 de cada variável em relação aos PCs e podem ajudar a visualizar quais variáveis são melhor representadas por quais componentes. Ainda, gráficos conhecidos como *biplot* combinam a representação dos PCs com os *loadings* das variáveis, permitindo uma visualização simultânea da relação entre variáveis e componentes, bem como a qualidade da representação das variáveis (Jolliffe, 2002).

Assim, o cálculo e a análise dos quadrados do cosseno são, portanto, recursos relevantes para interpretar os resultados da PCA e entender como as variáveis originais contribuem para os PCs selecionados. Embora o uso do cos^2 ofereça uma abordagem útil para avaliar esta contribuição, é importante complementá-lo com outras medidas e gráficos para obter uma compreensão ampla da eficácia da PCA na captura da variabilidade dos dados originais.

Em resumo, a análise de PCA é uma ferramenta poderosa na mão do analista, permitindo a redução eficaz da dimensionalidade de dados enquanto preserva a maior

parte da variabilidade (informação) original. Quando aplicada e interpretada corretamente, pode revelar a estrutura essencial dos dados originais, simplificar sua complexidade e facilitar análises subsequentes, tornando-se indispensável em diversas aplicações.

PARTE 2: APLICAÇÃO DO MÉTODO E ANÁLISE DOS RESULTADOS DO MAR DO NORTE

A expedição do cruzeiro marítimo a bordo do *RV Meteor* incluiu uma equipe de pesquisadores dedicados a rastrear metais críticos de alta tecnologia no Mar do Norte. Durante a viagem, os pesquisadores enfrentaram desafios logísticos e técnicos (TRAM, 2024), mas também momentos de descoberta científica e interações valiosas entre diferentes disciplinas. O foco foi a coleta de amostras de água em de diferentes profundidades para investigar a presença e o impacto de metais contaminantes emergentes, assim como a resistência de metais em comunidades microbianas. A equipe, composta por geoquímicos, microbiologistas e oceanógrafos, trabalhou em turnos para processar as amostras, usando técnicas como ultrafiltração para isolar frações coloidais e análises de águas para íons metálicos. A expedição ofereceu *insights* significativos sobre o ciclo de traços de metais em sistemas de estuário e o papel dos rios na introdução de metais no mar, destacando a importância da colaboração multidisciplinar em pesquisas oceanográficas (A. Koschinsky, 2021).

4. COLETA E PRÉ-TRATAMENTO DOS DADOS

Os dados foram coletados em andamento, ao longo do percurso do cruzeiro, com dois sistemas autônomos de termossalinografia (*thermosalinograph* - TSG), nomeados TSG1 e TSG2, consistindo no uso do TSG modelo SBE21 junto com um termômetro SBE38. Os sistemas funcionaram independentes um do outro durante todo o cruzeiro, garantindo a validação cruzada dos dados e a redundância na coleta. A temperatura foi medida na entrada de água a cerca de 2,5 m de profundidade, a salinidade é estimada no interior do TSG a partir da condutividade e da temperatura interior (Schlundt, 2022). As observações foram tomadas para cada 5 segundos no TSG1 e 10 segundos no TSG2. Nos dados ainda não processados dos sensores, foram feitas médias das medições minuto a minuto, descartando-se valores que extrapolam 2 vezes o valor do desvio padrão. Finalmente, o grupo fez a escolha do conjunto de dados TSG1 para publicação dos resultados, tendo os

valores de salinidade e temperatura muito próximos para ambos os sensores conforme demostrado nas séries temporais da Figura 6.

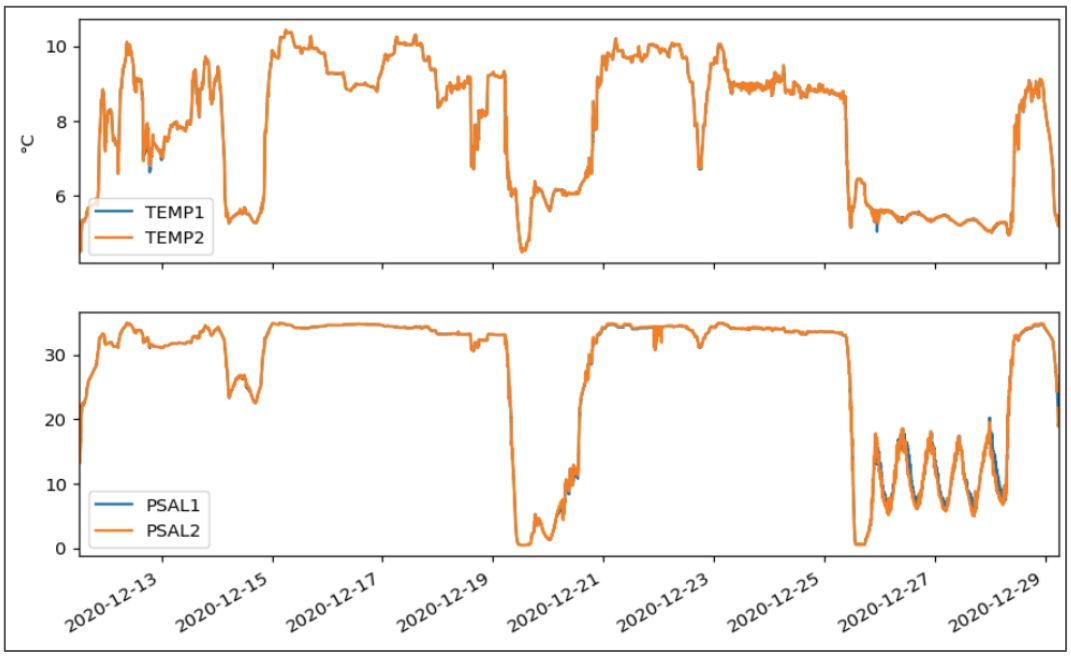


Figura 6: Séries temporais de salinidade e temperatura TSG1 e TSG2 (SeaDataNet, 2019)

Para detalhes técnicos específicos dos sensores e das etapas de controle de qualidade de dados, pode-se consultar o relatório de processamento de dados na íntegra fornecidos pela equipe de pesquisa (Schlundt, 2022), o qual foi utilizado como referência para o pré-tratamento dos dados do presente estudo.

4.1. Apresentação das Variáveis Estudadas e Divisão das Zonas

Embora o estudo anterior considerasse apenas os dados de salinidade e temperatura, a caracterização da área proposta pelo presente estudo considerou o valor das demais variáveis tomadas em cada observação pelo mesmo sensor. Os dados brutos do TSG1 foram então disponibilizados pela equipe em formato tab-delimitado (*tab-delimitted*) diretamente do software do equipamento. O arquivo possui 11 colunas (variáveis), sendo seus nomes e descrição:

- ***DateTime***: data e hora da observação

- ***Lat*** e ***Lon****:* posição (latitude e longitude) da observação
- ***Salinity****:* salinidade estimada no interior do TSG a partir da condutividade e da temperatura interior. A estimação é feita pelo próprio equipamento. A salinidade é expressa em Unidades de Salinidade Prática (PSU).
- ***TempExtern****:* temperatura externa da água, medida pelo termômetro SBE38.
- ***N.NO3****:* refere-se à concentração de íons de nitrato (NO_3^-) na amostra de água. A análise de nitrato é importante porque níveis elevados desse íon na água podem indicar poluição ou contaminação, o que pode ter efeitos prejudiciais nos ecossistemas aquáticos e representar riscos para a saúde humana (World Heath Organization, 2016). A concentração de N-NO3 é expressa em unidades de miligramas por litro (mg/L), que representam a massa de íons nitrato dissolvidos em um litro de água, calculada a partir da análise de espectro (LSA).
- ***DOCeq***: quantidade de 'Carbono Orgânico Dissolvido' (*Dissolved Organic Carbon*). Esta leitura é comumente usada em aplicações de monitoramento ambiental para avaliar a qualidade das amostras de água. O carbono orgânico dissolvido pode afetar a cor, o sabor e o odor da água, bem como sua capacidade de sustentar a vida aquática. Ao medir o carbono orgânico dissolvido, é possível identificar fontes de poluição e determinar a eficácia dos métodos de tratamento na remoção de carbono orgânico da água (Potter & Wimsatt, 2005). A concentração de DOC é expressa em unidades de miligramas por litro (mg/L), que representam a massa de íons nitrato dissolvidos em um litro de água. Esta variável está relacionada com a variável ABS*254* e é calculada a partir da análise de espectro (LSA).
- ***Abs210****,* ***Abs254****,* ***Abs360****:* variáveis correspondentes a absorção em cada comprimento de onda (210 nm, 254 nm, 360 nm, respectivamente). Estes parâmetros são medidos em unidades de absorção (AU).
- ***SAC254***: refere-se à Absorção Específica de Carbono a 254 nm (*Specific Absorbance of Carbon at 254 nm*). Este parâmetro é utilizado para avaliar a presença de compostos orgânicos na água, especialmente aqueles com

estruturas aromáticas, que tendem a absorver luz UV a um comprimento de onda de 254 nm. A SAC254 é um indicador útil da qualidade orgânica da água e pode ser usada para estimar a quantidade de matéria orgânica natural (NOM), incluindo ácidos húmicos e fúlvicos, presentes na amostra de água. Este parâmetro é particularmente relevante na monitorização da qualidade da água natural, pois altos níveis de NOM podem interferir com processos de tratamento e formação de subprodutos de desinfecção. (Baird, Eaton, Rice, Bridgewater, & Federation, 2017). Os resultados da análise NOM podem ajudar a identificar fontes de poluição e determinar a eficácia das medidas de remediação, sendo calculada a partir de ABS254, em 1/m.

As variáveis calculadas a partir da absorção nos comprimentos de onda podem ser visualizadas no espectro da Figura 7.

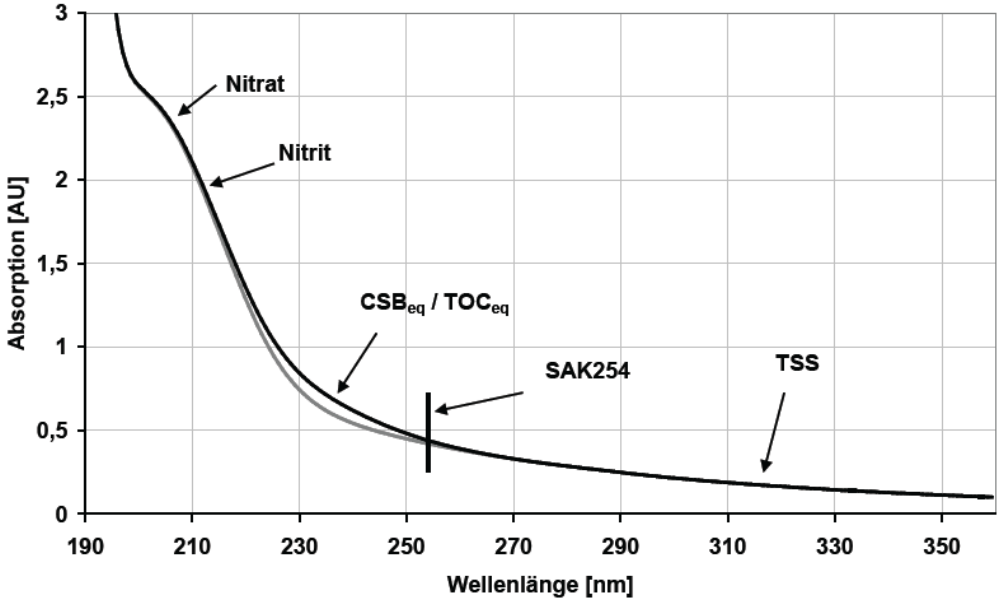


Figura 7: Regiões de Absorção das Diferentes Variáveis Calculadas (TriOS GmbH, 2017)

A explicação técnica dos princípios de como tal sensor funciona foge ao escopo desta dissertação. Porém, uma breve explicação pode ser dada, em que uma lâmpada de flash de xenônio é usada como fonte de luz de banda larga. A luz passa pelo caminho óptico

do sensor e é parcialmente absorvida por ele. O espectrômetro detecta a luz restante resolvida espectralmente e determina sua intensidade (I) em diferentes comprimentos de onda em uma faixa de comprimento de onda definida. O enfraquecimento da luz causado pela passagem pelo meio de medição é comparado ao enfraquecimento da luz causado pela passagem através de água ultrapura (TriOS GmbH, 2017). A medição em água ultrapura fornece a chamada intensidade básica (I_0). Usando a Equação 11 e a Equação 12, o sensor determina a transmissão (T) e a absorbância (A) para comprimentos de onda individuais na faixa de comprimento de onda definida:

$$T = \frac{I}{I_0} \tag{29}$$

$$A = -\log_{10} T \tag{30}$$

onde a transmissão (T) é dada em % da transmissão em água pura e a absorção é calculada em AUs (*absorbance unit*).

O conjunto de dados analisado foi dividido em três zonas de modo que as águas de cada zona tivessem o potencial de apresentar diferentes perfis nas variáveis estudadas. Em busca disto, foi utilizado a divisão oficial alemã apresentada na Figura 8 (Maritime Borders of the Federal Republic of Germany, 2019). As zonas são a considerada Águas Internas (AI), referente as águas doces e encosta próxima, Mar Territorial (MT), sendo a parte alemã do Mar do Norte, que abrange cerca de 22km da costa, e Mar do Norte (MN), sendo as águas internacionais. O esperado é que cada zona apresente diferenças características nas variáveis, como salinidade e temperatura, por exemplo. É preciso mencionar também que as zonas não estão sendo consideradas como áreas (polígonos) nos testes de estatística espacial, porém estão sendo apenas analisadas separadamente através dos pontos coletados dentro de cada região. Ainda na Figura 8 é apresentada a rota do cruzeiro no período considerado, representado pela linha pontilhada cinza, bem como pontos de cada região coloridos de acordo com sua localização.

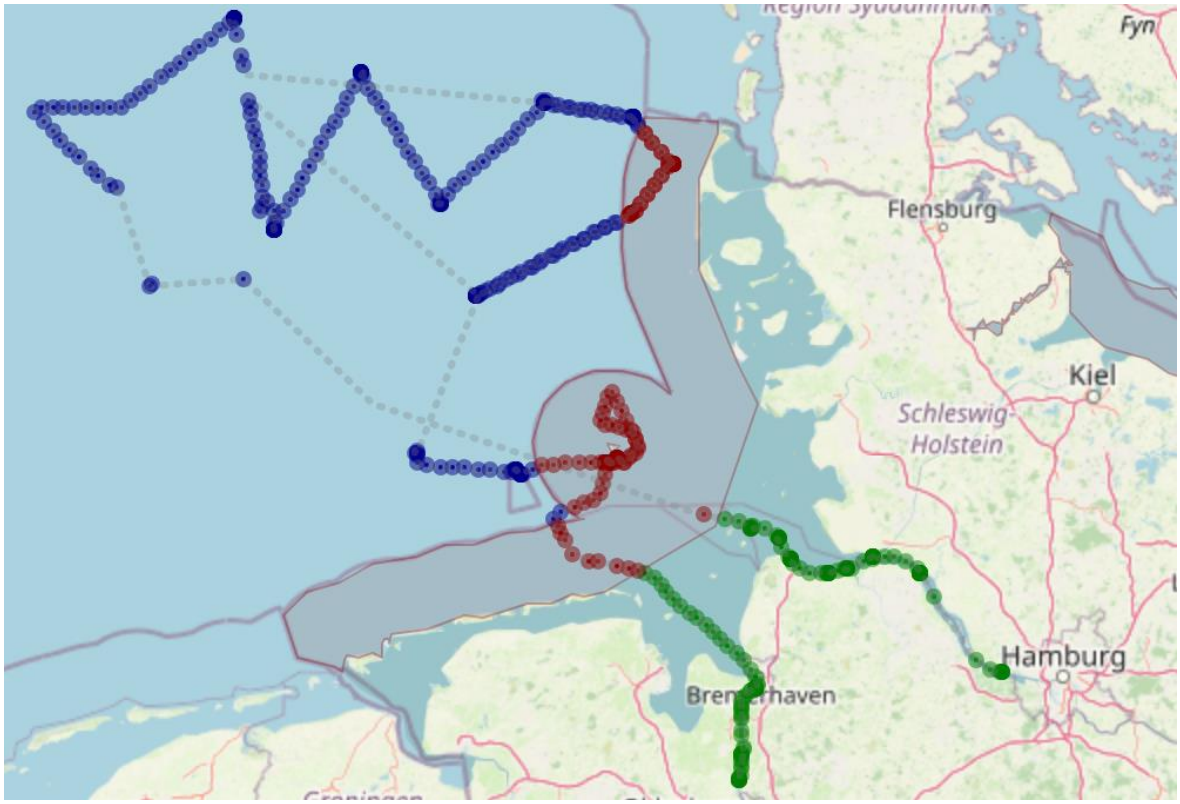


Figura 8: Zonas: Águas Internas (Verde), Mar Territorial (Vermelho) e Mar do Norte (azul)

Para que os dados sejam adequados para a posterior análise, foi realizado um pré-processamento que segue o protocolo de Controle de Qualidade de Dados da *SeaDataNet* (SeaDataNet, 2019), e que será descrito no próximo tópico.

4.2. O Pré-Processamento dos Dados

Os dados brutos continham cerca de 17 mil observações que foram obtidas no período de 7 dias (de 19/12 a 26/12/2020). As observações foram tomadas conforme referido no início deste capítulo.

O primeiro passo do pré-processamento foi avaliar se haviam dados não disponíveis no conjunto de dados. Como o foco da expedição era apenas sobre os dados de salinidade e temperatura, seria possível que as outras variáveis pudessem estar incompletas. As variáveis N.NO3, DOCeq e Abs210 foram as mais comprometidas, com cerca de 31.27%, 28.93% e 22.85% dos dados ausentes. A maior parte desses pontos faltantes na espectroscopia encontravam-se entre Glückstadt e Wedel, no Rio Elbe,

próximo a Hamburgo. Como são variáveis dependentes, removeram-se os pontos que não possuíam leitura de N.NO3, assim tendo cerca de 0.78% dos pontos DOCeq faltantes e 10.8% dos pontos Abs210. Os pontos removidos podem ser vistos no mapa da Figura 9 em vermelho. Conforme mencionado anteriormente, não há registros do cruzeiro sobre outras variáveis que não salinidade e temperatura externa por serem as variáveis alvo do seu estudo. Contudo, algumas possibilidades podem ter acontecido: o sensor podia não estar corretamente calibrado até aquele momento, pois era o início da viagem, ou as águas eram em demasiado turvas para a absorção em 210 nm. Essa segunda possibilidade é plausível, pois o manual do equipamento aponta limites de qualidade de absorção para o cálculo da variável N.NO3, porém a hipótese da calibração também não pode ser descartada. Há ainda alguns poucos pontos intercalados nas três regiões, porém não consecutivos, sendo os pontos ao redor suficientes para manter o padrão de cada seção.

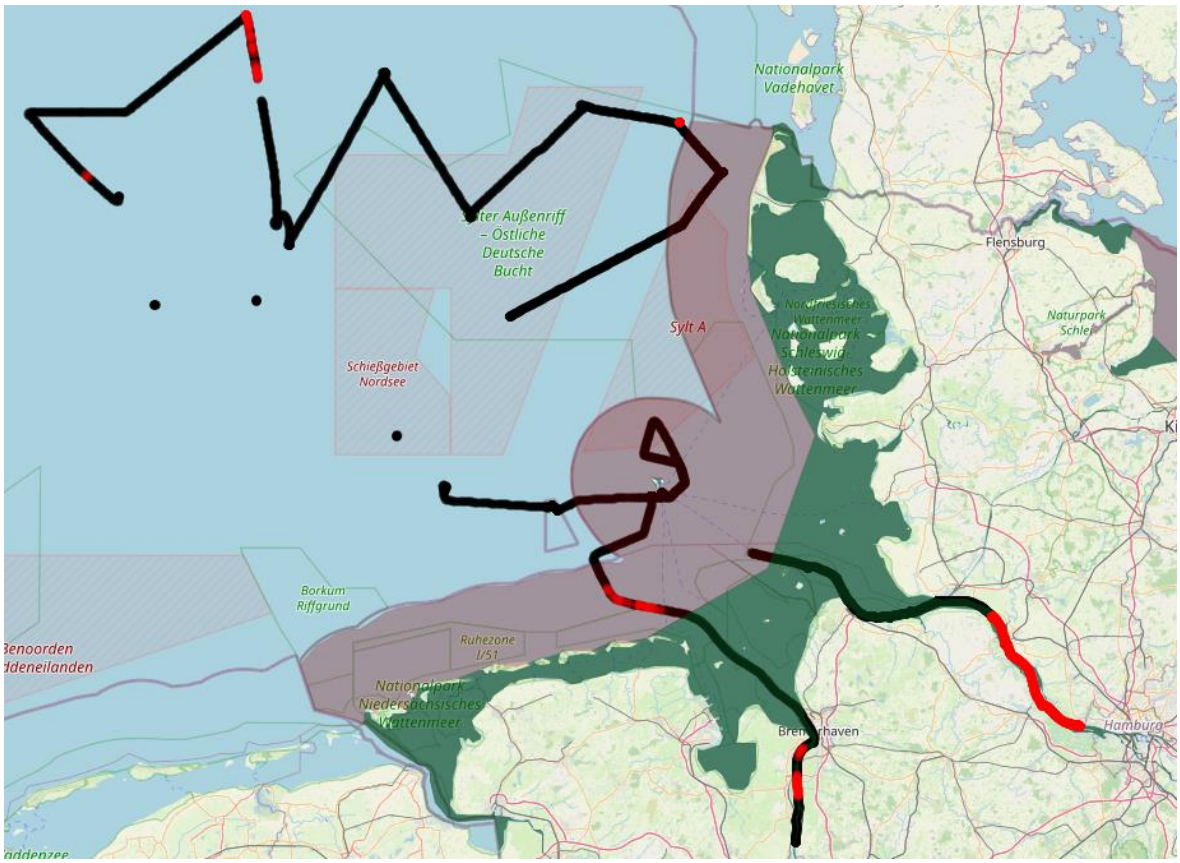


Figura 9: Observações com dados de N.NO3 ausentes removidos do conjunto de dados (em vermelho).

Para os pontos incompletos remanescentes, procedeu-se com a interpolação espacial dos dados. Segundo Longley (2005), a interpolação espacial é um processo de adivinhação inteligente, no qual o investigador procura uma estimativa razoável do valor de um campo contínuo em locais onde o campo não foi realmente medido (Longley & Et. Al., 2005). Em suas diversas aplicações, o único princípio subjacente a toda a interpolação espacial é a Lei de Tobler, citada no início do capítulo 2. O método escolhido para interpolação espacial foi a ponderação pela distância inversa (*IDW*). Nesse método, os valores em locais não amostrados são estimados como a média ponderada dos valores do restante dos locais com pesos inversamente proporcionais à distância entre os locais não amostrados e os amostrados. O cálculo é realizado da seguinte maneira (Moraga, 2023)

$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{n} Z(s_i) \times \left(1/d_i^{\beta}\right)}{\sum_{i=1}^{n}\left(1/d_i^{\beta}\right)} = \sum_{i=1}^{n} Z(s_i) \times \mathrm{w_i} \qquad (31)$$

onde $\hat{Z}(s_0)$ é o previsto em s_0, n é o número de locais amostrados, $Z(s_i)$ é o valor na localização s_i, e, por fim, d_i é a distância entre a localização s_i e s_0 onde se deseja prever. Os pesos são dados pelo termo $w_i = \frac{1/d_i^{\beta}}{\sum_{i=1}^{n}\left(1/d_i^{\beta}\right)}$, sendo β a potência da distância que determina o grau no qual distâncias mais próximas são preferidas sobre locais mais distantes.

Ao final desses dois passos do pré-tratamento, o conjunto de dados possuía cerca de 11750 observações, agora completas para todas as variáveis.

A próxima etapa do pré-processamento foi a consistência dos dados em relação as medições sequenciais. As diferenças entre medições sequenciais, onde uma medição é bastante diferente das adjacentes, apresentando um pico na grandeza e no gradiente. Esse tratamento é orientado pelos Procedimentos de Controle de Qualidade de Dados da SeaDataNet (SeaDataNet, 2010).

O teste para grandeza, conhecido como teste de pico (*spike test*), deverá falhar se uma observação se afastar muito dos pontos próximos após o conjunto de dados ser temporalmente ordenado. Assim, aplica-se aos pontos de uma variável o seguinte valor de teste

$$\text{Test value} = |\,\mathrm{V}_2 - \frac{\mathrm{V}_3 + \mathrm{V}_1}{2}\,| - |\frac{(\mathrm{V}_3 - \mathrm{V}_1)}{2}\,| \quad (32)$$

onde V_2 é o valor testado como pico, V_3 a observação seguinte e V_1 a observação anterior. As recomendações são que o valor de teste não deva exceder 0.9 PSU para salinidade e 6.0ºC para temperatura (SeaDataNet, 2019; Argo, 2009). Para as demais variáveis sem referência encontrada na literatura, tomou-se o valor do cálculo de pontos extremos, ou seja, a variação de $1.5 \times IQR$.

O valor tomado para este cálculo representa uma porção considerável da amplitude dos dados. É preciso considerar que ambos os testes, tanto o de pico quanto o de gradiente, consideram a variação local e não variação considerada de todos os pontos da variável. Mesmo que alguns valores se aproximem dos limites estabelecidos nesses testes e possam ser potenciais picos e gradientes que não correspondem a dados considerados bons para análise, sejam eles por erro de medição ou por alguma condição temporária anômala no local da tomada das medidas, tais picos ou gradientes serão dispersos quando tomadas a média por minuto das medições. Assim procuramos aqui eliminar os casos acentuados.

Os *outliers* das variáveis não serão removidos nesta etapa, pois podem representar pontos que apresentam uma concentração significativamente alta de fenômenos ou características a serem percebidas em comparação com as áreas circundantes que serão analisadas através da investigação de autocorrelações a partir do índice global de Moran. Em suma, tratando-se de uma análise espacial, utilizou-se apenas os testes de pico e gradiente para identificar pontos extremos em relação a seus vizinhos imediatos, e outliers espaciais por variável poderão ser identificados ao final através do diagrama de dispersão de Moran conforme referido no capítulo 2.

O teste de gradiente refere-se à taxa de mudança entre valores consecutivos em um conjunto de dados ordenados. Ao analisar como os valores mudam de um ponto de dados para o próximo, é possível detectar desvios que são inesperados ou fora dos padrões normais (Argo, 2009). Por exemplo, em um conjunto de dados que registra temperaturas ao longo do tempo, uma mudança abrupta e significativa de um valor para o próximo pode

indicar um erro de medição ou um problema com o sensor de temperatura. O valor de teste é calculado conforme a expressão

$$\text{Test value} = |\, \mathrm{V}_2 - \frac{\mathrm{V}_3 + \mathrm{V}_1}{2} \,| \qquad (33)$$

onde V_2 é o valor testado como pico, V_3 a observação seguinte e V_1 a observação anterior. As recomendações são que o valor de teste não deva exceder 1.5 PSU para salinidade e 9.0ºC para temperatura (SeaDataNet, 2019; Argo, 2009). Para as demais variáveis sem referência encontrada na literatura, tomou-se o valor de teste igual a 20% do alcance dos dados.

Argo (2009) e SeaDataNet (2019) estabelecem que pontos que falham nos testes de pico e gradiente devem ser sinalizados no conjunto de dados e tratados com cautela ou, se possível, considerados pontos ruins e descartados. Dependendo do contexto, o analista pode optar por realizar análises mais aprofundadas ou, se entender que valores reprovados são consequência de medidas ruins ou erro de medidas, inferir novos valores através de métodos de interpolação (IOC/IODE, 1993).

O número de observações invalidadas pelos testes de pico e gradiente para o conjunto de dados deste este estudo estão na Tabela 1.

Tabela 1: Número de observações invalidadas pelos testes de pico e gradiente.

Teste	***Salinity***	***TempExtern***	***N.NO3***	***DOCeq***	***Abs210***	***Abs254***	***Abs360***	***SAC254***
Pico	20	0	51	42	122	17	17	5
Gradiente	11	0	70	36	87	14	12	11

É preciso considerar que alguns pontos invalidados para determinadas variáveis são pertencentes a mesma leitura (observação) no conjunto de dados. O total de observações removidas que obtiveram sinalização somente de pico foi de 77, mostrando que algumas observações ainda foram marcadas como reprovadas em ambos os testes, totalizando 45 observações desse tipo. Ainda, 151 observações foram marcadas como unicamente reprovadas no teste de gradiente. O total de observações ruins removidas foi de 273.

O número de observações invalidadas pelos testes de consistência de dados em medições sequenciais corresponde a 2.3% do total do conjunto de dados. Assim, procedeu-se com a remoção de tais pontos e ignorou-se a possibilidade de interpolação para ajuste de dados. O total de observações ficou em 11470.

Para os passos finais do pré-tratamento de dados, foram consideradas as varáveis de tempo *DateTime* e o deslocamento do navio *Lat* e *Lon*. Foram tiradas as médias dos dados por minuto. Como descrito anteriormente, os dados foram tomados com alguns segundos de diferença em sua maioria. Procedeu-se com o cálculo da média por minuto de cada variável nas observações, deixando o conjunto de dados com cerca de 7700 observações. Em seguida foram utilizadas as geocoordenadas das observações e, para garantir um deslocamento mínimo da embarcação entre duas observações consecutivas, foi estipulado uma distância mínima de 50 metros entre elas. Caso a distância seja menor, a observação seguinte é descartada e a próxima é avaliada se está dentro deste critério.

Assim concluímos a etapa de pré-processamento, tendo os dados completos para todas as variáveis, consistentes, com separação mínima temporal de 1 minuto e espacial de 50 m. O total de observações no conjunto de dados ao iniciarmos a análise foi de 2772.

5. PRIMEIROS RESULTADOS DA AEDE

5.1. Análise Exploratória Inicial por Zona

Cada zona (ver seção 4.1) foi analisada individualmente para entender como as variáveis estariam distribuídas. Para tanto, calcularam-se as estatísticas de cada região, apresentadas na Tabela 2.

Tabela 2: Súmula Estatística por Zona

Zona "Águas Internas" (726 observações)								
	Salinity	**TempExtern**	**N.NO3**	**DOCeq**	**Abs210**	**Abs254**	**Abs360**	**SAC254**
Média	14.067	6.0	1.5808	8.754	3.7184	1.0874	0.7019	19.276
Min	0.509	4.5	0.0006	1.137	0.9493	0.1717	0.0986	1.413
Q1	4.951	5.6	0.4938	4.032	2.9151	0.5362	0.3374	9.151
Med	11.982	6.0	1.16226	8.252	3.8640	0.9325	0.5343	19.204
Q3	26.107	6.4	2.1989	11.610	4.4086	1.4142	0.8899	25.325
Max	32.646	7.7	12.5532	37.498	6.37	4.3996	4.3714	52.966
IQR	21.156	0.8	1.7051	7.578	1.4935	0.878	0.5525	16.174
DesvP	10.773	0.575	1.196	5.685	0.829	0.723	0.530	10.858

Zona "Mar Territorial" (639 observações)								
	Salinity	**TempExtern**	**N.NO3**	**DOCeq**	**Abs210**	**Abs254**	**Abs360**	**SAC254**
Média	33.04	8.464	0.05463	1.0129	2.376	0.1393	0.211	5.391
Min	31.01	6.8	0.0002	0.6041	2.257	0.0972	0.0462	2.317
Q1	32.98	8.5	0.04135	0.8347	2.34	0.1132	0.0619	2.559
Med	33.22	8.7	0.0528	0.8922	2.37	0.1176	0.0653	2.668
Q3	33.44	8.8	0.06105	1.1311	2.406	0.1594	0.07885	3.023
Max	34.17	9.4	0.1538	2.1929	2.583	0.3189	0.09245	3.366
IQR	0.46	0.3	0.0197	0.2964	0.066	0.0462	0.01695	0.464
DesvP	0.644	0.578	0.027	0.279	0.049	0.040	0.026	0.699

Zona "Mar do Norte" (1407 observações)								
	Salinity	**TempExtern**	**N.NO3**	**DOCeq**	**Abs210**	**Abs254**	**Abs360**	**SAC254**
Média	34.68	10.1	0.1671	4.7693	4.7179	2.1162	2.0756	16.455
Min	29.97	7.8	0.0038	0.0688	0.4764	0.0954	0.0506	0.1086
Q1	33.92	9.1	0.03005	0.7379	2.352	0.1053	0.0604	2.2081
Med	34.13	9.7	0.0402	0.8236	2.3879	0.1088	0.063	2.2867
Q3	34.04	9.489	0.04156	0.755	2.4086	0.1608	0.1131	2.3875
Max	34.32	9.9	0.0504	0.8676	2.4301	0.143	0.0901	2.3977
IQR	0.12	0.389	0.01151	0.0171	0.0566	0.0555	0.0527	0.1794
DesvP	0.406	0.542	0.018	0.311	0.256	0.202	0.201	0.598

As tabelas indicam que, com base nos valores de desvio padrão e na amplitude interquartil, as zonas MT e MN apresentam dados cujos valores são mais similares entre si, indicando uma maior proximidade ou compactação.

Podemos confirmar essa distribuição através do gráfico *boxplot* na Figura 10, onde temos a representação das distribuições das variáveis. Para podermos comparar todas as diversas variáveis simultaneamente, a padronização dos dados foi realizada. Isto porque, quando variáveis são expressas em diferentes unidades, a magnitude dos valores de algumas pode afetar de forma desproporcional as distâncias calculadas entre observações, prejudicando a análise em favor dessas variáveis de maior magnitude (Fávero & Belfiore, 2017). Uma discussão mais detalhada sobre a padronização de dados e seu cálculo foram escritos na Seção 3.1 acima.

É possível perceber as caixas no gráfico para as zonas mais compactadas MT e MN para a maior parte das variáveis. O gráfico não somente indica que a medida de dispersão estatística é semelhante, mas também que a grandeza dos valores padronizados em si são próximos em magnitude. Isto também pode ser constatado pelos valores próximos das médias e medianas na Tabela 2 para maioria das variáveis. Concluímos então que as zonas MT e MN são similares em amplitude para as variáveis e que a região AI apresenta uma distribuição estatística dissimilante. Também é possível perceber que as variáveis se parecem muito em sua distribuição quando padronizadas. Isto pode indicar que são variáveis altamente correlacionadas, o que iremos verificar mais adiante na análise de correlograma.

Sobre os *outliers*, conforme mencionado, não serão removidos nesta etapa, pois podem representar pontos com determinada característica agrupados espacialmente e que poderiam ser identificados na análise de autocorrelação espacial.

Vemos ainda no gráfico que a salinidade aumenta no sentido da zona AI para a MN, conforme esperado e que a salinidade na MT já é próxima da MN. A variável *TempExtern* foi a que melhor se caracterizou para cada zona, sendo distinta na MT e na MN, indicando o aumento da temperatura na direção da MN.

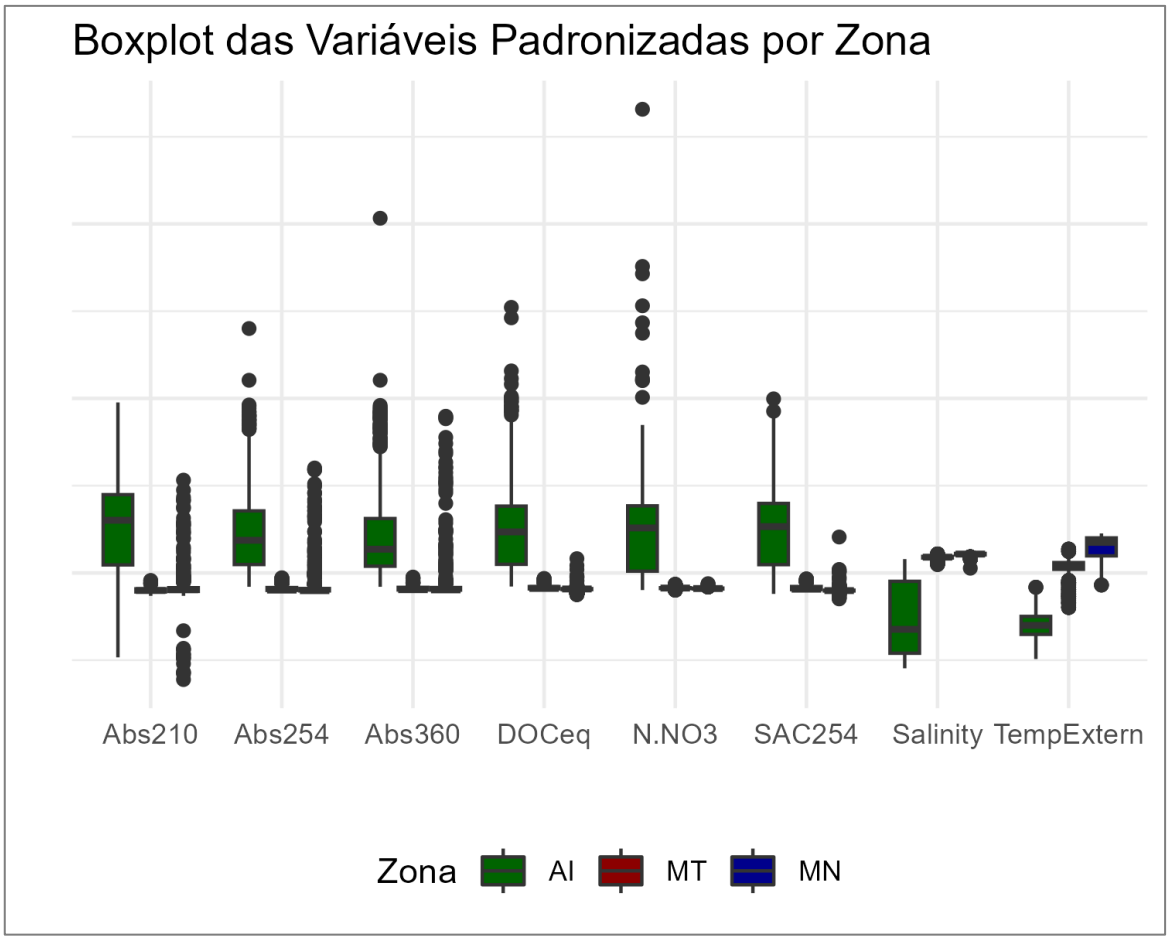


Figura 10: Distribuição das variáveis padronizadas por zona - Bloxplot

Os gráficos *boxplots* individuais para cada variável não padronizada encontram-se no Apêndice I para maior detalhamento, separados por zona.

5.2. Análise Exploratória Inicial dos Dados Globais

Ao final da análise inicial proposta por zonas, ainda podemos investigar o conjunto de dados sem esta separação e verificar quais características predominam para os dados globais. Assim, apresenta-se na tabela o resumo estatístico global na Tabela 3.

Tabela 3: Sumário Estatístico dos Dados Globais

Global (2772 observações)								
	Salinity	**TempExtern**	**N.NO3**	**DOCeq**	**Abs210**	**Abs254**	**Abs360**	**SAC254**
Média	28.579	8.339	0.4477	2.9094	2.7442	0.3985	0.2594	6.9573
Min	0.509	4.5	0.0002	0.0688	0.4764	0.0954	0.0462	0.1086
Q1	30.945	6.9	0.0371	0.8086	2.3593	0.1082	0.0617	2.2848
Med	33.433	8.8	0.0517	0.8752	2.4074	0.1311	0.0772	2.5744
Q3	34.133	9.7	0.1353	2.066	2.6175	0.3478	0.2146	5.2579
Max	34.676	10.1	12.5532	37.4984	6.37	4.3996	4.3714	52.966
IQR	3.188	2.8	0.0982	1.2574	0.2582	0.2396	0.1529	2.9731
DesvP	0.4063	0.5418	0.0178	0.3108	0.2565	0.2023	0.2007	0.5980

Conforme esperado, o sumário estatístico indica valores mais parecidos com as zonas preponderantes MT e MN, com desvios padrão e IQR baixos. A predominância dessas características é confirmada pelo gráfico de *boxplot*, que também indica o maior IQR para temperatura, sendo esta mais espalhada ao longo de sua amplitude, e pode ser visualizado na Figura 11.

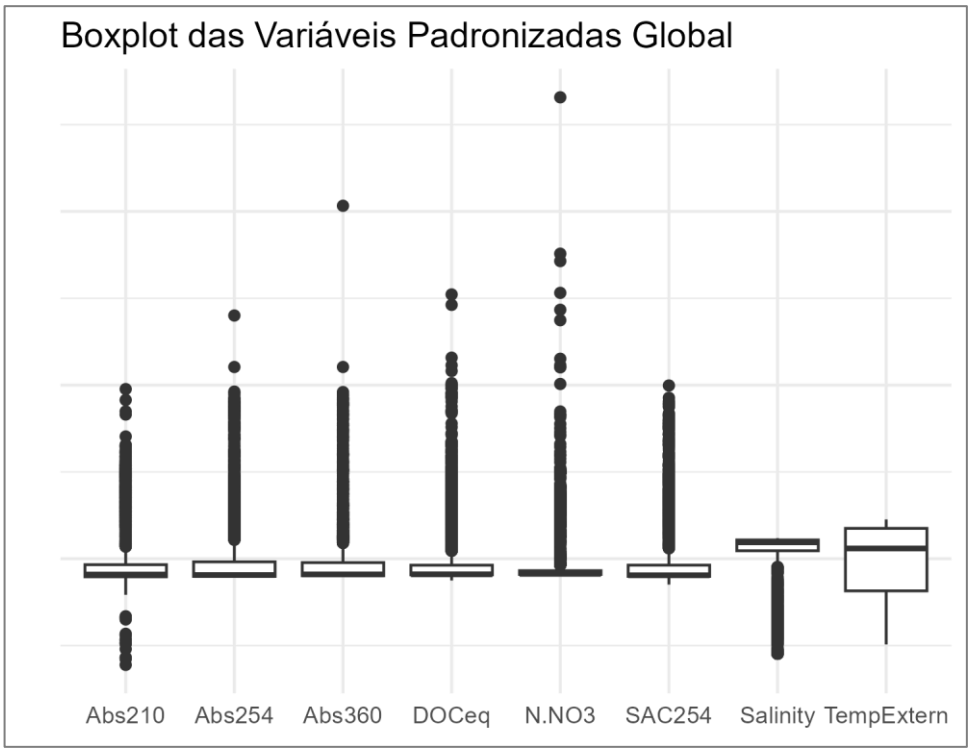


Figura 11: Boxplots das variáveis padronizadas globais

Após a observação da similaridade da distribuição e da variabilidade dos dados, prosseguiu-se com a construção de gráficos de correlação chamados de correlogramas conforme referencial teórico apresentado na seção 3.1 e apresentados na Figura 12. O coeficiente de correlação utilizado foi o de Pearson. O p-valor associado a esta correlação avalia a hipótese nula de correlação zero ou, em outras palavras, nenhuma correlação linear. Se o p-valor for inferior à significância considerada, sugere-se que é improvável que a correlação observada tenha ocorrido por acaso, sendo, portanto, considerada estatisticamente significativa (Zar, 2019). Valores para a significância de 0.005, 0.01 e 0.05 foram testados e nos três casos os resultados foram similares. Assim, o valor adotado foi de 0.05 para este estudo.

Na análise destes por zonas, percebemos que temperatura e salinidade apresentam correlação moderada em AI, porém significativamente forte para as outras duas regiões. Isto pode ocorrer pois talvez ambas variam diferentemente à medida que se avança para o Mar do Norte, ou podem também ter variação diferenciada nas águas mais distantes do mar, entrando no continente. Elas ainda apresentam correlação negativa moderada à forte para todas as outras variáveis. N.NO3 e Abs210 e possuem correlação de magnitude similar com as outras variáveis, e moderada entre elas. As variáveis de absorção apresentaram correlação fortemente positiva entre elas, o que pode ter influência da turbidez da água (TriOS GmbH, 2017).

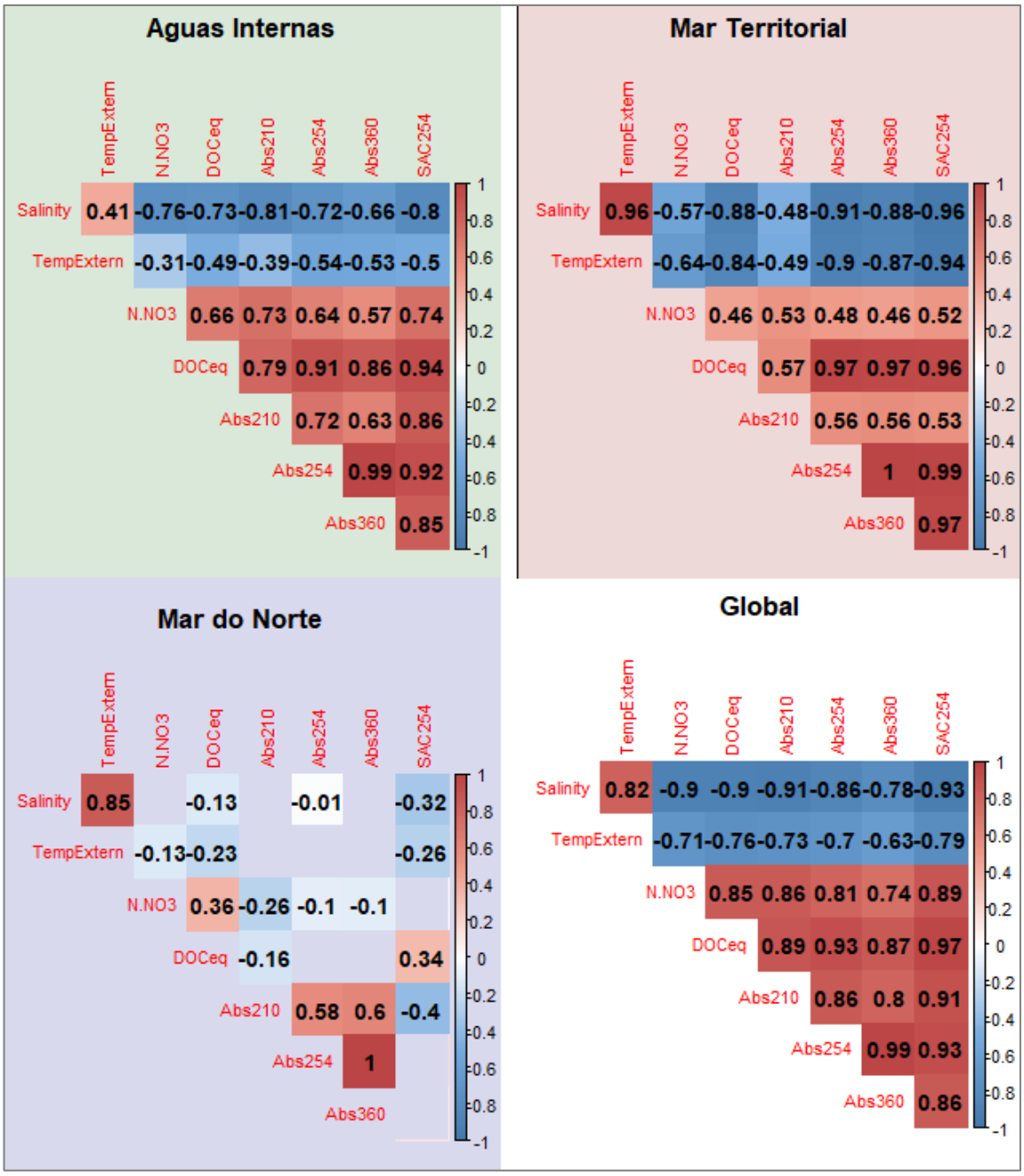


Figura 12: Correlogramas por zona e global

A região do MN apresentou correlações fracas e/ou não estatisticamente significativas para alguns pares de variáveis. Este resultado pode ser explicado pela compactação dos dados naquela zona. *N.NO3*, *DOCeq* e *SAC* são consideravelmente compactados e em sua pequena variação não observação de correlação linear entre elas. Ainda no MN, *Salinity* e *TempExtern* apresentam correlação forte positiva para esta zona, indicando uma correlação linear significativamente mais forte que na zona AI.

O correlograma global indicou que, ao considerar o conjunto completo de dados, as variáveis apresentam uma forte correlação entre si. Isso é esperado, pois ao observamos os gráficos *boxplot* gerados para os dados por zonas, vemos as variáveis *Salinity* e *TempExtern* aumentar de maneira semelhante à medida que analisamos da zona AI para MN. Observamos também que todas as demais variáveis exibem padrões de comportamento parecido, porém diminuindo conjuntamente. No entanto, elas se diferenciam por estarem muito mais compactadas nas áreas MT e MN.

Desta análise inicial dos dados globais, podemos então concluir que temos majoritariamente dois grupos de variáveis que fortemente correlacionam positivamente entre si: o grupo com *Salinity* e *TempExtern* e o grupo das demais variáveis. Também vemos que entre esses dois grupos há uma correlação fortemente negativa.

6. A REDUÇÃO DA DIMENSIONALIDADE DOS DADOS - PCA

Após a construção dos correlogramas, sendo gráficos baseados na matriz de correlações, observamos que as variáveis possuem uma estrutura de alta correlação que a PCA pode explorar, segundo a seção 3.1 do referencial teórico. Ainda, aquela seção esclarece que o uso da matriz de correlação é preferida em PCAs devido à sua invariância à escala das variáveis, assim permitindo a comparabilidade entre diferentes análises. Desta forma, abordaremos as PCAs de cada zona e dos dados globais, bem como faremos comparações e interpretações, em busca da simplificação dos dados originais. Ainda, de acordo a mesma seção (3.1), foi realizada a padronização dos dados e estes foram utilizados nas análises que se sucedem.

Para avaliar a qualidade da PCA foram utilizados a estatística KMO que indicou valor superior a 0.8 e o teste de Bartlett para esfericidade obtendo-se p-valor < 0.01, sendo os dados apropriados para PCA.

6.1. PCA de Zonas

Procedeu-se com a PCA para as diferentes zonas em busca de entendermos se o padrão de agrupamento nas PCs para as diferentes zonas se mantém ou se diferencia. Assim, obteve-se os resultados apresentados nas tabelas a seguir:

Tabela 4: Heatmap - Loadings da PCA por zona

	Águas Internas		Mar Territorial		Mar do Norte	
	PC1	PC2	PC1	PC2	PC1	PC2
Salinity	-0.3526	-0.2834	-0.3772	0.11432	0.20344	-0.4937
TempExtern	-0.2371	0.71746	-0.3759	0.01752	0.30041	-0.4558
N.NO3	0.32304	0.44178	0.24799	0.70834	-0.1808	0.13156
DOCeq	0.38411	-0.0428	0.37805	-0.1823	-0.0476	0.40351
Abs210	0.35595	0.30206	0.25221	0.5962	0.48835	0.01666
Abs254	0.38692	-0.1942	0.38763	-0.1816	0.53044	0.31932
Abs360	0.36518	-0.2789	0.38144	-0.1862	0.54275	0.29665
SAC254	0.39722	0.0338	0.39066	-0.1689	-0.1378	0.42204

Tabela 5: Variâncias e porcentagem da variabilidade por zona

	Águas Internas		Mar Territorial		Mar do Norte	
	PC1	PC2	PC1	PC2	PC1	PC2
Variância	5.964	0.843	6.353	0.852	2.613	2.17
% da Variância	74.545	10.542	79.408	10.644	32.664	27.119
% Variância acumulada	74.545	**85.087**	79.408	**90.052**	32.664	**59.784**

Na zona AI, os loadings indicam que nenhuma variável ou grupo de variáveis é majoritariamente carregada no PC1, embora esta contribua com cerca de 74.5% do total da variabilidade dos dados originais. No PC2, destaca-se a os carregamentos das variáveis *TempExtern* (~ 0.7), seguida de N.NO3 (~ 0.45). Os dois primeiros PCs somados representam 85% da variabilidade das variáveis originais.

Na zona MT, os loadings no PC1 indicam um comportamento similar à AI. Porém, o PC2 representou mais eficazmente as variáveis N.NO3 (~ 0.7) e Abs210 (~ 0.6).

Os dois primeiros componentes somados conservam 85% da variabilidade das variáveis originais.

Houve diferenças significativas na zona MN, conforme previsto na análise do correlograma (Figura 12). Vimos que a maior parte das variáveis não apresentava correlação linear significativamente forte e sabemos da seção 3.1 que essa característica diminui a efetividade da PCA em representar a variabilidade dos dados em um número pequeno de PCs. Vemos que a % da variância explicada agora está mais dispersa entre os dois PCs escolhidos e que o total da variabilidade explicado foi de cerca de 60%. Na observação dos loadings por componente, vemos que o PC1 foi em maior parte representada pelas variáveis de Absorção e que a variabilidade dos pares *Salinity/TempExtern* e *DOCeq/SAC254* carregam a segunda variável com sinais opostos. Voltando ao correlograma para esta zona, vemos as variáveis de absorção se correlacionando de moderada a fortemente, explicando sua participação no PC1. Os pares que contribuem para o PC2 também se correlacionam positivamente entre seus membros, porém negativamente entre os pares.

Apresentaremos em seguida os gráficos da PCA conforme seção 3.3.

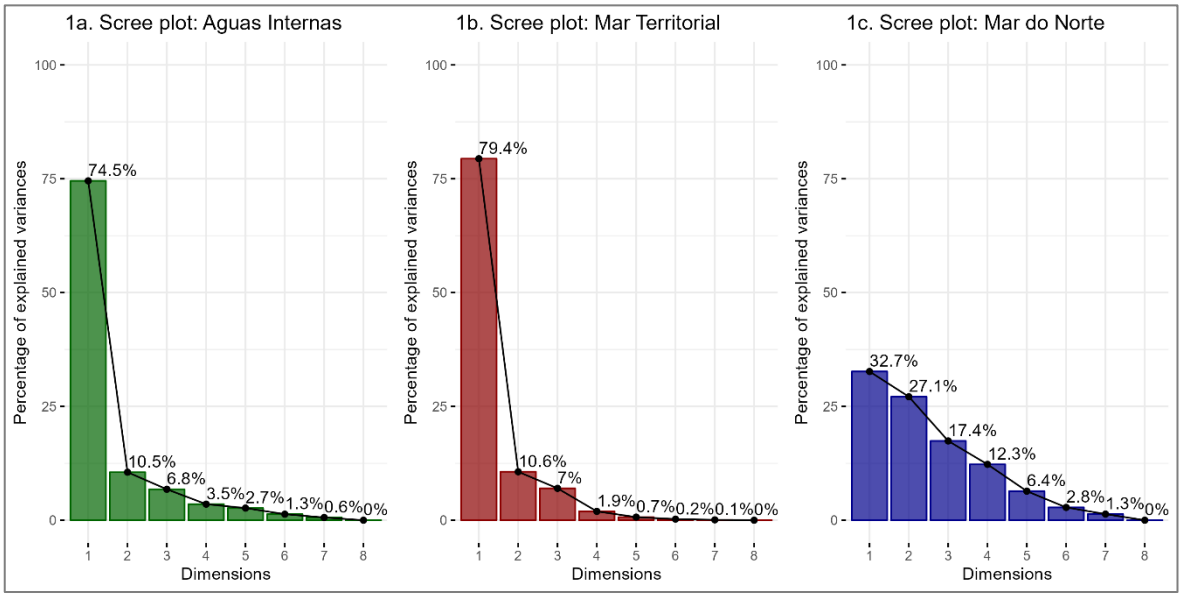


Figura 13: Gráficos scree das PCAs por zona

Os gráficos *scree* mostram claramente a predominância do PC1 nas zonas AI e MT. Nestas zonas, as observações das variáveis mostraram-se fortemente correlacionadas

linearmente entre si nos seus correlogramas, o que explica o resultado, seção 3.1. Com menor correlação linear, mas significativa, entre as variáveis, a variabilidade das observações no MN ficou mais dispersa entre as PCs.

Passemos agora a analisar a qualidade da representatividade dos dados.

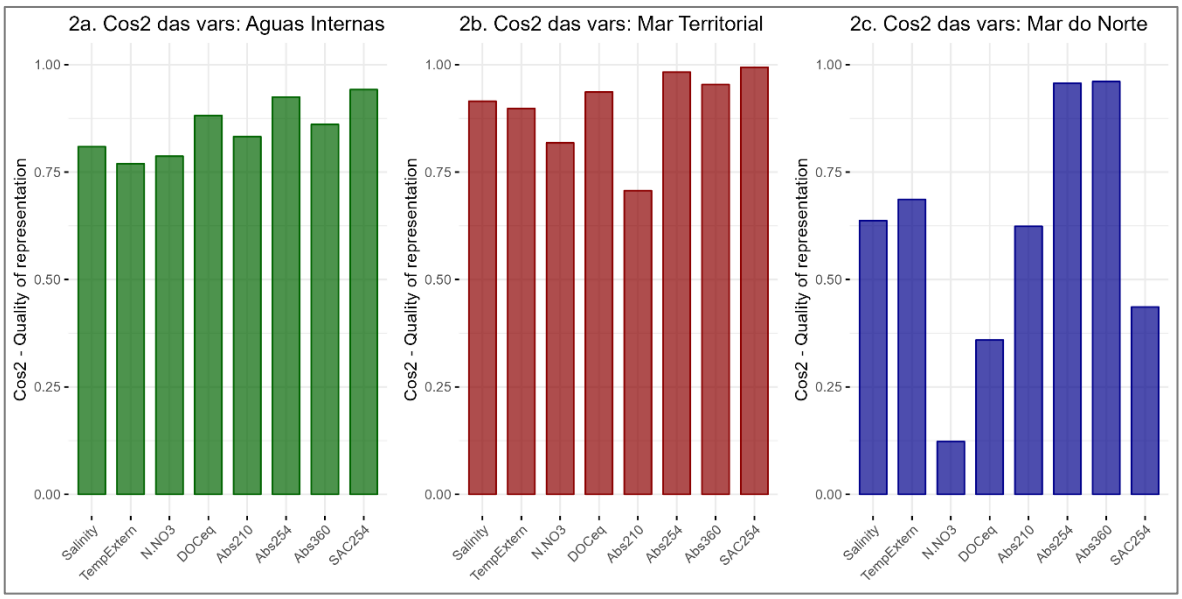


Figura 14: Cos² das variáveis por zona

Os valores do $\cos^2$ para a representação da variabilidade dos dados indicam um excelente resultado para os dois primeiros PCs em AI e MT, retomando que este valor possui máxima de 1. Logo, aqueles componentes foram capazes de representar adequadamente a variabilidade de todas as variáveis sem a necessidade de agregarmos mais PCs. A variável *Abs210* foi a menos representada, pois apresentou menor correlação linear com outras variáveis, porém ainda moderada (~ 0.5).

Por não ter demonstrado uma estrutura de dados com correlações lineares tão robustas quanto as das duas primeiras zonas, a zona MN não teve a variabilidade de todas as suas variáveis representadas adequadamente. Vemos que a variabilidade das variáveis de absorção *Abs254* e *Abs360* são bem representadas pelos dois primeiros PCs. Isto ocorreu pois estas são parte do grupo de variáveis que bem se ajustaram ao PC1 e por terem correlação linear forte entre si, maior do que *Abs210*. Também encontramos *Salinity* e *TempExtern* com uma boa qualidade de representação, onde o PC2 demonstrou valores

consideráveis de carregamento para estas variáveis. *N.NO3* foi a variável que apresentou qualidade de representação mais baixa. Percebemos esta variável com fraca correlação linear com todas as outras variáveis, sendo preciso considerar um maior número de PCs para incluir significativamente sua variabilidade nos resultados.

6.2. PCA dos Dados Globais

Os resultados da PCA para o conjunto de dados globais são demonstrados a seguir, seguindo estrutura similar da seção anterior.

Tabela 6: Heatmap - Loadings (global)

	Global	
	PC1	PC2
Salinity	-0.3631	0.2442
TempExtern	-0.3122	0.65956
N.NO3	0.34589	-0.1737
DOCeq	0.36757	0.09276
Abs210	0.35708	-0.0467
Abs254	0.36301	0.39596
Abs360	0.34225	0.55452
SAC254	0.37355	0.01002

Tabela 7: Variâncias e % da variabilidade

	Global	
	PC1	PC2
Variância	6.912	0.47
% da Variância	86.402	5.876
% Variância acumulada	86.402	**92.278**

O PC1 apresenta a mesma separação das variáveis encontradas no correlograma. Esta característica é esperada pela forte correlação linear que os dados apresentam quando comparados de maneira global. Esta forte correlação também justifica a grande porção da variabilidade explicada no primeiro PC, 86.4%. O PC2 foi caracterizado pelas variáveis *TempExtern, Abs254* e *Abs360*, porém esta representou apenas 5.9% da porcentagem da variância total.

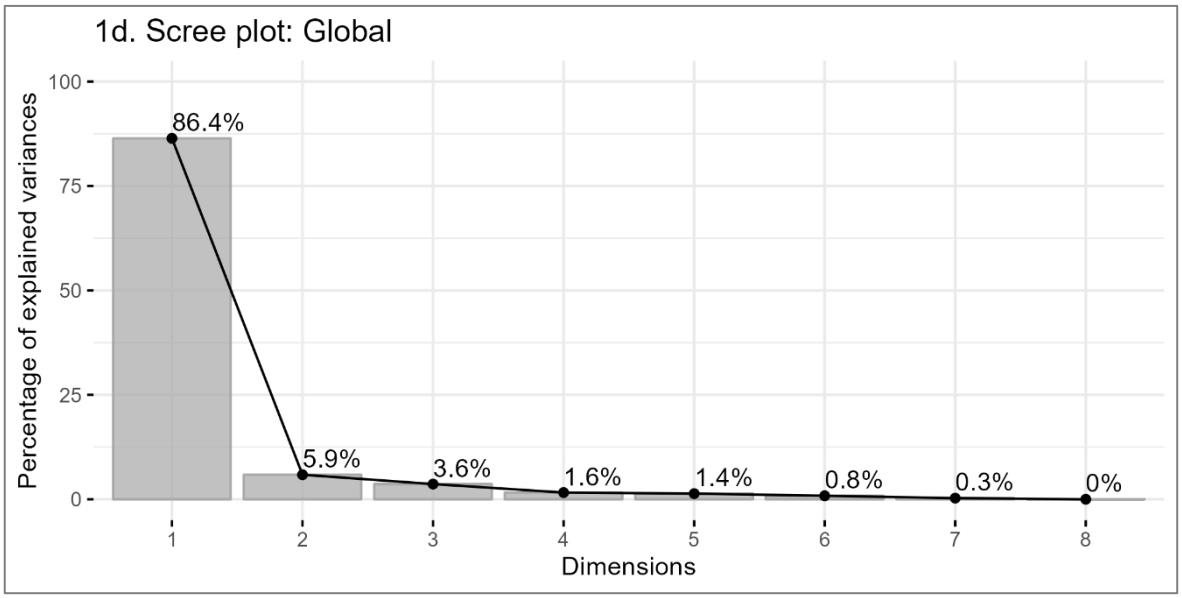


Figura 15: Gráfico scree da PCA dos dados globais

O gráfico *scree* indica como a forte correlação linear carregou fortemente o PC1. Ele também mostra que considerar outras PCs não irá influenciar largamente a representação da variabilidade dos dados globais originais.

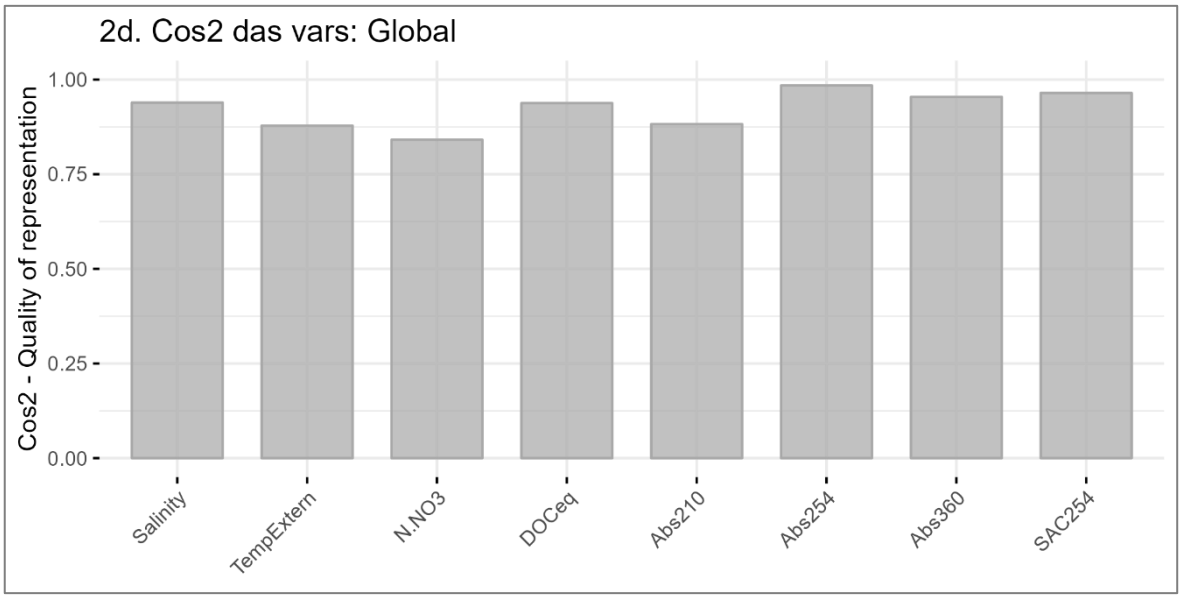


Figura 16: Cos² das variáveis dos dados globais

Sobre a qualidade da representação, o gráfico do Cos^2 das variáveis indica que a variabilidade de todas as variáveis do conjunto de dados é fortemente representada pelos dois primeiros PCs.

6.3. Visualização e Análise dos Resultados da PCA

Foi plotado o gráfico do PC1 x PC2 (gráficos *biplot,* conforme mencionado na seção 3.3), com a indicação de $\cos^2$ para cada PCA (zonas e global). Este gráfico auxilia a visualização das variáveis (ou seja, os vetores de variáveis) no espaço dos PCs. Assim:

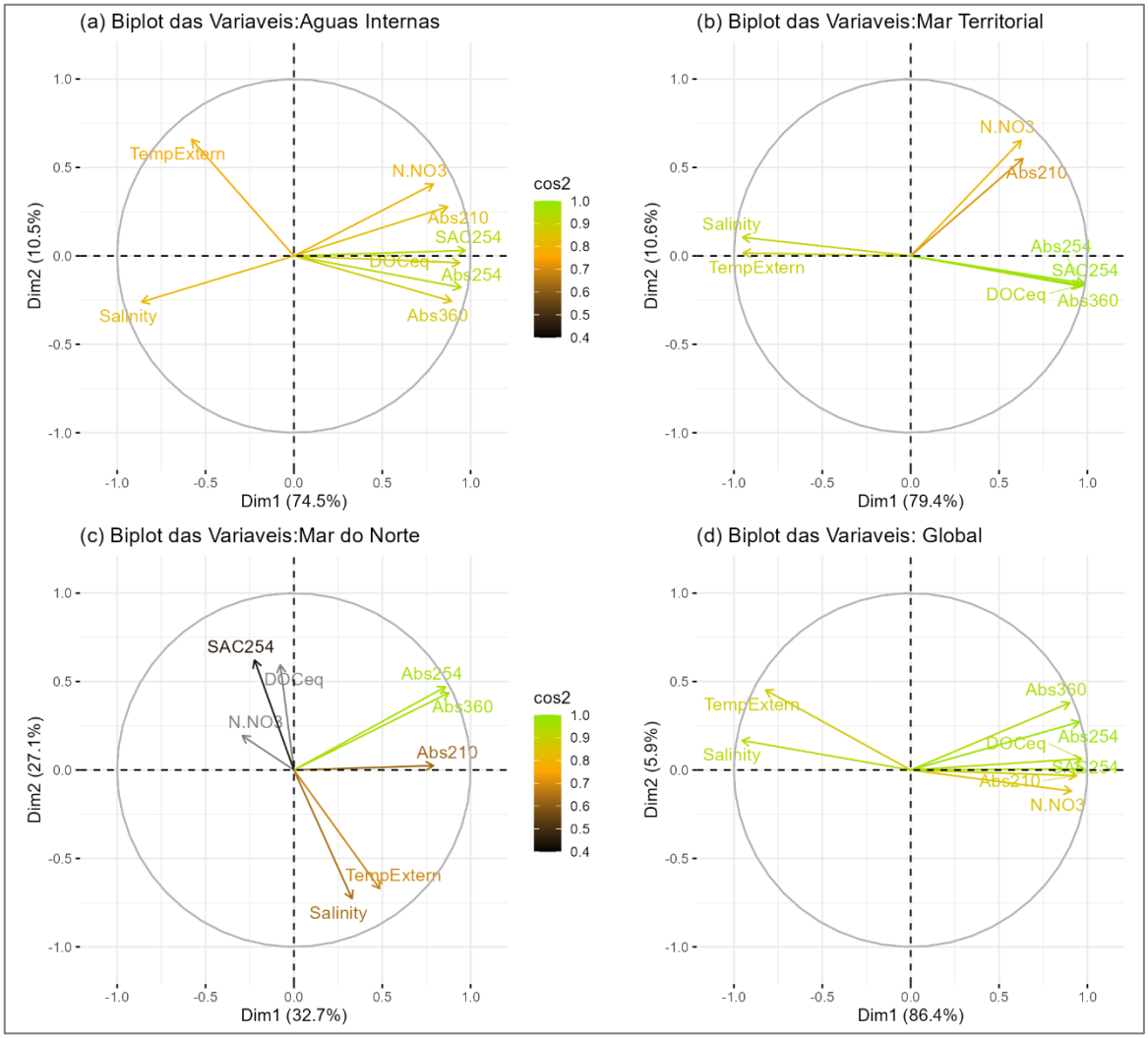


Figura 17: Gráfico biplots das variáveis com cos²

No eixo das abscissas é representado PC1 e no eixo das ordenas o PC2. Os vetores das variáveis mostram a direção e a magnitude no espaço definido pelos PCs selecionados. Isso ajuda a interpretar a relação entre as variáveis e os PCs. A direção do vetor indica a direção da máxima variação influenciada pela variável, e o comprimento do vetor sugere a força da correlação da variável com o componente principal. A função utilizada plota um círculo de correlação, facilitando a avaliação da correlação entre as variáveis e os PCs. Variáveis próximas uma da outra são correlacionadas positivamente, variáveis opostas no círculo são correlacionadas negativamente, e variáveis ortogonais (em ângulo de 90°) são não correlacionadas. Então é possível visualizar a contribuição das variáveis para os PCs, o que ajuda a identificar quais variáveis são mais importantes para explicar a variação capturada por cada PC (Jolliffe, 2002).

Na Figura 17, os gráficos *biplot* AI e MT caracterizam logo as correlações negativas de *Salinity* e *TempExtern* em relação as outras variáveis. Vemos também as variáveis que possuem alta correlação entre si se posicionarem próximas, conforme mencionado. Em ambos os gráficos (AI e MT), as magnitudes dos vetores das variáveis se mostraram longos. As coordenadas correspondem ao valor do *loading* naquele PC multiplicado pelo desvio padrão do PC. Assim, quanto maior a contribuição daquela variável (*loading*) no PC, maior será a magnitude do vetor naquela direção. Observa-se ainda que a maior parte das variáveis ficou próxima ao eixo do PC1, confirmando que este PC se ajusta melhor linearmente aos dados podendo assim representar maior parte da variabilidade dos dados originais. Os valores de $\cos^2$ indicados pelas cores dos vetores confirmam que a representatividade da variabilidade de todas as variáveis foi satisfatória tomando-se apenas 2 PCs.

Na zona MN devemos recordar que as variáveis não se mostraram altamente correlacionadas como nas outras zonas (Figura 12). Assim, os 2 PCs representados não conseguem representar corretamente a variabilidade dos dados de *N.NO3*, *DOCeq* e *SAC254*. Percebemos sua magnitude menor e por consequência seus valores de $\cos^2$ foram baixos. Percebemos no gráfico as fortes correlações mencionadas entre as variáveis de absorção, além da correlação entre *Salinity* e *TempExtern*. Observamos também que as coordenadas das variáveis de absorção são maiores no PC1 indicando que foram

priorizadas nesse componente frente ao outro grupo por obter maiores autovalores, conforme referencial teórico, seção 3.2.

O gráfico *biplot* das variáveis globais indicou claramente as fortes correlações lineares entre as variáveis. Os vetores apresentados se encontram ao redor do eixo do PC1, confirmando sua eficiência na representatividade da variabilidade dos dados originais, tendo pouca perda dessa informação ($\sim 86\%$). Os valores de $\cos^2$ suportam essa afirmação, tendo valores ao redor de 0.9 para maioria das variáveis.

Percebe-se que a análise de PCA apresentou resultados excelentes para a área global dos dados. Por se tratarem de são zonas com perfis de água diferenciado, analisaremos agora o gráfico *biplot* de indivíduos para entender como os PCs estão a representar a variabilidade dos dados originais.

O gráfico de biplot dos indivíduos é apresentado na Figura 18. Este gráfico exibe os *scores* dos PCs que foram gerados a partir da análise dos dados globais. No gráfico percebemos a maior variação observada na zona AI e a grande concentração de valores dos dados nas regiões MT e MN, o que corresponde aos gráficos apresentados na seção 5.1 (**Erro! Fonte de referência não encontrada.** e Figura 10). O gráfico ainda apresenta os v alores de $\cos^2$.

Relembre do Cap. 3 que a contribuição dos indivíduos nos resultados é interpretada em como são projetados no espaço multidimensional das PCs. Iremos analisar como os indivíduos influenciaram na variação capturada pelos PCs. Sendo indivíduos os resultados das combinações lineares das variáveis originais, sua posição relativa próxima pode indicar similaridade em termos das variáveis originais. Ainda, indivíduos com scores altos (positivos ou negativos) em um componente principal têm características que têm uma contribuição significativa para a variação capturada por aquele PC. Indivíduos situados mais distantes da origem no gráfico biplot das PCs contribuem mais para a variação explicada pelos PCs em que estão distantes. Eles são considerados influentes na definição da direção e magnitude dos PCs. Assim, a contribuição de um indivíduo para um PC pode ser medida pela sua contribuição à variação total capturada pelo PC.

Percebemos no gráfico que os pontos foram representados em suas zonas. Como conhecemos os perfis dos dados originais de cada zona (seção 5.1), podemos

comparar as conclusões dessa análise com os dados originais. Assim vemos a maior variabilidade da zona AI demonstrada no decorrer da variação dos indivíduos em PC1 ao mesmo tempo que vemos as zonas MT e MN compactadas para aquele PC. Percebemos isto pois as áreas das zonas mais escuras no mapa representam concentração de dados ou, em outras palavras, as combinações das variáveis originais que são semelhantes. Isso é um forte indício de que elas possuem valores próximos naqueles pontos. Vemos ainda que os centros dos pontos mais claros representam valor cosseno quadrado mais próximos de 1. Nesse contexto, obtemos a contribuição aquele individuo para variância dos PCs plotados. Assim, indivíduos que contribuem pouco podem representar *outliers* ou combinações lineares das variáveis originais que não têm uma contribuição significativa para variância total dos PCs representados. Recordemos que nesta análise a variabilidade do PC1 permitiu representar 86% da variabilidade dos dados originais. Logo, pontos próximos de PC1 = 0 (e pontos próximos a origem), em geral, serão mais escuros por não terem uma contribuição significativa nesse importante PC. Vemos também alguns pontos mais afastados e que são possíveis *outliers,* apresentado cos^2 baixo.

Os *loadings* das PCs também indicam sobre como a variação dos indivíduos está ocorrendo. No PC1 vemos *Salinity* e *TempExtern* com sinais negativos, indicando correlação negativa com o PC1. Portanto, observamos que, à medida que os pontos na zona AI se aproximam das demais zonas e reduzindo o valor do score para o PC1, aquelas variáveis estão a aumentar de valor, como vimos anteriormente. O oposto ocorre no retorno do navio para a zona AI, indicado pelas observações 2400 e 2600 tomadas, mostrando que *Salinity* e *TempExtern* estão a reduzir. Em contrapartida, ainda fazendo a análise dos loadings, por apresentarem correlações positivas com o PC1, as outras variáveis indicam que se reduzem na direção das zonas salgadas MT e MN, porém não variam muito naquelas zonas. A variação no PC2 é significativa para enfatizar a variação da *TempExtern*, que agora é positivamente correlacionada com essa componente, e também as variáveis de absorção Abs254 e Abs360, sendo as três apresentando *loadings* altos naquele PC e, por consequência, forte correlação com ele.

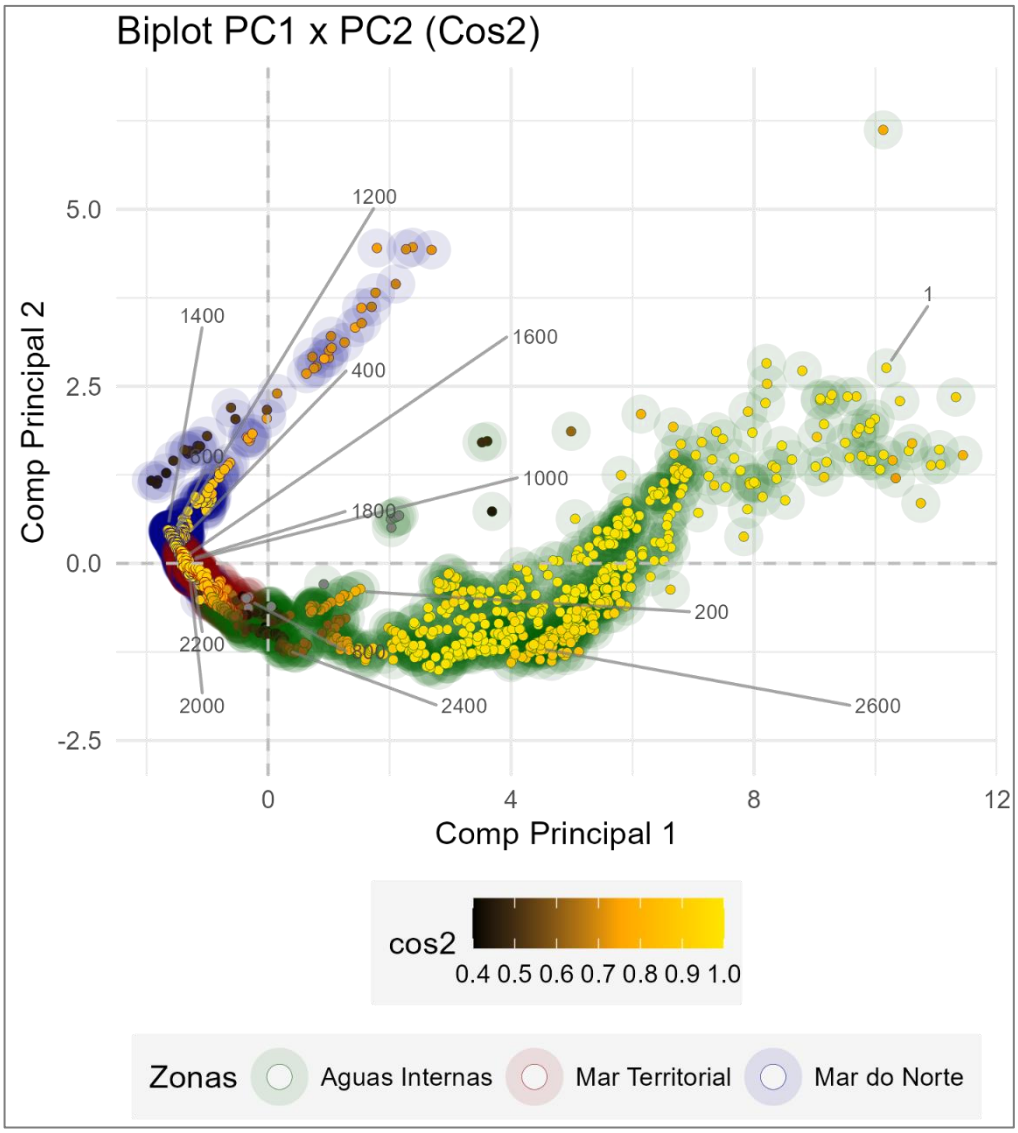


Figura 18: Biplot PC1 x PC2 dos indivíduos por zona com cos² dos dados globais

Em suma, a PCA revelou algumas particularidades dos dados e tirou proveito da forte correlação linear existente entre as variáveis nas zonas AI e MT, bem como nos dados tomados globalmente. Com apenas dois PCs selecionadas, foi possível representar 85%, 90% e 65% da variabilidade dos dados originais nas zonas AI, MT e MN, respectivamente. Nos dados globais, cerca de 92% da variabilidade foi explicada pelos dois primeiros componentes. Assim, o uso do espaço com número de variáveis reduzido se torna útil para simplificação da análise desse conjunto de dados em análises posteriores, tendo a PCA demonstrado ser ferramenta útil para sua simplificação.

7. A ANÁLISE DA AUTOCORRELAÇÃO ESPACIAL

7.1. A Matriz de Pesos Espaciais

Como foi discutido na seção 2.1, a determinação do número de vizinhos para matriz de pesos é um processo crítico e pode afetar consideravelmente a análise da autocorrelação espacial.

A escolha para este estudo foi o método de k-vizinhos próximos (KNN). Foram plotados gráficos com diferentes parâmetros para avaliar as conexões entre os pontos e verificar a existência de pontos isolados ou ilhas que não participem do arranjo total. Os parâmetros escolhidos foram o número de vizinhos próximos (k) e o tempo de separação entre a coleta de dois pontos. No intervalo de tempo de coleta considerado, foram tomadas as medianas das observações por variável. Assim, variáveis que foram observadas com número significativo de *outliers* (vide Figura 10: Distribuição das variáveis padronizadas por zona - Bloxplot) não irão ser enviesadas por eles, como no caso do uso da média do intervalo. Os gráficos, com seus pontos separados nas zonas, são apresentados no Apêndice II.

O arranjo escolhido foi de espaçamento temporal de 60 min e 5 k-vizinhos próximos, exibido na Figura 19. As zonas estão representadas pelas cores conforme legenda. Tal arranjo com 49 pontos permitiu uma cobertura ampla da área navegada, evita aglomerações densas de pontos em uma única região, ao mesmo tempo em que as conexões são realizadas sem deixar 'ilhas'. Pela quantidade de pontos, a carga computacional do cálculo da matriz de pesos não é sobremaneira onerosa em tempo e consumo de máquina. Além disso, podemos concluir que considerar um grande número de vizinhos próximos pode enviesar os resultados para predições mais uniformes devido ao potencial efeito de diluição que possa ocorrer entre os muitos pontos, reduzindo a sensibilidade local.

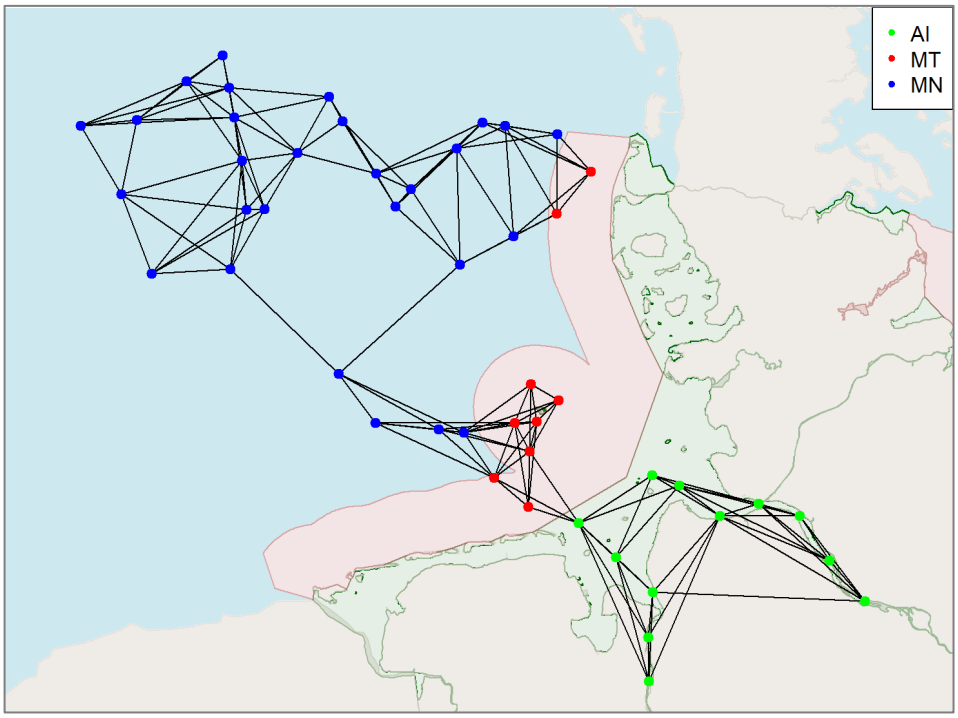


Figura 19: Arranjo escolhido para cálculo da Matriz W (KNN)

Para entendermos melhor o arranjo acima, podemos visualizar a distribuição distâncias no seguinte gráfico *boxplots:*

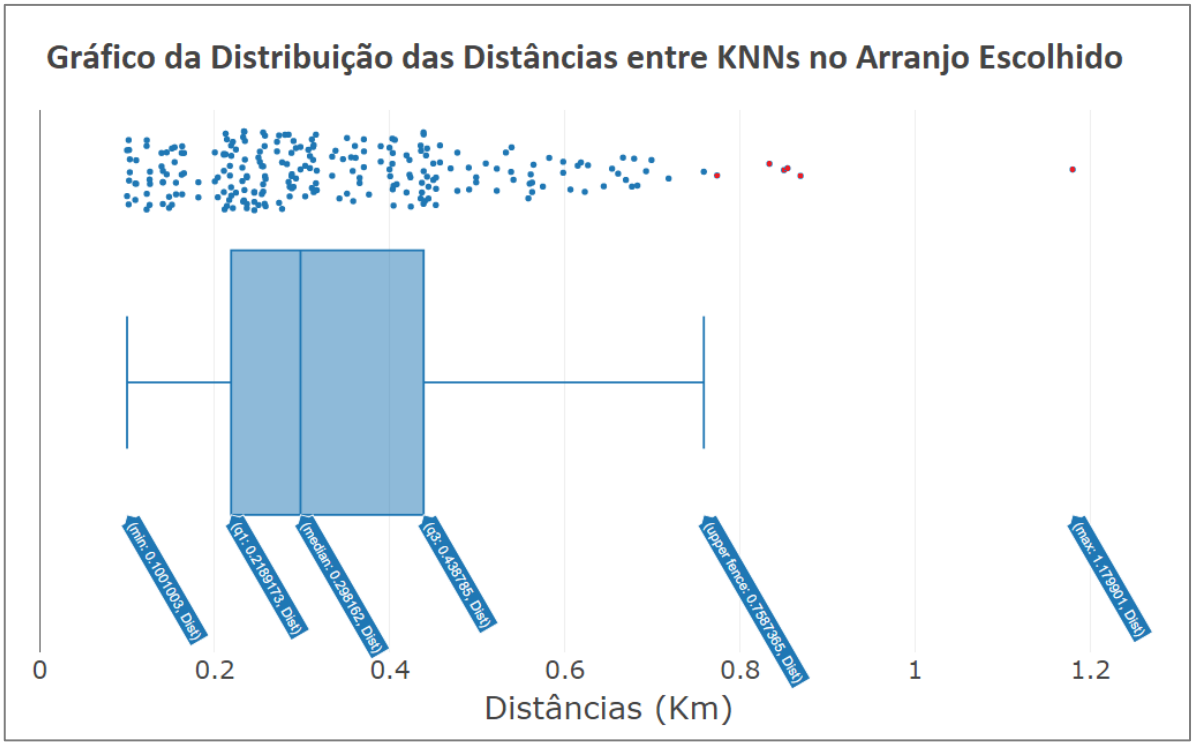


Figura 20: Boxplot da Distribuição das Distâncias dos Pontos no Arranjo Escolhido

Vemos no gráfico *boxplot* da Figura 20 que temos uma distribuição aproximada para pontos até 500 km. Alguns vizinhos influenciarão com menor intensidade o cálculo do índice de correlação espacial ao terem distâncias maiores que 600 m. Vemos ainda um pequeno agrupamento de pontos com distâncias próximas a 100 min, sendo estes os vizinhos próximos mais influentes para o cálculo da autocorrelação espacial segundo o inverso da distância a ser calculado para matriz W. Ainda, sobre as medidas de variação, a mediana dos pontos foi de 298.1 m, o IQR calculado foi de 219.9 m e a amplitude entre os percentis 10-90 foi de 450.7 m, sendo o desvio padrão das distâncias de 177.4 m. O vizinho mais distante considerado distou 1180 m.

Assim, para o cálculo da matriz de pesos considerando o arranjo acima, foi utilizada a distância inversa conforme Equação (3) com métrica do grande círculo para o cálculo de distância conforme Equação (7).

Devemos analisar o impacto da normalização da matriz de pesos de distâncias inversas e verificar se devemos considerar a normalização nas linhas da matriz, conforme seção 2.1 discute. Gráficos indicando os valores através de tonalidades de cores (*heatplots*) podem ajudar a visualizar as diferenças nas matrizes da distância inversa (W), calculada a partir da Eq. (3), e a mesma matriz da distância inversa, porém normalizada nas linhas, conforme Eq. (4).

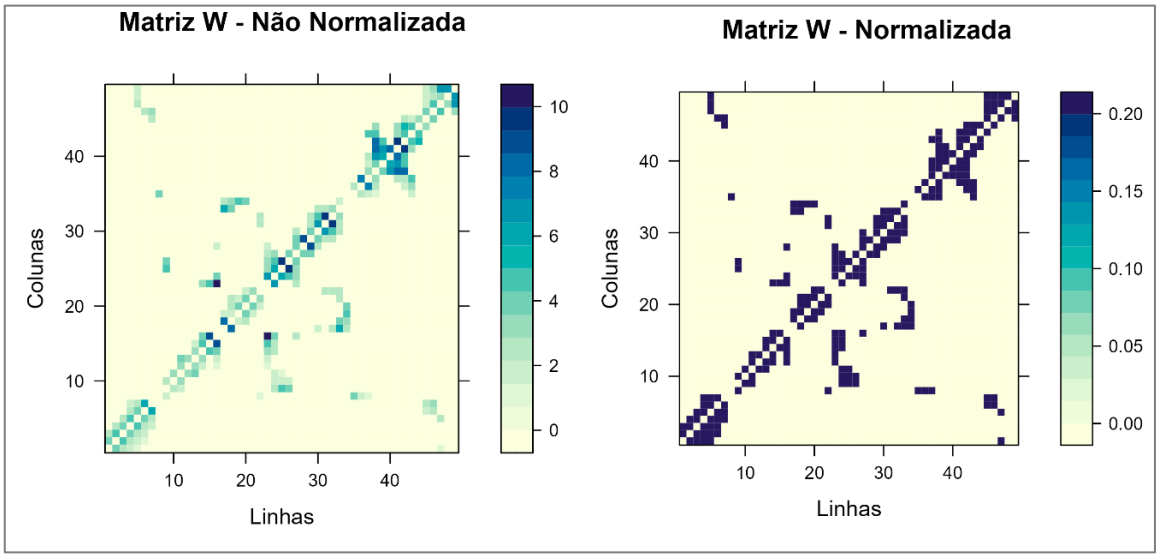


Figura 21: Heatmaps das matrizes *W* e *W* normalizada na linha

O que notamos ao comparar os *heatplots* da Figura 21 é que a matriz normalizada perde informação sensível às distâncias dos pontos. O que observamos são pontos muito intensos sendo representados por cores escuras com os pontos mais claros da matriz não normalizada sendo considerados como se fossem pontos próximos.

Portanto, por termos utilizado a distância inversa dos KNN para o cálculo da matriz de pesos, esta não foi normalizada, considerando o impacto da normalização, e preservando as ponderações dos inversos das distâncias, de acordo com o que foi discutido na seção 2.1.

7.2. O Cálculo da Estatísticas I de Moran e c de Geary Global

Tomando-se a matriz de pesos calculados a partir das distâncias inversas conforme seção anterior, faz-se o cálculo da Estatística I de Moran e c de Geary global para as diversas variáveis conforme Equações (8) e (13). Os resultados são apresentados na Tabela 8.

Tabela 8: Resultados para I de Moran e c de Geary por variável

	I de Moran	z-score	p-valor	p-valor MC	c de Geary	z-score	p-valor
Salinity	0.65463	7.2445	2.17E-13	0.001	0,10494	5,9322	1,49E-09
TempExtern	0.80615	8.5925	4.25E-18	0.001	0,06023	8,0086	5,80E-16
N.NO3	0.50486	5.7203	5.32E-09	0.001	0,25428	4,5658	2,49E-06
DOCeq	0.39864	4.9059	4.65E-07	0.001	0,27938	3,4127	3,22E-04
Abs210	0.36964	4.1697	1.52E-05	0.002	0,43431	3,8525	5,85E-05
Abs254	0.45485	5.3906	3.51E-08	0.001	0,27875	3,7459	8,99E-05
Abs360	0.42591	4.9868	3.07E-07	0.001	0,36146	3,4989	2,34E-04
SAC254	0.41866	5.0087	2.74E-07	0.001	0,27772	3,6833	1,15E-04
PC1	0.55069	6.3064	1.43E-10	0.001	0,18138	4,6997	1,30E-06
PC2	0.57698	6.4722	4.83E-11	0.001	0,49732	3,1593	7,91E-04

O valor esperado para o I de Moran global calculado foi de $E[I] = -0.0208$ conforme equação (9), indicando que valores de I superiores sugerem uma autocorrelação espacial positiva.

Na análise das variáveis, vemos que todos os valores de I representaram autocorrelação espacial positiva, com p-valores inferiores a *0.05* utilizando-se do z-score calculado de acordo com a equação (11). Na função utilizada na linguagem R, o teste foi feito para verificar a presença de autocorrelação espacial positiva, ou seja, com as hipóteses $H_0: I \leq E[I]\ e\ H_1: I > E[I]$. Sendo p-valor $< \alpha$ para todas as variáveis, logo rejeita-se a hipótese nula de autocorrelação espacial negativa ou ausência de autocorrelação, então concluímos que os dados fornecem evidências para autocorrelação espacial positiva.

A mesma conclusão é obtida se utilizarmos uma abordagem de Monte Carlo para avaliar a significância de I, conforme seção 2.2. Conduzimos uma abordagem de randomização de Monte Carlo definindo o número de simulações para $nsim = 999$. A Figura 22 mostra os histogramas dos valores do I de Moran para cada um dos padrões simulados, bem como o I de Moran obtido com os dados reais nas linhas vermelhas (da Tabela 8). Vemos os valores reais nas extremidades da distribuição. Observamos p-valores inferiores a 0.05 novamente, indicando que os dados apresentam autocorrelação espacial positiva, sendo esta significante, conforme teste de hipóteses similar ao anterior.

Os resultados da estatística c de Geary são similares ao I de Moran, também indicando correlação espacial positiva para todas as variáveis.

Com a significância dos valores calculados para o I de Moran verificada, podemos agora dar andamento na AEDE e analisar os valores de I calculados. A variável com o maior valor foi a *TempExtern* que, de acordo com *boxplot* das distribuições das variáveis por zona (Figura 10), indicava um aumento gradual na direção da zona MN. O segundo maior valor ficou com *Salinity* que, conforme o mesmo *boxplot*, teve um aumento na mesma direção. Porém, se observarmos com cautela, vemos que a variação desta variável é muito para as zonas MT e MN, se mantendo razoavelmente constante naquelas duas zonas. Assim sendo, podemos agora constatar que a variação destas duas variáveis é constante e crescente na direção da zona MN. O reduzido número de *outliers* observado

para estas duas variáveis também é um potencial fator que contribui para uma autocorrelação espacial mais proeminente.

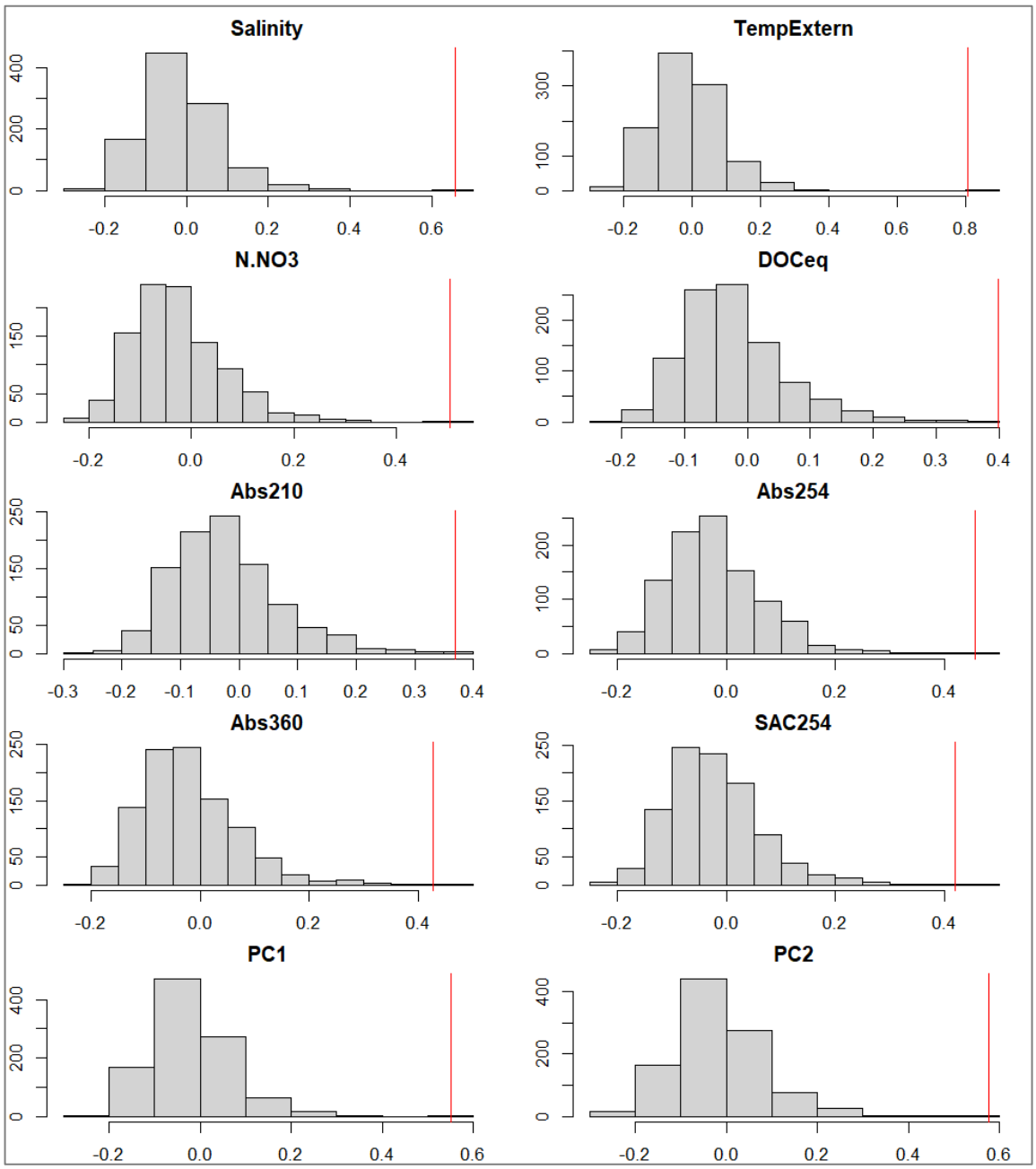


Figura 22: Valores de I para cada um dos padrões simulados por variável

Para as outras 6 variáveis do conjunto de dados original, ou seja, desconsiderando *PC1* e *PC2*, observamos valores mais moderados para o I de Moran. Da

AEDE inicial (seção 5.1), constatamos que elas se mostram fortemente compactadas em um curto alcance nas zonas MT e MN, assim sendo, possuindo naquelas zonas valores próximos e com variância muito pequena, reduzindo a variância da variável (ver Tabela 3: Sumário Estatístico dos Dados Globais). A variância é parte do denominador no cálculo do índice de Moran Global conforme Equação (8). Assim, pequenas variações espaciais nos dados podem resultar em um Índice de Moran mais sensível. Isso significa que mesmo leve autocorrelações espaciais podem ser detectadas como significativas. Outro ponto importante a se considerar foi a quantidade considerável de *outliers* gerado pela compactação dos dados naquelas zonas. Devemos recordar que os dados foram concentrados a uma separação de 60 min, tendo sido considerada a mediana dos dados naquela hora. Assim sendo, *outliers* não contribuíram com suas informações para o índice I de Moran Global.

Na consideração do *PC1* e *PC2,* podemos observar um valor moderado de I de Moran Global para ambas. Como essas variáveis representam a variabilidade do conjunto de dados como um todo, é interessante verificarmos que seu valores ficaram intermediários aos valores das outras variáveis. Podemos interpretar aqui que o conjunto de dados, como um todo, apresenta uma correlação espacial significativa, em especial pela observação dos resultados do PC1 pelo seu alto carregamento da variabilidade dos dados originais. Assim, consideramos que o uso das PCs nessa análise auxilia na interpretação do resultado conjunto da autocorrelação espacial das variáveis estudadas, indicando que elas, quando consideradas juntas, não somente são fortemente correlacionadas (vide Figura 12: Correlogramas por zona e global), mas apresentam autocorrelação espacial positiva.

Podemos ter uma visão mais abrangente da estatística I fazendo uso do Diagrama de Dispersão de Moran:

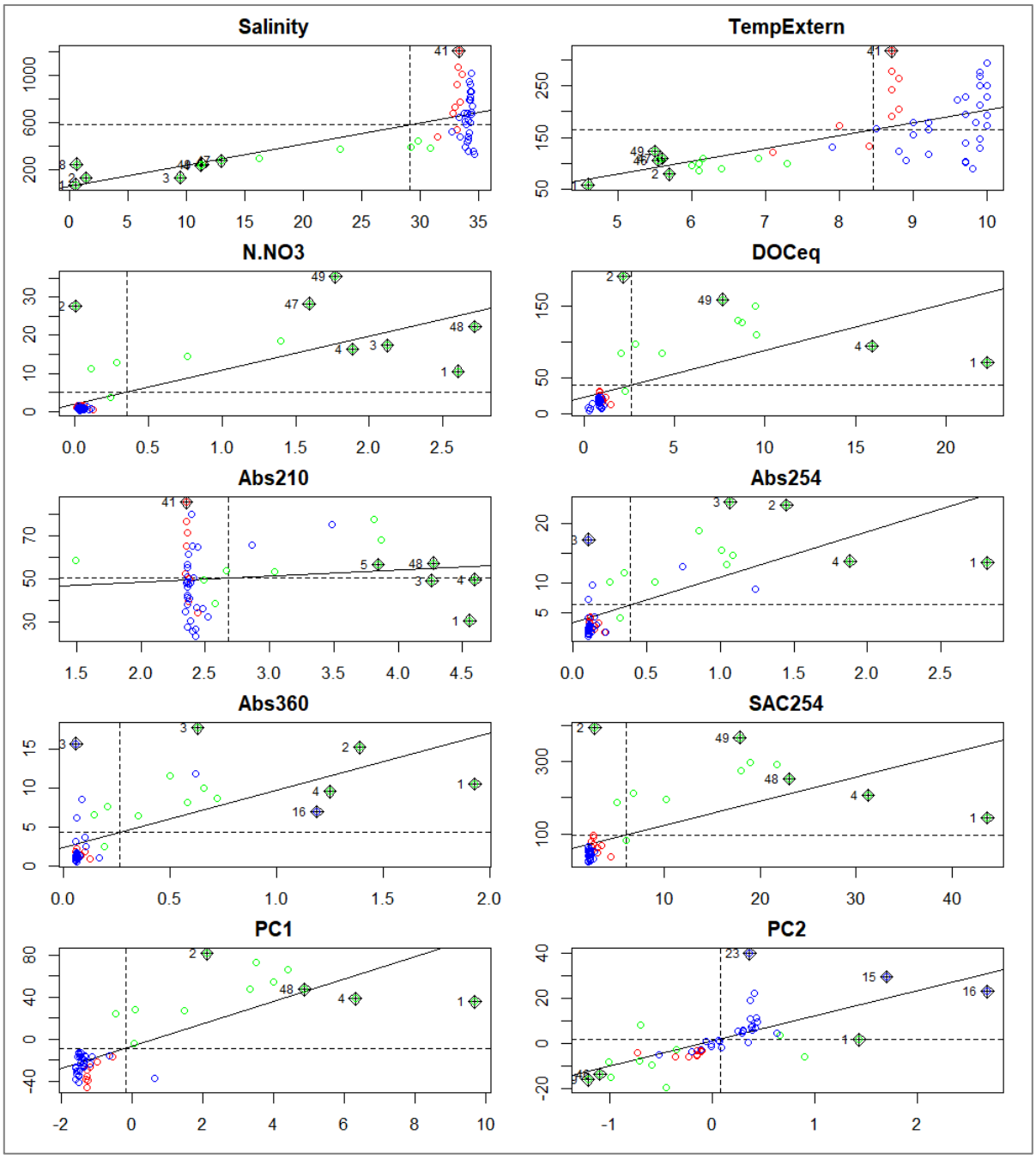


Figura 23: Diagramas de Dispersão de Moran por Variável

Na Figura 25 vemos as variáveis *N.NO3, DOCeq, Abs254, Abs360 e SAC254* com um padrão muito semelhantes em seus gráficos. Constatamos que as zonas MT e MN povoam o quadrante LL por apresentar os valores (eixo horizontal) e defasagem espacial baixos (abaixo de suas médias) e próximos, resultado da compactação de dados comentada anteriormente. Esse foi o padrão carregado no PC1, como podemos observar. A existência

de *outliers* espaciais é presente em todos eles por apresentarem pontos nos quadrantes LH e HL, indicando pontos circundados por vizinhos com valores que distam do valor deles. Há ainda alguns pontos de alavancagem identificados, em especial os pontos na zona da AI. Isto ocorre, como vimos, pois os valores naquela zona variam com maior intensidade. Pontos mais distantes das zonas salgadas foram os pontos de alavancagem comuns entre os gráficos, como os pontos 1, 2, 3 e 4, além do ponto 48.

A variável *Abs210* apresentou um padrão único, tendo o menor I de Moran do grupo. Um número significativo de outliers localizados no quadrante LH (pontos baixos circundados por pontos de valor alto) são representados, sendo estes majoritariamente das zonas MT e MN. Há ainda 3 pontos na zona contrária (HL), estes pertencentes a zona AI. O p-valor desta variável ficou abaixo de 0.05, porém apresentou o maior valor dentre todas as variáveis, indicando uma maior força no sentido da aceitação da H_0 de autocorrelação espacial. Esta oscilação que percebemos no gráfico de dispersão de Moran pode ser uma constatação de erro provocado por calibração do aparelho ou pela turbidez da água naquelas zonas, pois esses fatores podem influenciar essa variável (TriOS GmbH, 2017).

Salinity e *TempExtern* apresentam-se como tendo seus valores máximos na zona MN. Por ter variação mais ampla nos seus valores, os pontos foram mais espalhados ao longo do eixo horizontal, tendo *TempExtern* indicado uma distribuição que se ajusta melhor a reta. Vemos o ponto 41 da zona MT como sendo um ponto de alavancagem identificado para as duas zonas, principalmente pelo seu valor de defasagem espacial. Alguns *outliers* espaciais são encontrados ainda no quarante HL para ambas as variáveis.

O *PC2*, variável da PCA carregada principalmente pela *TempExtern* e *Abs360* apresentou uma configuração diferente das demais variáveis. O que podemos interpretar é que a combinação majoritária daquelas duas variáveis resultou em poucos *outliers* espaciais e alguns pontos de alavancagem relevantes, como o ponto 23. É preciso lembrar aqui que o PC1 representa a variabilidade destas duas variáveis com maior intensidade. Assim, apenas podemos considerar o resultado do PC2 como uma referência para o restante da variabilidade não explicada no PC1.

Podemos concluir então que, devido à baixa variabilidade dos dados nas zonas MT e MN, sendo a grande parte das variáveis com valores restringidos a um pequeno alcance, a sensibilidade das análises de I de Moran para a maioria das variáveis parece ter sido afetada. Mesmo assim, o cálculo de I mostrou que temos autocorrelação espacial positiva e, para todas as variáveis, verificamos que os valores foram significantes. De fato, por apresentarem pontos com valores próximos nas regiões, certa autocorrelação espacial era esperada *a priori*.

7.3. O I de Moran Local (LISA)

Procederemos com a análise dos valores de I de Moran locais. Recordamos da seção 2.3 que o objetivo da LISA é indicar *clusters* espaciais, significativos estatisticamente, para cada variável. Assim, calculamos os valores de I_i utilizando a Equação (14). Os números referentes aos 3 primeiros e 3 últimos pontos para a variável *Salinity* são exibidos na Tabela 9 como exemplo dos cálculos realizados. Os resultados para os valores do I de Moran Local e para os p-valores dos pontos da PC1 são exibidos no mapa da Figura 24. Os demais mapas para PC1 e as outras variáveis estão disponibilizados no Apêndice IIIa: Mapas do Cálculo do $\boldsymbol{I}$ de Moran Local (LISA).

Tabela 9: Alguns resultados dos cálculos da LISA para variável *Salinity*

	I_i	$E[I_i]$	$Var[I_i]$	$Z[I_i]$	$Pr(z > E(Ii))$
1	64.5589	-1.7705	1.85E+02	4.88240	5.24017E-07
2	78.6476	-2.3023	3.06E+02	4.63025	1.8261E-06
3	60.2940	-1.2364	1.79E+02	4.59838	2.12899E-06
...	...	...	...	...	...
47	45.7320	-1.0661	2.08E+02	3.24109	0.000595361
48	63.5253	-2.7984	4.86E+02	3.00975	0.001307335
49	65.1301	-1.4144	3.31E+02	3.65843	0.000126885

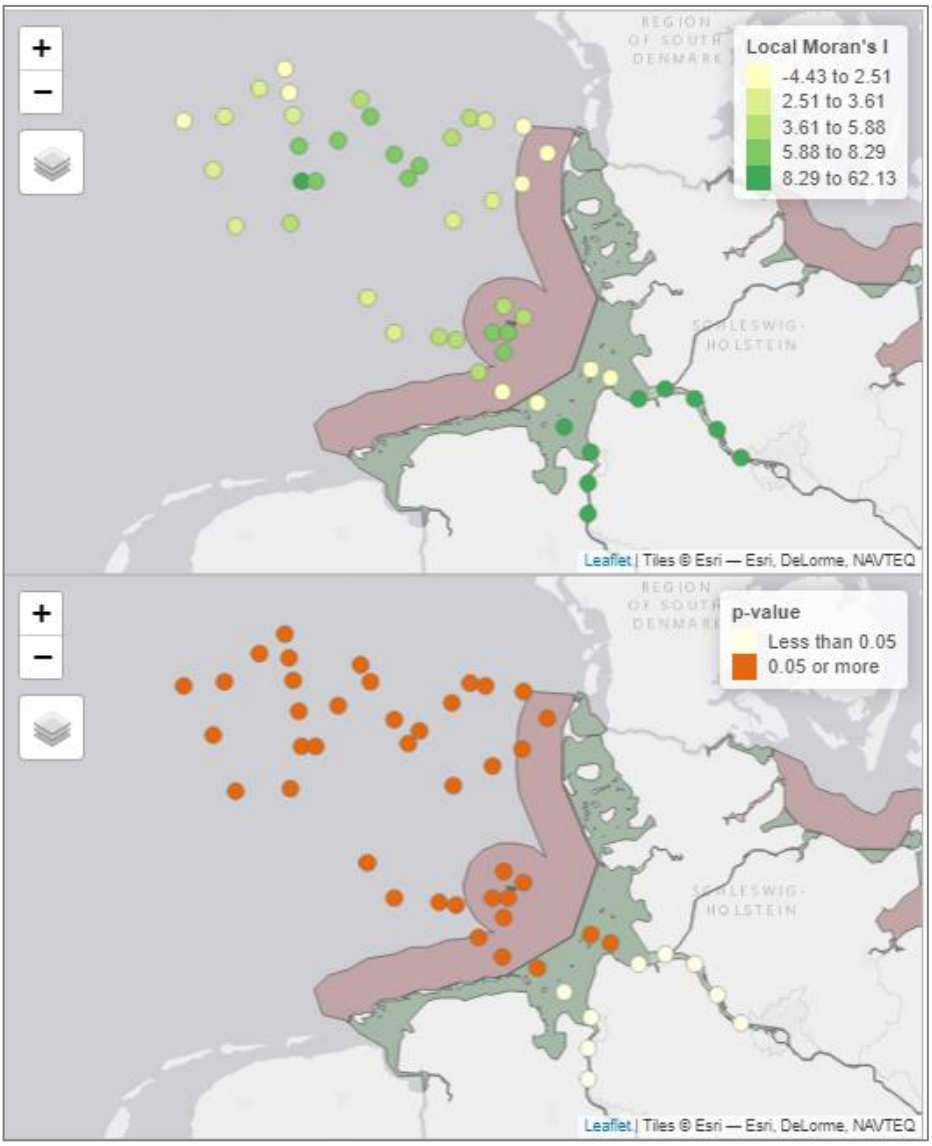


Figura 24: Mapas com valores para I de Moran Local (LISA) e p-valores

Observamos dos mapas plotados que os pontos nas zonas MT e MN falham em apresentar autocorrelação espacial positiva significativa para o teste de hipóteses, com seus p-valores superior a 0.05 para maioria das variáveis. Isto pode estar relacionado com a baixa variabilidade das variáveis naquela região. As observações feitas na zona AI apresentam pontos com p-valor menor que 0.05. Logo, podemos concluir que, para significância escolhida, rejeitamos a hipótese nula, havendo autocorrelação positiva significativa naquela área para as variáveis estudadas. Além disso, observações de *TempExtern* apresentaram p-valores inferiores a 0.05 na zona MN, apresentando

autocorrelação espacial positiva significativa naquela zona, diferenciando-se das outras variáveis.

A presença de autocorrelação espacial positiva não nos permite avaliar se há autocorrelação espacial negativa ou se não há autocorrelação espacial. Para investigar o que ocorre naqueles pontos, podemos plotar os mesmos mapas e ajustar o teste de hipóteses para H_0: autocorrelação espacial inexistente e H_1: autocorrelação espacial positiva ou negativa. Sendo este teste de 2 caudas (*two tailed test)*, o intervalo de valores para os z-scores mudam para $]-\infty, -1.96]\ U\ ]1.96, +\infty[$ na Equação (11), pois sua distribuição é normal e queremos manter o valor de α. Os resultados para todas as variáveis são apresentados no Apêndice IIIb. O gráfico para o PC1 é apresentado na Figura 25.

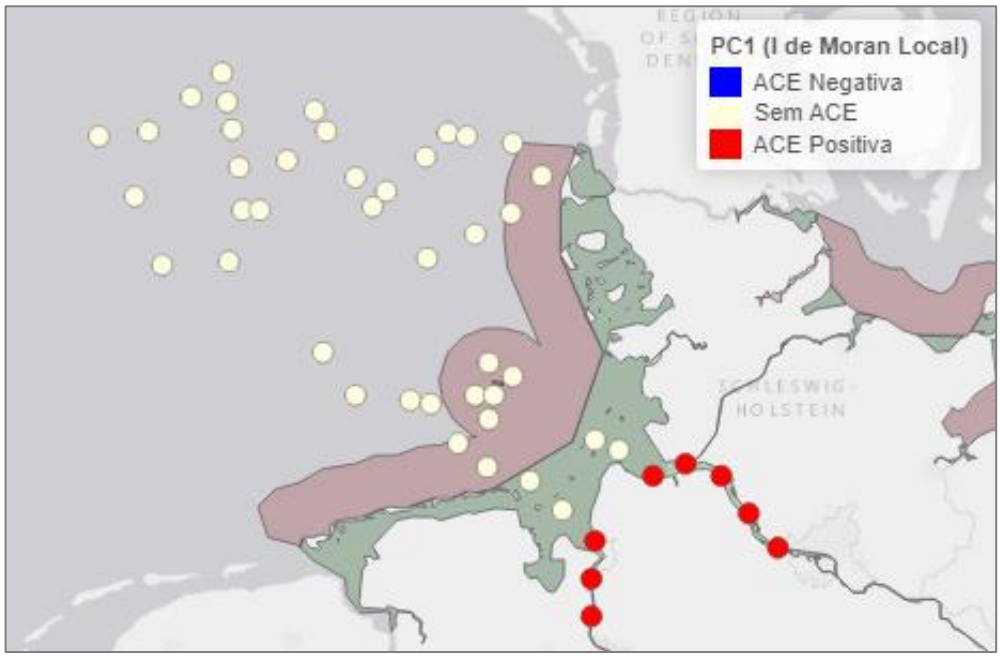


Figura 25: Resultados do I de Moran Local para autocorrelações Espaciais (ACE) significativas no PC1

Na análise desses resultados, confirmamos que há ausência de autocorrelação espacial significativa nas zonas MT e MN para maioria das variáveis. Também confirmamos que não houve autocorrelação espacial significativa entre os pontos observados sem autocorrelação espacial positiva em teste anterior, com exceção de alguns *outliers* encontrados em algumas variáveis, em espacial o segundo ponto da direita para esquerda

no mapa (observação 2). Isso indica que a pouca variabilidade nas zonas não permitiu uma análise sensível o suficiente para detectar algum tipo de autocorrelação espacial significativa entre aqueles pontos na maioria das variáveis. Também concluímos que as observações da zona AI apresentaram autocorrelação espacial positiva para maioria das variáveis. Este perfil foi refletido no PC1, componente que explicou a maior parte da variabilidade dos dados originais no teste de PCA (Figura 25).

Por fim podemos analisar *clusters* dos pontos que não têm uma contribuição significativa para a autocorrelação das diferentes variáveis. Mapas foram plotados para cada uma das variáveis e se encontram no Apêndice IIIc. A diferença entre estes e os mapas anteriores é que apresentam não apenas o tipo de autocorrelação espacial para os pontos significativos (positiva ou negativa), mas também apresenta as relações dadas pelo quadrante do diagrama de dispersão de Moran Local para a variável. Assim, podemos classificar estes clusters em HH, BB, HL e LH, e verificar o seu tipo, de acordo com o explicado seção 2.2. O mapa de exemplo tomado aqui foi o da variável N.NO3.

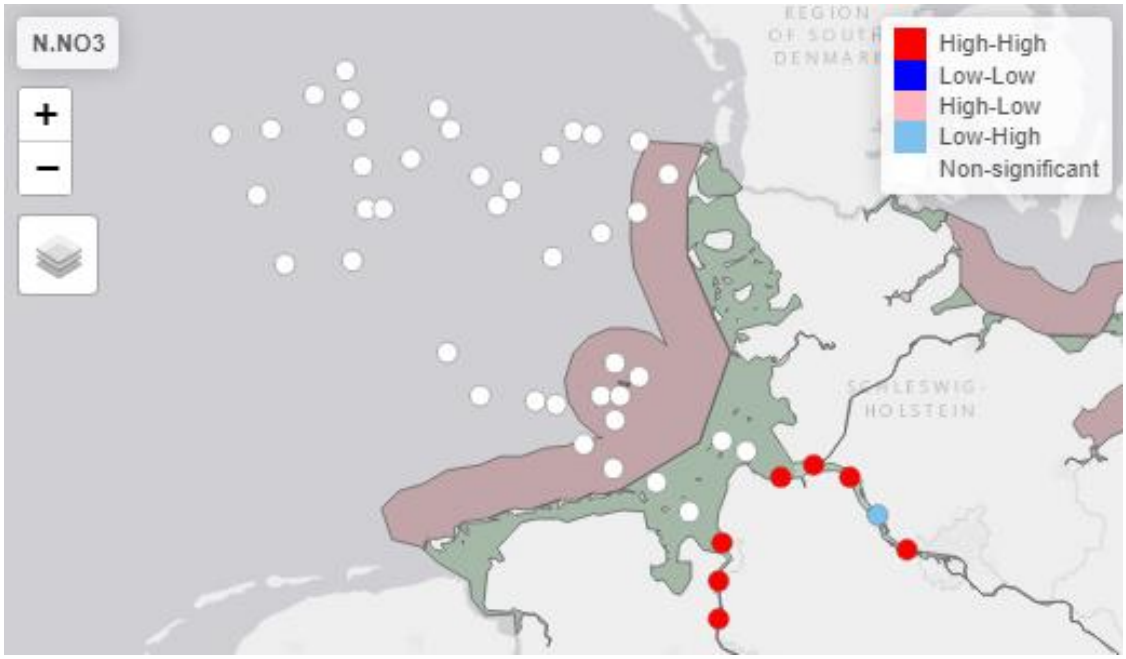


Figura 26: Tipos de Clusters Significativos da Variável N.NO3

Assim, com exceção das variáveis *Salinity*, *TempExtern* e *PC2* (carregada fortemente por *TempExtern*), foram encontrados *clusters* do tipo HH na zona AI. Como esperado, pois tais variáveis de exceção são inversamente correlacionadas com as outras.

Os *outliers* significativos encontrados são do tipo LH, indicando que possuem valor baixo e são circundados por pontos de valores altos. A variável *TempExtern* mostrou seu cluster LL em AI e HH em MN. Nenhum ponto significativo para a autocorrelação espacial foi encontrado na zona MT para variáveis estudadas.

CONCLUSÃO

Buscando investigar possível autocorrelação espacial na análise de alguns componentes em águas no Mar do Norte, esta dissertação focou em métodos analíticos geoestatísticos utilizando-se de técnicas multivariadas como ferramenta para simplificação dos dados em busca de *insights* significativos sobre os dados coletados na expedição ocorrida em dezembro de 2020 a bordo do RV Meteor.

Algumas questões de interesse se destacam. Uma delas é traçar o perfil espacial das variáveis estudadas através de seu mapeamento e estudo de sua autocorrelação espacial para entender como que estas poderiam ou não estar associadas a outros fenômenos naturais. Outra é verificar se os resultados obtidos nesse estudo podem acrescentar aos resultados obtidos nas análises biológicas, de metais ou de outros componentes nas águas realizados pela equipe da expedição. Ainda, caso sejam encontrados padrões nas autocorrelações espaciais, se podemos inferir quaisquer conclusões de suas origens e causas através dos resultados obtidos.

Após a realização do pré-tratamento de dados, a AEDE inicial por zona revelou o achatamento da maior parte das variáveis em um pequeno intervalo nas zonas MT e MN, apresentando um desvio padrão baixo, bem como IQR, e um número considerável de *outliers* devido ao achatamento. Além disso, as distribuições das variáveis padronizadas se mostraram semelhantes à exceção de *Salinity* e *TempExtern*, que se diferenciaram das outras variáveis, mas foram semelhantes entre si. A zona AI foi a zona que apresentou maior variância para todas as variáveis. Os resultados iniciais por zona indicavam haver uma possível correlação entre as diferentes variáveis devido a suas distribuições semelhantes. Também indicaram que uma possível autocorrelação espacial exista da zona AI na direção MN.

Ainda sobre a AEDE inicial, realizou-se o mesmo estudo, porém tomando as variáveis globalmente. Os desvios padrões com valores baixos e achatamento das variáveis no gráfico *boxplots* indicam que as propriedades das zonas MT e MN prevaleceram sobre o conjunto global.

Os correlogramas por zona apresentados indicam correlações semelhantes entre as zonas. Em todas as zonas, *Salinity* e *TempExtern* se mostraram correlacionadas

positivamente. As variáveis de absorção também apresentaram correlação significativa positiva em todas as zonas. As zonas AI e MT foram muito similares em relação à correlação de suas variáveis, enquanto MN apresentou diversas correlações não significativas o suficiente para serem consideradas. A baixa variabilidade dos dados naquela região produziu este resultado, pois esta característica reduz a força da relação linear entre as variáveis e esta não pode ser estimada corretamente. Além disso, tal característica também torna o cálculo do coeficiente de correlação mais sensíveis a *outliers*, sendo que pontos ligeiramente diferentes dos outros podem ter um grande impacto no valor final. Estes efeitos também podem conduzir a conclusões enganosas, pois se duas variáveis têm pouca variação, podem eventualmente produzir correlação significativa e que, na verdade, não representa a real relação entre elas. Tais considerações também devem ser feitas para as correlações calculadas na zona MT, pois também apresentou características semelhantes. É possível que essa zona tenha apresentado correlações lineares fortes e semelhantes à zona AI devido a sua proximidade espacial, porém apenas alguns pontos seriam mais significativos no cálculo da correlação de Pearson, aqueles que se afastam da média e se aproximam da distribuição da zona AI, o que poderia explicar a semelhança dos correlogramas da zona MT e AI.

No cálculo da matriz de correlações dos dados globais, temos um resultado final que se assemelha a região AI e MT. Uma análise inicial sem a consideração do fator espacial não pode ser definitiva, porém é preciso ter cautela na tomada desses valores como absolutos ou mesmo preponderantes para toda a região, pelos mesmos motivos considerados na análise das correlações por zona.

A PCA foi então realizada por zona e globalmente e percebemos o impacto das fortes correlações lineares calculadas nas zonas AI e MT e dos dados globais. Para estes casos, com apenas 2 PCs foi possível explicar 85% ou mais da variabilidade dos dados originais, tendo o PC1 considerável peso, explicando cerca de 75% do total. Conforme esperado dos correlogramas, a variabilidade representada pelos dois primeiros PCs da zona MN foi menor, com cerca de 60%. Os *loadings* significativos foram: zona AI, *TempExtern* no PC2; zona MT, *N.NO3* e *Abs210* no PC2; zona MN, as variáveis de absorção no PC1 e *Salinity* e *TempExtern* no PC2. A fraca correlação linear verificada para a zona MN resultou uma

qualidade baixa da representatividade de *N.NO3*, *DOCeq* e *SAC254*. Para os dados globais, os *loadings* no PC1 foram similares para todas as variáveis, porém o PC2 foi apresentou valores de *loadings* mais altos para *TempExtern* e *Abs360*.

Os resultados da PCA foram interpretados com o devido cuidado dada a baixa variabilidade dos dados originais em algumas zonas. O gráfico *biplot* dos scores das PCs apresentado (Figura 18) indica a compactação dos dados nas zonas MT e MN. Também representa a maior variabilidade dos dados na zona AI. Zonas próximas do centro, especialmente aquelas próximas do PC mais representativa (PC1) e *outliers* são exibidos com cos^2 baixo, indicando que aqueles pontos das PCs não têm uma contribuição significativa na representação das variáveis originais. Assim, os dois primeiros PCs foram tomadas como variáveis a serem consideradas na análise de autocorrelação espacial.

Na continuidade a AEDE, procedeu-se com a construção da matriz de pesos espaciais utilizando-se do método de k-vizinhos próximos. As distâncias inversas foram então calculadas e evitou-se a normalização nas linhas pela perda da sensibilidade espacial percebida ao comparar ambas matrizes.

O cálculo do I de Moran global para cada variável e para os dois PCs indicou autocorrelação espacial positiva significativa em todas as variáveis. Nos diagramas de dispersão de Moran foi possível perceber arranjos compactados de pontos devido à baixa variabilidade nas regiões MT e MN. Isto pode ter ocasionado possíveis valores de Moran que conduzem a conclusões errôneas, pois a variância é parte do denominador no cálculo do índice de Moran Global e assim pequenas variações espaciais nos dados podem resultar em um I de Moran mais sensível. Isso significa que mesmo leve autocorrelações espaciais poderiam ser detectadas como significativas, semelhantemente ao ocorrido com a correlação linear.

Os resultados do cálculo do I de Moran local para cada variável corroboraram com a suspeita anterior e indicaram que nas regiões de dados compactados os dados não apresentaram autocorrelação linear significativa. A exceção foi a variável com maior amplitude para as três zonas, *TempExtern* e, consequentemente, o PC2, por ser fortemente carregada por *TempExtern*, que chegaram a apresentar correlação positiva em alguns pontos da zona MN.

Assim, concluímos que as autocorrelações espaciais apontadas pelo I de Moran global estão localizadas majoritariamente na zona AI. *Clusters* LL foram encontrados para *Salinity* e *TempExtern* na região AI enquanto *clusters* HH foram encontrados na mesma zona para as demais variáveis originais. Os resultados das PCs representaram bem o conjunto de dados como um todo, tendo o PC1 apresentado *clusters* semelhantes às variáveis originais como um todo e o PC2 semelhante a *TempExtern.* O mesmo ponto foi determinado como *outlier* espacial LH para 4 variáveis, indicando que pode haver uma variação de interesse naquela zona, pois há uma baixa no valor de algumas das variáveis em relação aos vizinhos próximos.

Dentre as limitações deste estudo, podemos citar o método da coleta dos dados. Os dados foram coletados durante a viagem, ao longo percurso do cruzeiro. Assim, não temos observações simultâneas nos diferentes pontos considerados, o que potencialmente influenciou os resultados deste estudo devido a marés e/ou outras dinâmicas marítimas e meteorológicas locais. Coletas realizadas por estações fixas poderiam prover dados mais consistentes nesse sentido. Outra limitação diz respeito ao breve período da coleta dos dados, sendo este curto para aplicação de métodos de análise espaço-temporais mais robustas que seriam convenientes para compreender as mudanças das variáveis consideradas ao longo de um período mais significativo. Desequilíbrios sazonais na região, como a mencionada *Great Salinity Anomaly*, poderiam levar um tempo considerável até serem percebidos, talvez anos, se não décadas. Além disso, os elementos responsáveis por iniciar ou intensificar esses eventos podem acumular-se ao longo do tempo, até que finalmente desencadeiem os fenômenos observados.

Enquanto perspectivas futuras para pesquisadores que desejam investigar temas semelhantes e avançar em relação à análise da qualidade da água na região, orientamos o cruzamento dos resultados deste estudo com outros resultados científicos obtidos naquela área. Uma possível abordagem inicial poderia utilizar outras observações da equipe do RV Meteor. Um dos achados da expedição foi que, em áreas de alta salinidade, há uma maior presença e diversidade de bactérias que resistem ao zinco (TRAM, 2024). Os resultados para a variável Salinity apresentados neste trabalho poderiam ajudar a mapear os locais onde tais bactérias se proliferam com mais intensidade e fazer um

estudo espacial comparativo entre a salinidade e a quantidade de bactérias resistentes a zinco. Também comentam sobre o potencial papel que a luz UV tem na seleção desse tipo de bactéria, assim outra proposta seria um estudo mais aprofundado sobre a relação da quantidade partículas suspensas na água e a quantidade de bactérias resistentes a zinco. Ainda, a região em que o *outlier* espacial foi encontrado pode ser investigada mais detalhadamente para entendermos quais são as dinâmicas locais que fazem aquela região se diferenciar das regiões que a circundam nas variáveis estudas. Possíveis estudos futuros também poderiam incluir técnicas de modelagem de dependência espacial, como o Modelo Espacial Autoregressivo (SAR) e o Modelo de Erro Espacial (SEM), úteis para predições e entendimento do impacto de variáveis explicativas no espaço.

BIBLIOGRAFIA

A. Koschinsky, K. S. (2021). *Tracing origin and distribution of geogenic and anthropogenic dissolved and particulate critical high-technology metals in the southern North Sea.* Emden, Germany: Jacobs University Bremen.

Abdi, H., & Williams, L. (7 de 2010). Principal Component Analysis. *WIREs Computational Statistics,2,4*, pp. 433-459.

Akimova, A., Núñez-Riboni, I., Kempf, A., & Taylor, M. (1 de 10 de 2016). Spatially-Resolved Influence of Temperature and Salinity on Stock and Recruitment Variability of Commercially Important Fishes in the North Sea. *PLoS ONE 11(9)*.

Almeida, E. (2012). *Econometria Espacial Aplicada.* Campinas, SP: Editora Alínea.

Anselin, L. (1988). *Spatial Econometrics : Methods and Models.* Kluwer Academic.

Anselin, L. (1995). Local Indicators of Spatial Association - LISA. *Geographical Analysis 27*, pp. 93–115.

Argo. (2009). *Argo Quality Control Management (V2.4).* Argo Data Management.

Baird, R., Eaton, A., Rice, E., Bridgewater, L., & Federation, W. E. (2017). *Standard Methods for the Examination of Water and Wastewater.* Washington, Dc: American Public Health Association.

Belkin, I., Levitus, S., Antonov, J., & Malmberg, S.-A. (Jan de 1998). "Great Salinity Anomalies" in the North Atlantic. *Progress in Oceanography, 41*, pp. pp. 1-66 .

Cliff, A., & Ord, J. (1981). *Spatial Processes.* London: Lion Limited.

Dronen, S. O., & Gjengedal, K. (11 de 03 de 2020). *Legal Puzzle in the North Sea*. Fonte: University of Bergen - Marine Research: https://www.uib.no/en/news/89171/legal-puzzle-north-sea

Elhorst, J. (2014). *Spatial Econometrics: from Cross-Sectional Data to Spatial Panels.* New York: Springer.

Fávero, L., & Belfiore, P. (2017). *Manual de Análise de Dados.* Rio de Janeiro: Elsevier Brasil.

Fu, Z., Wu, F., Zhang, Z., Hu, L., Zhang, F., Hu, B., . . . Liu, R. (26 de 02 de 2021). Sea Surface Salinity Estimation and Spatial-Temporal Heterogeneity Analysis in the Gulf of Mexico. *Remote Sens, 2021,13, 881*.

Getis, A. (7 de 2008). A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective. *Geographical Analysis,40,3*, pp. 297-309.

Goodchild, M. (1986). *Spatial Autocorrelation.* Norwich: Geo Books.

Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal 69*, 211-221.

Hair, J., Black, W., Babin, B., & Anderson, R. (2019). *Multivariate Data Analysis (8th ed.).* Cengage Learning.

Huthnance J, W. R. (2016). Recent change—North Sea. Em C. F. Quante M, *North Sea Region Climate Change Assessment* (pp. pp. 85-136). Cham: Springer International Publishing.

IOC/IODE. (1993). *IOC Manuals and Guides No.26: Manual of Quality Control Procedures for Validation of Oceanographic Data .*

Isard, W. (1969). *General Theory.* Cambridge: MIT Press.

Johnson, R., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis.* Upper Saddle River, N.J.: Pearson Prentice Hall.

Jolliffe, I. (2002). *Principal Component Analysis (2nd ed.).* New York: Springer-Verlag.

Kim, W., Yeager, S., & Danabasoglu, G. (11 de 2022). Revisiting the Causal Connection between the Great Salinity Anomaly of the 1970s and the Shutdown of Labrador Sea Deep Convection. *Journal of Sea Research, 189*, p. p. 102281.

Knauss, J. (2005). *Introduction to Physical Oceanography.* Long Grove: Waveland Press.

Kopczewska, K. (2021). *Applied Spatial Statistics Data Analysis in R.* New York, NY: Routledge.

Li, H., Calder, C., & Cressie, N. (2007). Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model. *Geographical Analysis*, 357-375.

Longley, P., & Et. Al. (2005). *Geographical Informtion Systems and Science.* Chichester: John Wiley & Sons, Cop.

Maritime Borders of the Federal Republic of Germany. (2019). Fonte: https://gdk.gdi-de.org/geonetwork/srv/api/records/631b3032-8a57-4354-ba6b-7571ac38fcd7

Miller, H. (2004). Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers. 94 (2)*, 284-289.

Moraga, P. (2023). *Spatial Statistics for Data Science: Theory and Practice with R.* Chapman & Hall/CRC Data Science Series.

Potter, B., & Wimsatt, J. (2005). *Method 415.3 - Measurement Of Total Organic Carbon, Dissolved Organic Carbon And Specific UV Absorbance At 254 Nm In Source Water And Drinking Water.* Washington, DC: U.S. Environmental Protection Agency.

Rigby, D., Barber, G., & Burt, J. (2009). *Elementary Statistics for Geographers.* New York; London: Guilford Press.

Schlundt, M. (14 de 09 de 2022). *Continuous thermosalinograph oceanography along RV Meteor cruise M169 - Data Processing Report*. Acesso em 22 de 03 de 2023, disponível em PANGEAEA: https://download.pangaea.de/reference/110986/attachments/DAM_DataProcessingReport_M169.pdf

SeaDataNet. (2010). *Data Quality Control Procedures (2.0).*

SeaDataNet. (07 de 06 de 2019). *Data Quality Control*. Acesso em 22 de 03 de 2023, disponível em Pan-European Infrastructure For Ocean & Marine Data Management: https://www.seadatanet.org/Standards/Data-Quality-Control

Sicilia, G., Rivera, M., & Navarro, J. (2017). Métodos gráficos de análisis exploratorio de datos espaciales con variables espacialmente distribuidas. *Cuadernos Latinoamericanos de Administración, vol. XIII, núm. 25*, 92-104.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Vol. 46, Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods*, 234-240.

TRAM. (06 de 02 de 2024). Fonte: Prof. Dr. Andrea Koschinsky: https://andrea-koschinsky.org/research-projects/tram/

TriOS GmbH. (2017). *OPUS Operating Instructions.* Rastede, Germany.

Tyszler, M. (2006). Econometria Espacial: Discutindo Medidas para a Matriz de Ponderação Espacial.

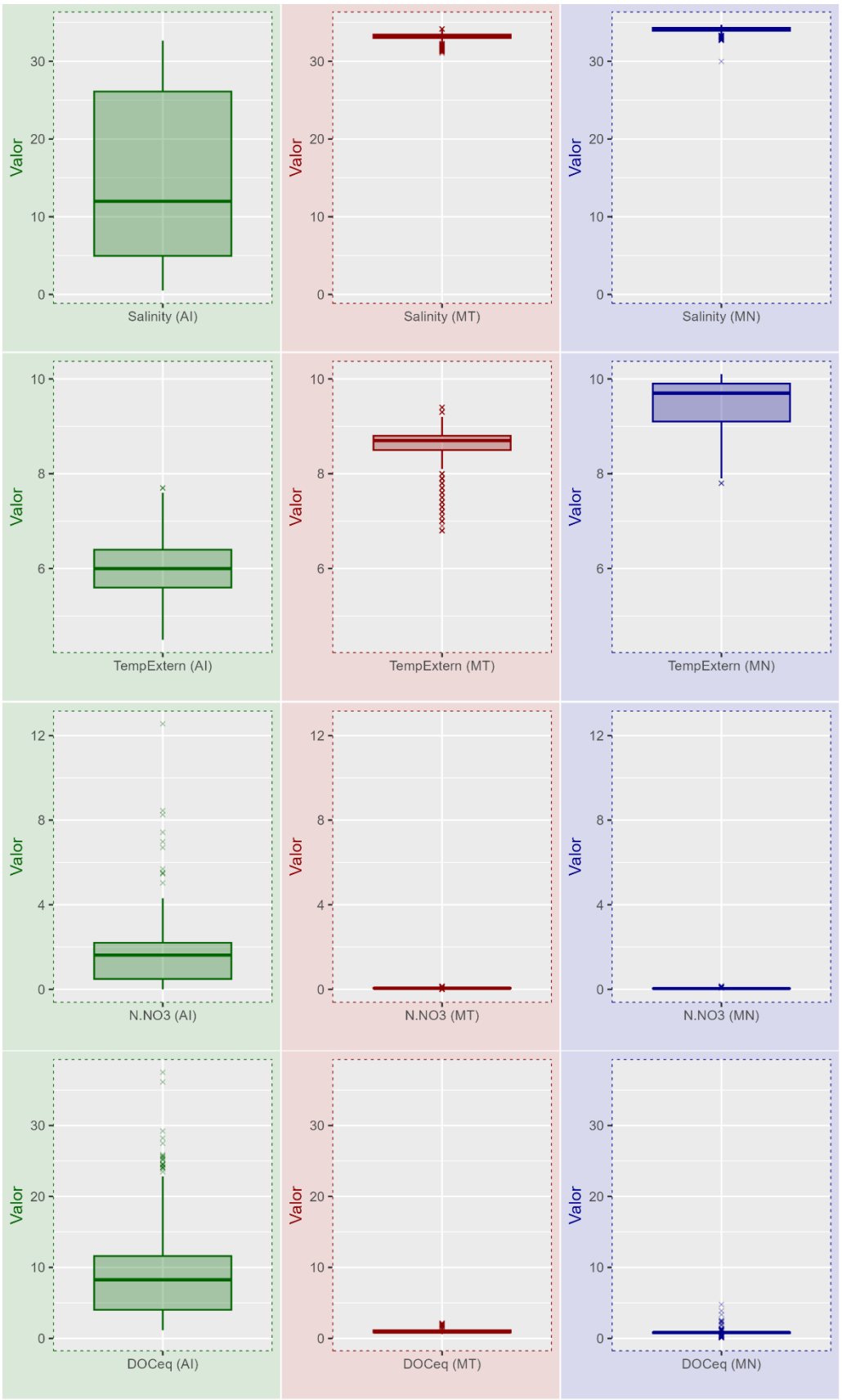
Walday, M., & Kroglund, T. (2008). The North Sea: Bottom Trawling and Oil/Gas Exploitation. *European Environment Agency*.

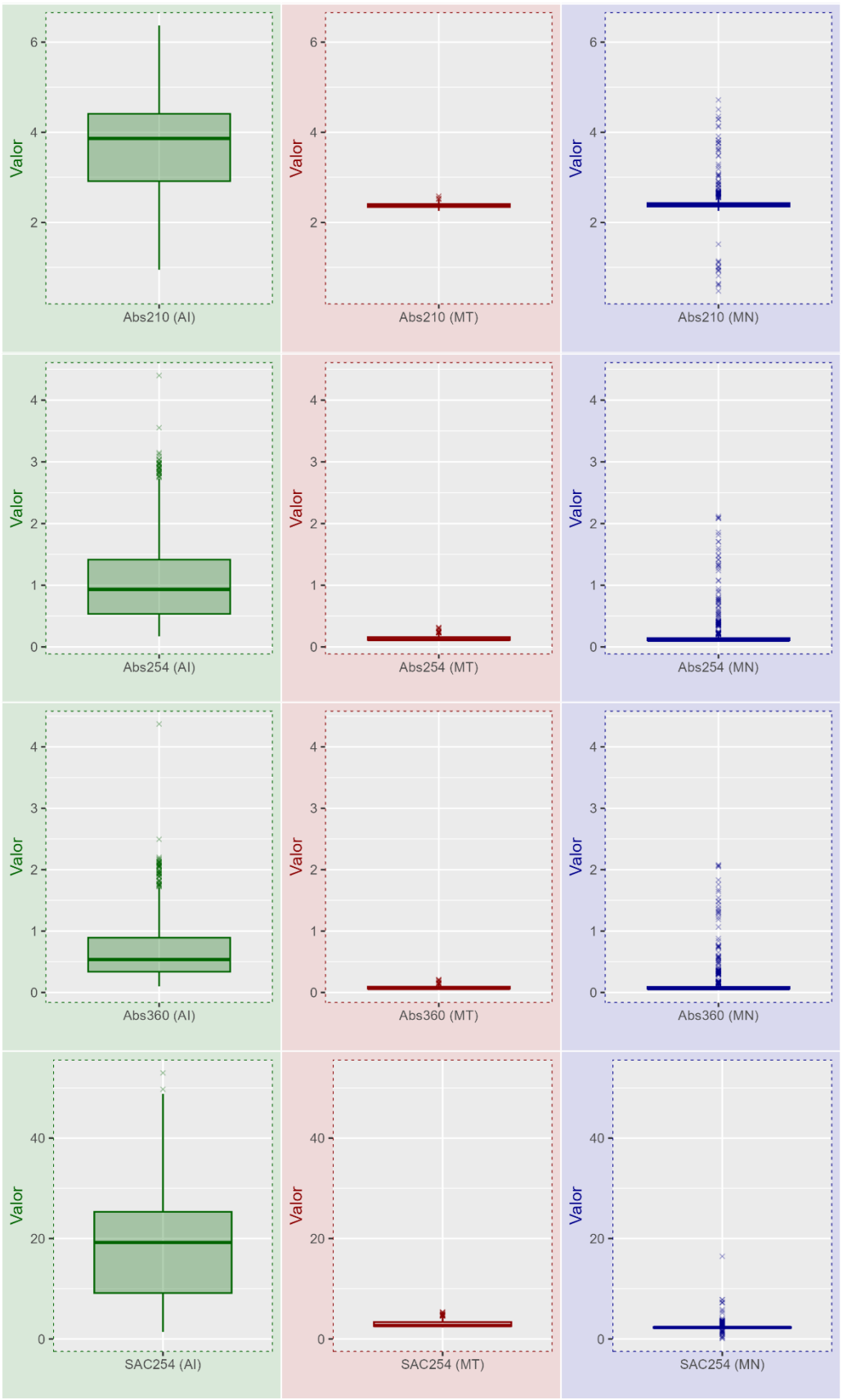
World Heath Organization. (2016). *Nitrate and Nitrite in Drinking-water Background Document for Development of WHO Guidelines for Drinking-water Quality.*

Zar, J. (2019). *Biostatistical Analysis.* Upper Saddle River, New Jersey: Pearson Education.

APÊNDICES

Apêndice I Distribuição das Variáveis por Zona




Valor
Abs210 (AI)
Abs210 (MT)
Abs210 (MN)
Abs254 (AI)
Abs254 (MT)
Abs254 (MN)
Abs360 (AI)
Abs360 (MT)
Abs360 (MN)
SAC254 (AI)
SAC254 (MT)
SAC254 (MN)

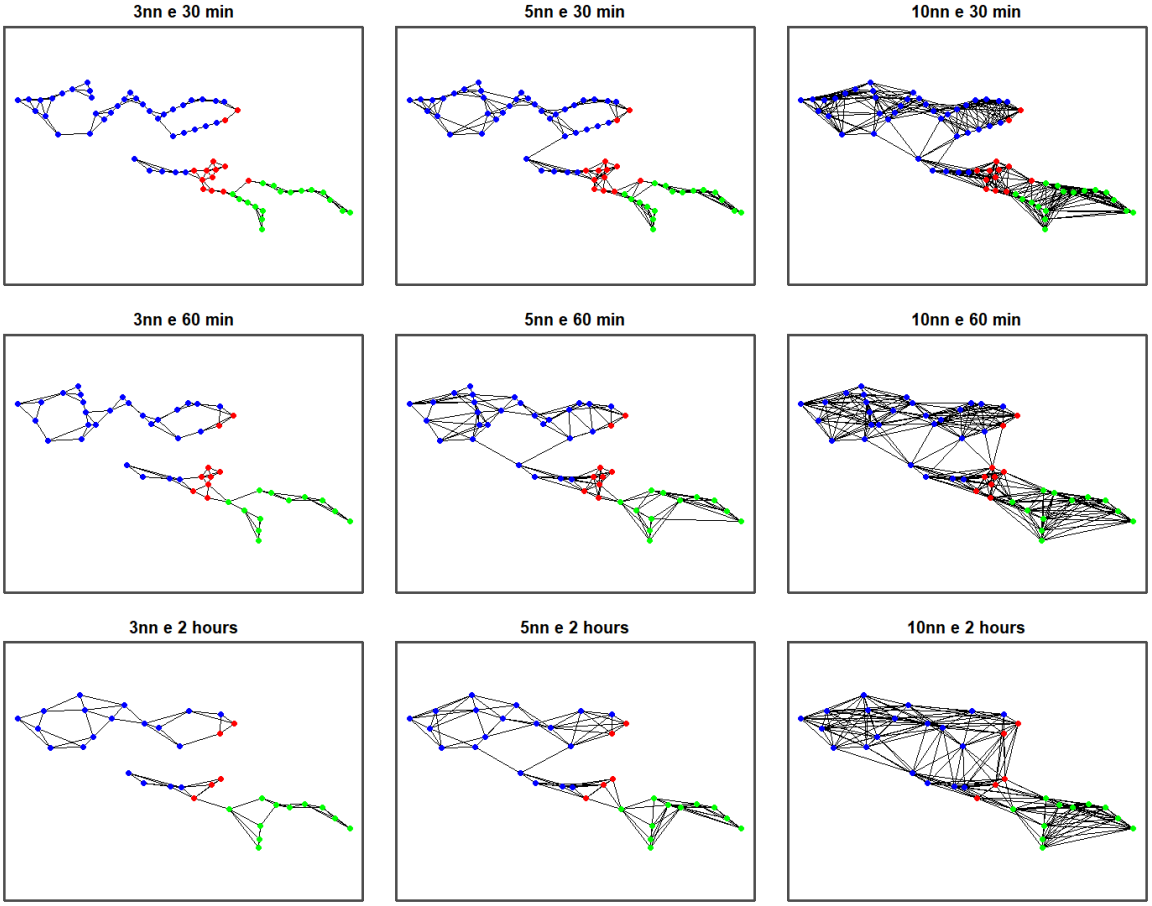
Apêndice II Diferentes Arranjos para a Matriz de Pesos

Diferentes arranjos foram investigados para a matriz de KNN:

- 3, 5 e 10 vizinhos
- Pontos temporalmente espaçados em 30, 60 e 120min

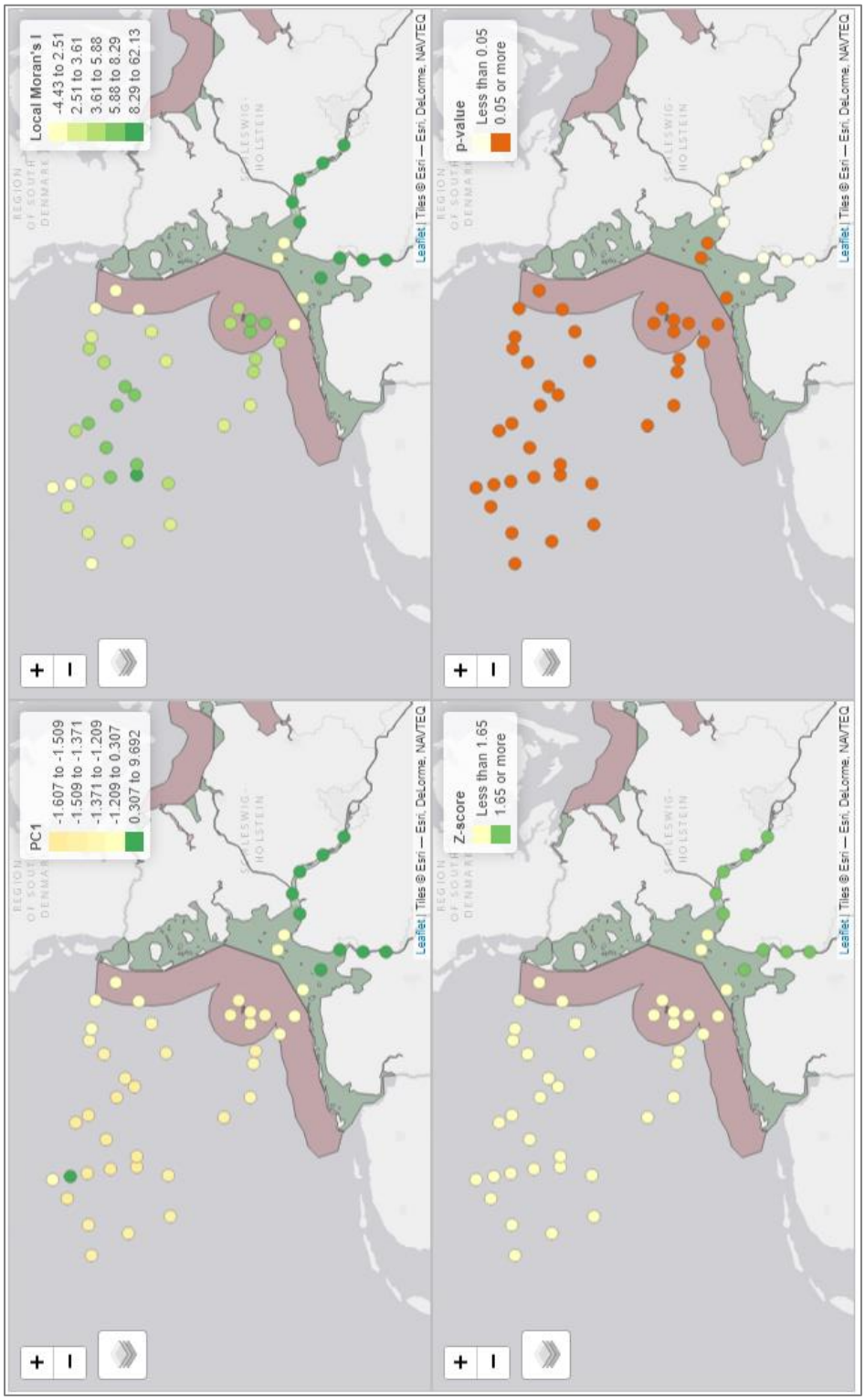
Os pontos nas regiões estão representados da seguinte maneira:

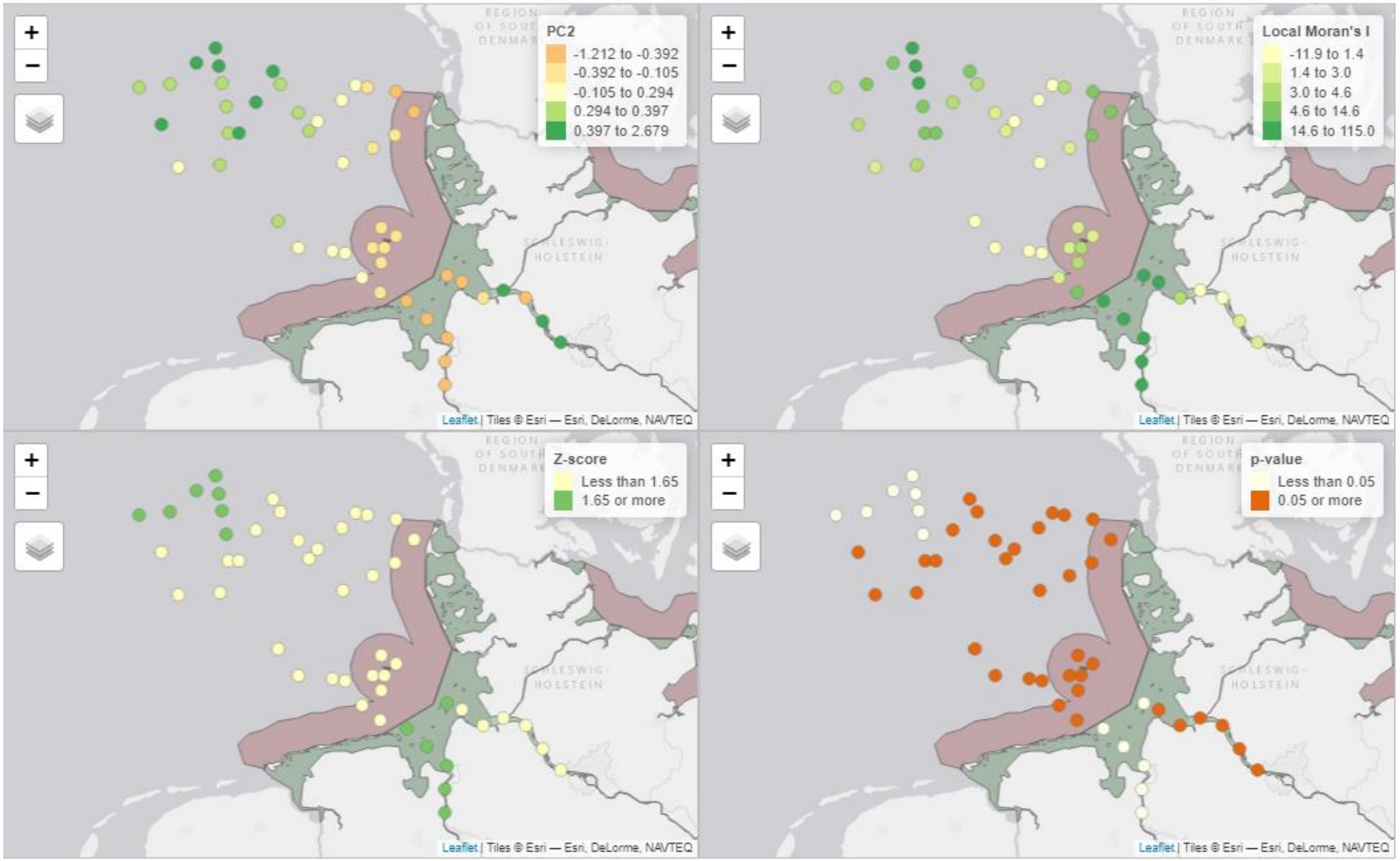
- Zona AI: pontos em verde
- Zona MT: pontos em vermelho
- Zona MN: pontos em azul



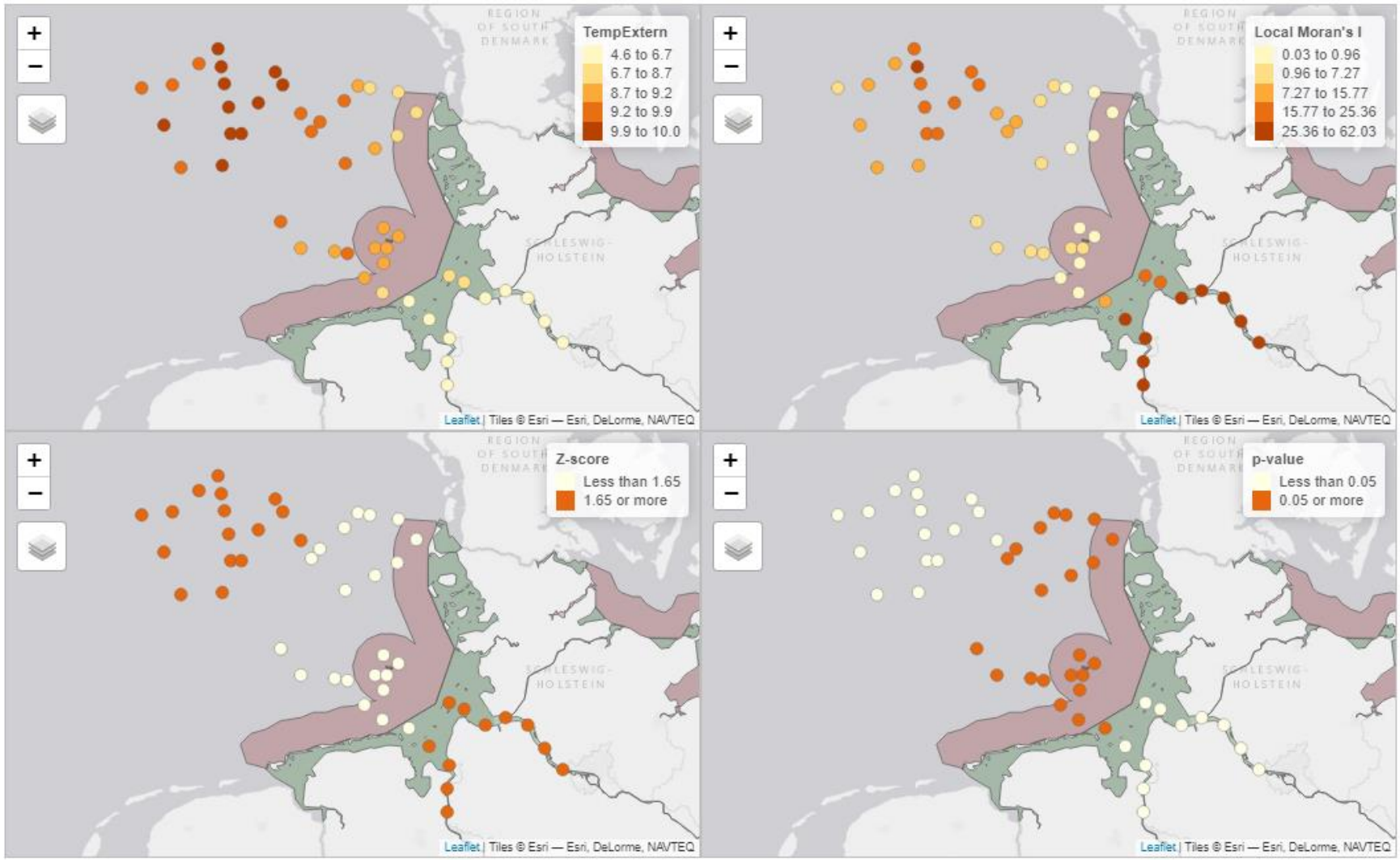
Apêndice III Mapas do Cálculo do I de Moran Local (LISA)

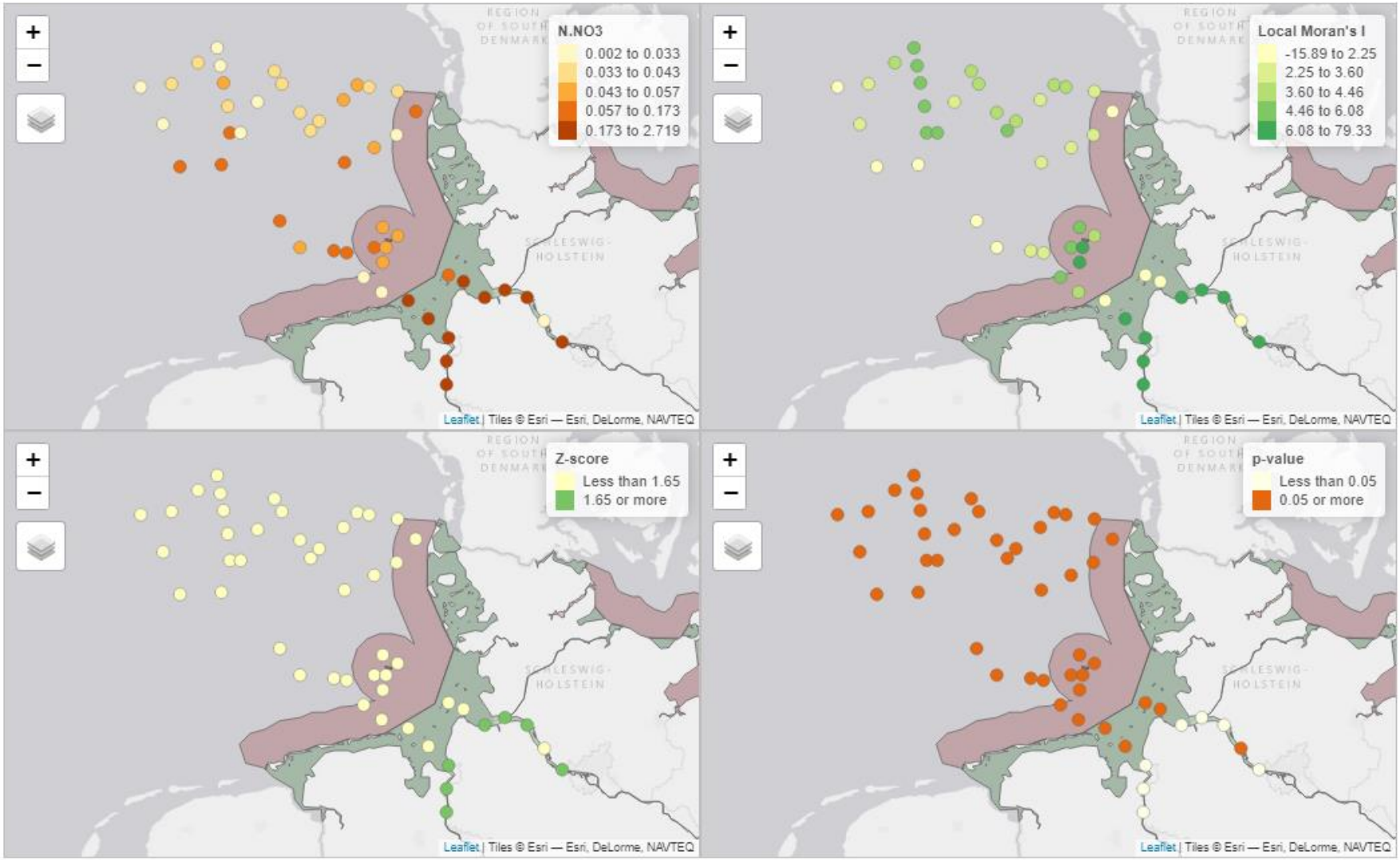
a. Para autocorrelação espacial positiva



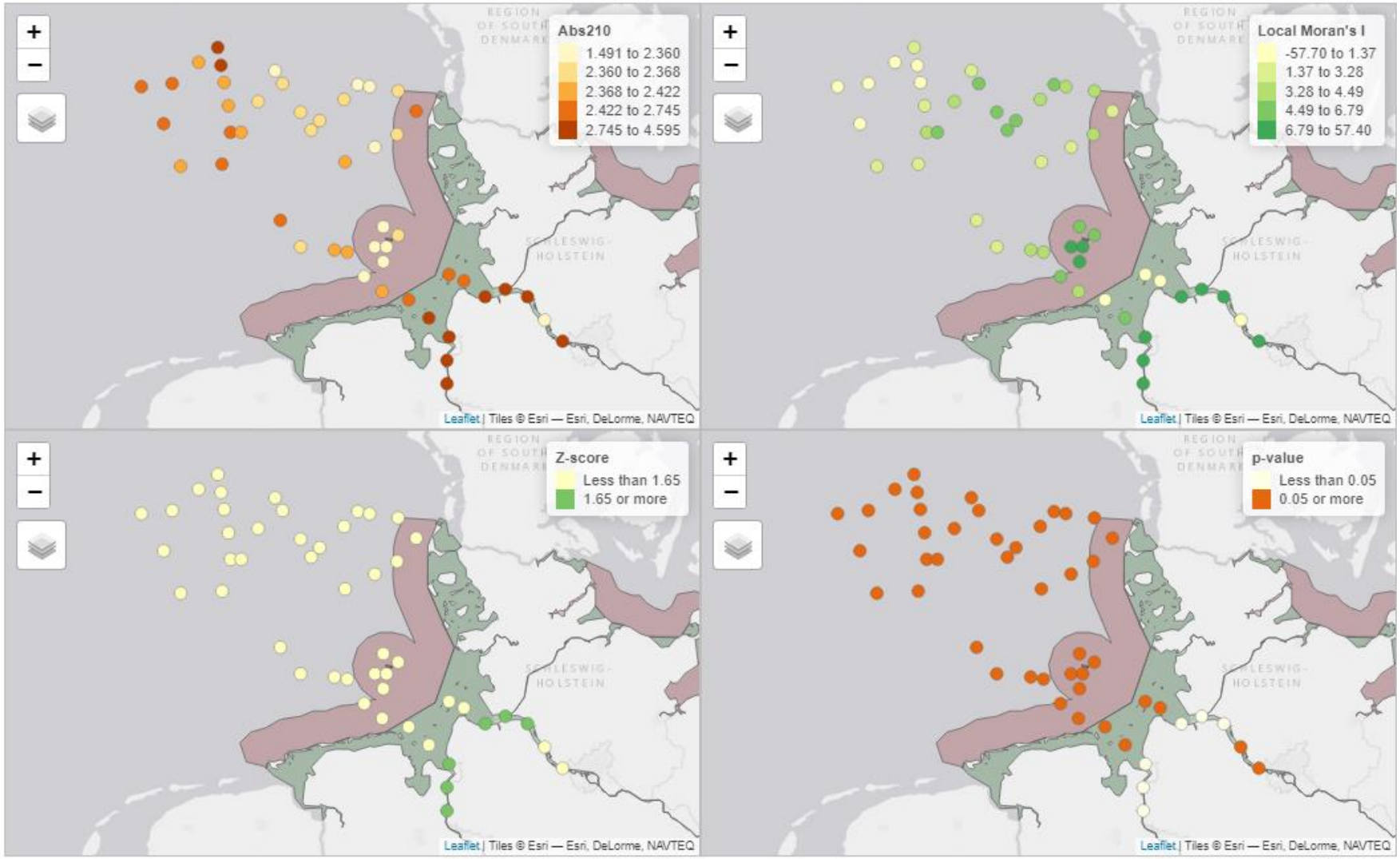

PC2
-1.212 to -0.392
-0.392 to -0.105
-0.105 to 0.294
0.294 to 0.397
0.397 to 2.679
Local Moran's I
-11.9 to 1.4
1.4 to 3.0
3.0 to 4.6
4.6 to 14.6
14.6 to 115.0
Z-score
Less than 1.65
1.65 or more
p-value
Less than 0.05
0.05 or more
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

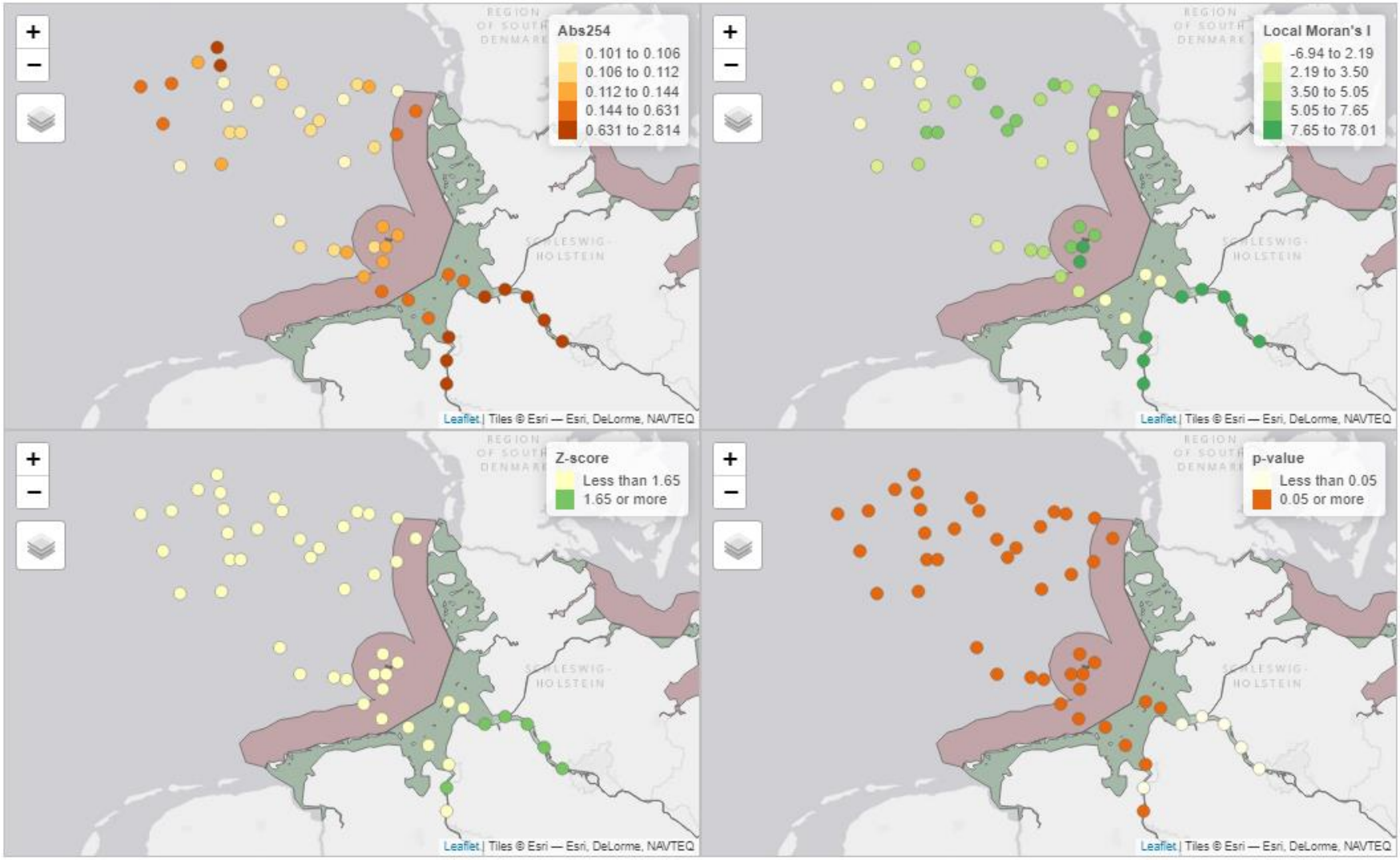
Local Moran's I
-1.40 to 2.77
2.77 to 4.17
4.17 to 5.75
5.75 to 7.47
7.47 to 78.65
p-value
Less than 0.05
0.05 or more
Salinity
0.51 to 29.59
29.59 to 33.30
33.30 to 34.01
34.01 to 34.29
34.29 to 34.68
Z-score
Less than 1.65
1.65 or more
Leaflet | Tiles @ Esri — Esri, DeLorme, NAVTEQ

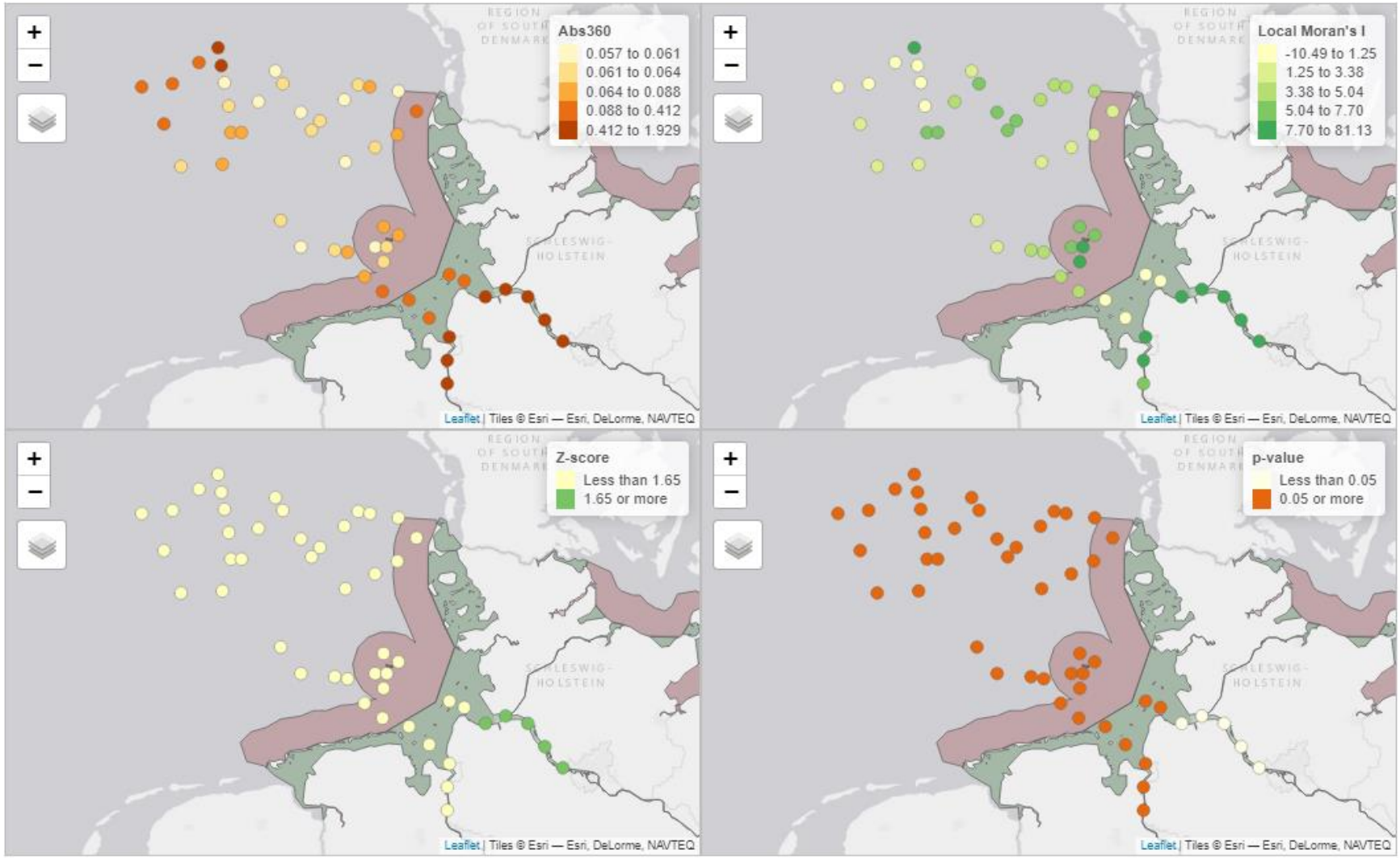

TempExtern
4.6 to 6.7
6.7 to 8.7
8.7 to 9.2
9.2 to 9.9
9.9 to 10.0
Local Moran's I
0.03 to 0.96
0.96 to 7.27
7.27 to 15.77
15.77 to 25.36
25.36 to 62.03
Z-score
Less than 1.65
1.65 or more
p-value
Less than 0.05
0.05 or more
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

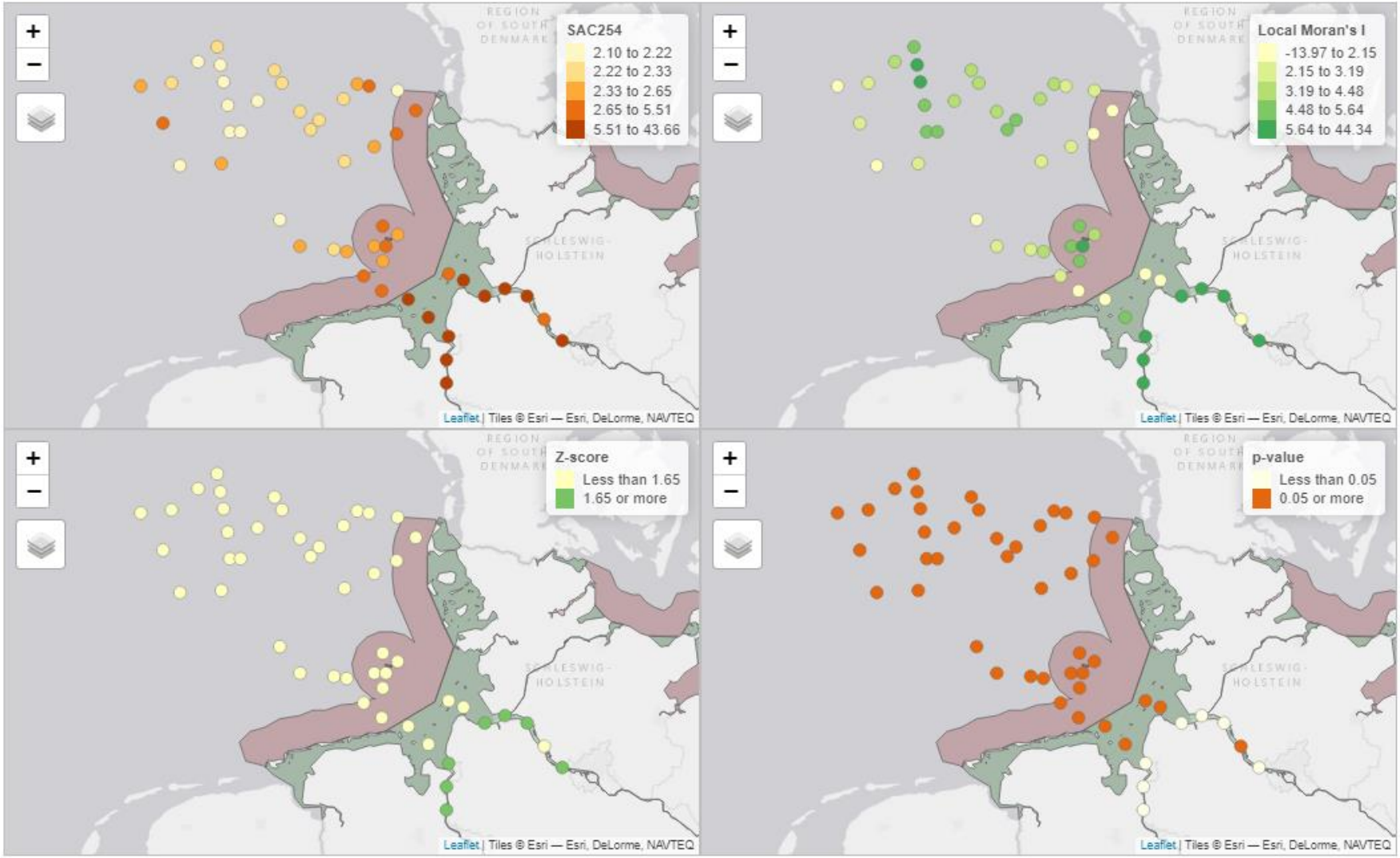

N.NO3
0.002 to 0.033
0.033 to 0.043
0.043 to 0.057
0.057 to 0.173
0.173 to 2.719
Local Moran's I
-15.89 to 2.25
2.25 to 3.60
3.60 to 4.46
4.46 to 6.08
6.08 to 79.33
Z-score
Less than 1.65
1.65 or more
p-value
Less than 0.05
0.05 or more
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

Local Moran's I
-3.58 to 2.25
2.25 to 3.19
3.19 to 3.83
3.83 to 5.02
5.02 to 50.00
p-value
Less than 0.05
0.05 or more
DOCeq
0.20 to 0.82
0.82 to 0.85
0.85 to 0.93
0.93 to 2.20
2.20 to 22.31
Z-score
Less than 1.65
1.65 or more
Leaflet | Tiles @ Esri — Esri, DeLorme, NAVTEQ

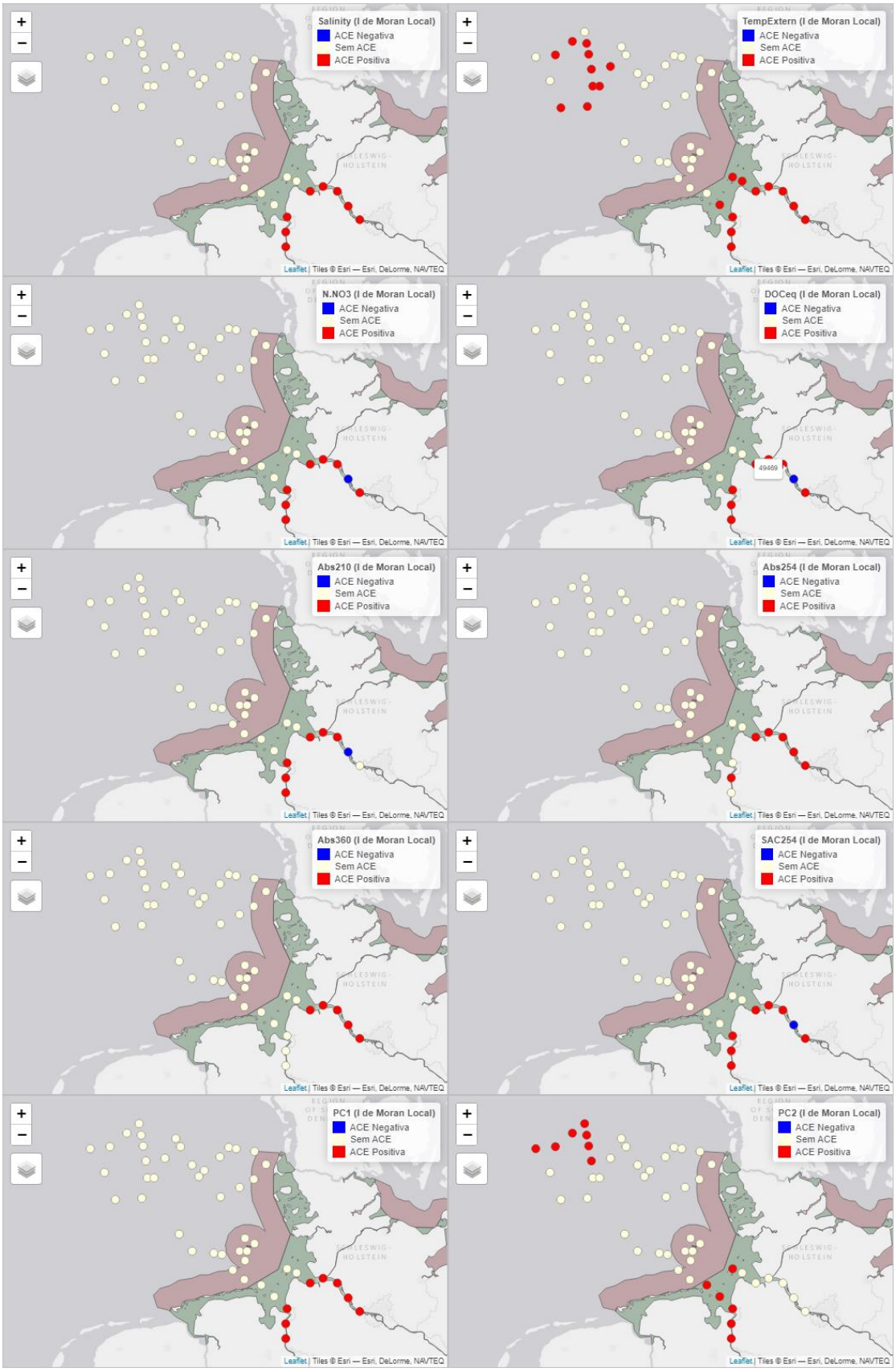

Abs210
1.491 to 2.360
2.360 to 2.368
2.368 to 2.422
2.422 to 2.745
2.745 to 4.595
Local Moran's I
-57.70 to 1.37
1.37 to 3.28
3.28 to 4.49
4.49 to 6.79
6.79 to 57.40
Z-score
Less than 1.65
1.65 or more
p-value
Less than 0.05
0.05 or more
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


Abs254
0.101 to 0.106
0.106 to 0.112
0.112 to 0.144
0.144 to 0.631
0.631 to 2.814
Local Moran's I
-6.94 to 2.19
2.19 to 3.50
3.50 to 5.05
5.05 to 7.65
7.65 to 78.01
Z-score
Less than 1.65
1.65 or more
p-value
Less than 0.05
0.05 or more
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


Abs360
0.057 to 0.061
0.061 to 0.064
0.064 to 0.088
0.088 to 0.412
0.412 to 1.929
Local Moran's I
-10.49 to 1.25
1.25 to 3.38
3.38 to 5.04
5.04 to 7.70
7.70 to 81.13
Z-score
Less than 1.65
1.65 or more
p-value
Less than 0.05
0.05 or more
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ


SAC254
2.10 to 2.22
2.22 to 2.33
2.33 to 2.65
2.65 to 5.51
5.51 to 43.66
Local Moran's I
-13.97 to 2.15
2.15 to 3.19
3.19 to 4.48
4.48 to 5.64
5.64 to 44.34
Z-score
Less than 1.65
1.65 or more
p-value
Less than 0.05
0.05 or more
Leaflet | Tiles © Esri — Esri, DeLorme, NAVTEQ

b. Para autocorrelação espacial



c. Mapa de *Clusters* de pontos significativos