UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Carlos Takeshi Kudo Yasojima

Modelo de Krigagem Automática Baseada em Agrupamento

Belém

2020

Carlos Takeshi Kudo Yasojima

Modelo de Krigagem Automática Baseada em Agrupamento

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas e Naturais como requisito parcial para obtenção do título de doutor.

Universidade Federal do Pará

Orientador: Jefferson Magalhães de Morais

Coorientador: Nelson Cruz Sampaio Neto

Belém

2020

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

Y11m Yasojima, Carlos Takeshi Kudo
Modelo de Krigagem Automática Baseada em Agrupamento / Carlos Takeshi Kudo Yasojima. — 2020.
86 f. : il. color.

Orientador(a): Prof. Dr. Jefferson Magalhães de Morais
Coorientador(a): Prof. Dr. Nelson Cruz Sampaio Neto
Tese (Doutorado) - Programa de Pós-Graduação em Ciência da Computação, Instituto de Ciências Exatas e Naturais, Universidade Federal do Pará, Belém, 2020.

1. Krigagem. 2. Algoritmos Bioinspirados. 3. Variograma. 4. Clusterização. 5. Interpolação Espacial. I. Título.

CDD 006.3

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CARLOS TAKESHI KUDO YASOJIMA

MODELO DE KRIGAGEM AUTOMÁTICA BASEADA EM AGRUPAMENTO

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal do Pará como requisito para obtenção do título de Doutor em Ciência da Computação, defendida e aprovada em 30/03/2020, pela banca examinadora constituída pelos seguintes membros:

Prof. Dr. Jefferson Magalhães de Morais
Orientador – PPGCC/UFPA

Prof. Dr. Nelson Cruz Sampaio Neto
Co-orientador – PPGCC/UFPA

Prof. Dr. Bianchi Serique Meiguins
Membro Interno – PPGCC/UFPA

Prof. Dr. João Marcelo Brazão Protázio
Membro Externo – UFPA

PAULO CERQUEIRA (26 de May de 2020 11:03 ADT)

Prof. Dr. Paulo Cerqueira dos Santos
Membro Externo – UFPA

Prof. Dr. Reginaldo Cordeiro dos Santos Filho
Membro Externo – UFPA

Visto:
Prof. Dr. Nelson Cruz Sampaio Neto
Coordenador do PPGCC/UFPA

Dedico esta tese a minha família.

Agradecimentos

Agradeço a minha esposa Waneila e filho Yuki, vocês são a minha maior felicidade e motivação.

Aos meus pais Edson, Rozinha e meu irmão Koiti, obrigado pela minha formação como pessoa e apoio nos momentos bons e difíceis durante toda minha vida.

Obrigado aos professores Morais e Terezinha por terem me acolhido durante o mestrado e nos primeiros anos do doutorado.

Obrigado aos professores Jefferson e Nelson pela orientação e minha formação como pesquisador. Ensinamentos e filosofias de trabalho que levarei para o resto da vida.

Obrigado ao professor Protázio por todo apoio na área da geoestatística, fundamental para este tese.

Obrigado ao professor Bianchi, pelo apoio na visualização da informação.

Obrigado aos professores Reginaldo e Paulo pela participação na banca avaliadora e pelo aprimoramento da tese.

A todos que diretamente ou indiretamente participaram em algum ponto da minha trajetória, deixo também meus agradecimentos.

"A persistência é o caminho do êxito"
(Charles Chaplin)

Resumo

A krigagem é uma técnica de interpolação da geoestatística que realiza a predição de medições e observações em localidades desconhecidas com base em dados previamente coletados. A qualidade da predição deste método é diretamente ligada à qualidade da modelagem do variograma teórico. O método convencional e muito utilizado da modelagem do variograma teórico, consiste na utilização de conhecimento especialista e estudo aprofundado para determinar quais são os parâmetros adequados para a modelagem. No entanto, essa situação não é sempre possível, e nesses casos, torna-se interessante a aplicação de um processo automático. Diante deste cenário, este trabalho propõe um modelo para automatizar etapas do processo de krigagem incluindo a modelagem do variograma teórico. O modelo proposto baseia-se em técnicas de pré-processamento, clusterização de dados, algoritmos bioinspirados e a classificação via K-vizinhos mais próximos. O desempenho do modelo foi avaliado utilizando duas bases de dados, sendo os resultados comparados com de outras técnicas de otimização consolidadas na literatura de krigagem. Os impactos da etapa de clusterização na hipótese da estacionariedade também é investigada por meio da aplicação de técnicas de remoção de *trends*. Os resultados demonstraram que nesta proposta automatizada, a clusterização alcança os melhores resultados na predição da krigagem. No entanto, a divisão da base de dados em subgrupos por consequência gera dados não estacionários. Algoritmos genéticos e bioinspirados em geral são facilmente configurados com base em uma heurística para definir os *ranges* (limites máximos e mínimos) das variáveis em comparação com outras técnicas estudadas. A classificação via K-vizinhos mais próximos é satisfatória em solucionar problemas causados pela tarefa de clusterização e alocando pontos desconhecidos nos *clusters* previamente definidos.

Palavras-chave: Interpolação espacial. Modelagem do Variograma. Clusterização. Algoritmos bioinspirados. Krigagem.

Abstract

Kriging is a geostatistical interpolation technique that performs the prediction of observations in unknown locations through previously collected data. The modelling of the variogram is an essential step of the kriging process because it drives the accuracy of the interpolation model. The conventional method of variogram modelling consists of using specialized knowledge and in-depth study to determine which parameters are suitable for the theoretical variogram. However, this situation is not always possible, and in this case, it becomes interesting to use an automatic process. Thus, this work aims to propose a new methodology to automate the estimation of theoretical variogram parameters of the kriging process. The proposed methodology is based on preprocessing techniques, data clustering, genetic algorithms and the K-Nearest Neighbor classifier. The performance of the methodology was evaluated using two databases and it was compared to other optimization techniques widely-used in the literature. The impacts of the clustering step on the stationary hypothesis was also investigated with and without trends removal techniques. The results showed that in this automated proposal, the clustering process increases the accuracy of the kriging prediction. However, it generates groups that might not be stationary. Genetic algorithms are easily configurable with the proposed heuristic when setting the variable ranges in comparison to other optimization techniques, and the KNN method is satisfactory in solving some problems caused by the clustering task and allocating unknown points into previously determined clusters.

Keywords: Spatial interpolation. Variogram fitting. Clustering. Bioinspired algorithms. Kriging.

Lista de ilustrações

Figura 1. Fluxo do processo de krigagem. Adaptado de (JAKOB; YOUNG, 2016) 17
Figura 2. Exemplo de um modelo final de variograma. 24
Figura 3. Exemplo de anisotropia. 25
Figura 4. Exemplo de classificação utilizando o algoritmo KNN (Adaptado de JOSÉ, 2017). 28
Figura 5. Ciclo de execução de um algoritmo genético. 30
Figura 6. Protocolo da revisão sistemática adaptada de Kitchenham, 2004. 39
Figura 7. Artigos científicos publicados no *Science Direct* - Palavra chave: Kriging. 42
Figura 8. Artigos científicos publicados no IEEExplore - Palavra chave: Kriging. 42
Figura 9. Artigos científicos publicados no Science Direct - Palavra chave: Kriging + Variogram + Optimization. 43
Figura 10. Artigos científicos publicados no IEEExplore - Palavra chave: Kriging + Variogram + Optimization. 43
Figura 11. Áreas de pesquisa objetivo de cada publicação. 44
Figura 12. Cluster-Kriging (ABEDINI; NASSERI; ANSARI, 2008). 47
Figura 13. Funções de custo utilizadas nas pesquisas. 49
Figura 14. Tipos de dados utilizados nas pesquisas. 50
Figura 15. Esquema geral das etapas de pré-processamento, clusterização e ajuste de modelos para cada *cluster* separadamente. 52
Figura 16. Exemplo da aplicação do algoritmo KNN para minimização da sobreposição de clusters. (a) Antes da normalização + KNN. (b) Após normalização + KNN. 54
Figura 17. Testes realizados com diferentes vizinhos utilizando o algoritmo genético para otimização dos parâmetros do variograma teórico. 54
Figura 18. Esquema do cromossomo utilizado na configuração do AG. 55
Figura 19. Estrutura espacial original da base de dados *Meuse*. 56
Figura 20. Estrutura espacial da base de dados *Meuse* indicando, em vermelho, os *outliers* identificados. 57
Figura 21. Estrutura espacial da base de dados *Meuse* final após etapa de normalização e detrending. 58
Figura 22. Pontos selecionados para treino (azul) e teste (vermelho) para uma iteração do modelo. 58
Figura 23. Clusterização da base de dados com 2 clusters utilizando a técnica de agrupamento K-Means + KNN. 59
Figura 24. Clusterização da base de dados com 3 clusters utilizando a técnica de agrupamento K-Means + KNN. 59

Figura 25. Pontos desconhecidos (destacados com a borda preta) alocados via KNN com 3 vizinhos mais próximos para a base com 2 clusters. 60
Figura 26. Pontos desconhecidos (destacados com a borda preta) alocados via KNN com 3 vizinhos mais próximos para a base com 3 clusters. 61
Figura 27. Estrutura espacial da base de dados Wolfcamp. 63
Figura 28. Estrutura espacial da base de dados Meuse. 63
Figura 29. Curvas de convergência para 1 até 5 clusters. Cada linha representa a média de 10 execuções do AG e PSO. 66
Figura 30. SPD para AG e PSO. Média de 10 execuções para configuração com 1 *cluster*. 66
Figura 31. Boxplot dos MSE considerando a etapa de classificação para o AG e PSO. Média de 10 execuções para cada número de *cluster*. 67
Figura 32. Testes com diferentes valores de limite máximo para o parâmetro sill para a otimização do AG. O eixo Y representa o erro e o eito X representa 4 valores de limites máximos testados: A variância da variável objetivo multiplicado por 1, 5, 10 e 15. 69
Figura 33. Variogramas experimentais das bases Meuse e Wolfcamp, antes e depois do processo de *detrending*. 70
Figura 34. Base de dados Meuse. Valores de zinco nas coordenadas X e Y - (a) Antes do processo de *detrending* e (b) após o processo de *detrending*. . 71
Figura 35. Base de dados Wolfcamp. Valores de nível piezométrico nas coordenadas X e Y - (a) Antes do processo de *detrending* e (b) após o processo de *detrending*. 71
Figura 36. Representação treemap do índice NMSE nas diversas configurações testadas durante os experimentos utilizado as bases de dados Meuse e Wolfcamp. 73
Figura 37. Mapas de krigagem, estimativas e variância, para a base de dados Meuse. 77
Figura 38. Mapas de krigagem, estimativas e variância, para a base de dados Wolfcamp. 78

Lista de tabelas

Tabela 1. *Scores* de três grupos correspondentes sob quatro condições. 35
Tabela 2. Comparativo entre os modelos dos artigos levantados. 45
Tabela 3. Coordenadas X e Y e valores dos outliers identificados. 57
Tabela 4. Informações das bases de dados. 62
Tabela 5. Parâmetros do AG e PSO. 64
Tabela 6. Melhor Fitness/MSE, média dos fitness e desvio padrão dos fitness. . . 65
Tabela 7. Lista dos algoritmos de otimização e funções de custo utilizados nos experimentos. 67
Tabela 8. *Seed* inicial para os algoritmos GN-ILS1, GN-ILS2 e LM-WLS. 68
Tabela 9. Heurística aplicada ao AG proposto. 68
Tabela 10. Porcentagem da ocorrência de *trends* nos grupos com base no índice de Mann-Kendall, considerando as configurações com 1 à 3 clusters. . . . 72
Tabela 11. Estatísticas descritivas com e sem a etapa de *detrending*. 74
Tabela 12. Estatísticas descritivas dos algoritmos de clusterização. 74
Tabela 13. Estatísticas descritivas considerando o número de clusters. 75
Tabela 14. Resultados do teste Bonferroni considerando diferente número de clusters. 75
Tabela 15. Estatísticas descritivas considerando os algoritmos de otimização. . . . 76
Tabela 16. Resultados do teste Bonferroni com diferentes algoritmos de otimização. 76

Lista de abreviaturas e siglas

AG	Algoritmo Genético
IA	*Inteligência artificial*
KNN	*K-Nearest neighbor*
ML	*Maximum likelihood*
MSE	*Mean squared error / Erro médio quadrático*
NMSE	*Normalized mean squared error / Erro médio quadrático normalizado*
OK	*Ordinary kriging / Krigagem ordinária*
PSO	*Particle swarm optimization*
SPD	*Standard population diversity / Diversidade padrão da população*
Vario. Param.	*Variogram parameters / Parâmetros do variograma*

Lista de símbolos

χ	Letra grega chi
Γ	Letra grega Gama
κ	Letra grega kappa
λ	Letra grega lambda
σ	Letra grega sigma
θ	Letra grega theta
φ	Letra grega phi

Sumário

1 INTRODUÇÃO 16
1.1 Contextualização 16
1.2 Motivação e Justificativa 20
1.3 Objetivos 21
1.4 Publicações 21
1.5 Organização da tese 22

2 REFERENCIAL TEÓRICO 23
2.1 Krigagem 23
2.2 Funções de custo 26
2.3 Hipótese da estacionariedade 26
2.4 K-vizinhos mais próximos 27
2.5 k-means 29
2.6 Algoritmos genéticos 29
2.7 Enxame de partículas 31
2.8 Métrica de avaliação 33
2.9 Índice de diversidade da população 34
2.10 Testes Estatísticos 34
2.10.1 Teste de Friedman 34
2.10.2 Teste de Mann-Kendall 35
2.10.3 Teste de Shapiro Wilk 36
2.10.4 Teste T Pareado e One-Way Anova para medidas Repetidas 37
2.10.5 Teste de Bonferroni 38

3 TRABALHOS RELACIONADOS 39
3.1 Revisão sistemática 39
3.2 Atividade: Planejamento 40
3.2.1 Identificação da necessidade de revisão 40
3.2.2 Desenvolvimento do protocolo 40
3.2.3 Condução 41
3.2.4 Selecionar estudos primários 41
3.2.5 Extração dos dados 41
3.2.6 Síntese dos dados 41
3.3 Resultados 41
3.3.1 Visão geral dos estudos 42
3.3.2 Discussão geral das pesquisas 45

3.3.3 Discussão das pesquisas: Questão secundária 1 . . . 46
3.3.4 Discussão das pesquisas: Questão secundária 2 . . . 47
3.3.5 Discussão das pesquisas: Questão secundária 3 . . . 48
3.3.6 Discussão das pesquisas: Questão secundária 4 . . . 49

4 MODELO PROPOSTO . . . 51
4.1 Pré-processamento dos dados . . . 53
4.2 Clusterização dos dados . . . 53
4.3 Otimização . . . 55
4.4 Classificação . . . 55
4.5 Passo a passo . . . 56
4.5.1 Dados originais . . . 56
4.5.2 Remoção de outliers, normalização e detrending . . . 56
4.5.3 Particionamento . . . 57
4.5.4 Clusterização . . . 58
4.5.5 Otimização . . . 60
4.5.6 Alocação de pontos desconhecidos . . . 60
4.5.7 Krigagem e geração de mapas da krigagem . . . 60

5 EXPERIMENTOS E RESULTADOS . . . 62
5.1 Bases de dados . . . 62
5.2 Comparativo das técnicas bioinspiradas: AG e PSO . . . 64
5.3 Organização dos experimentos . . . 66
5.4 Discussão . . . 69
5.5 Processo de *detrending* . . . 73
5.6 Clusterização dos dados . . . 73
5.7 Algoritmos de otimização . . . 75
5.8 Mapas de krigagem . . . 76

6 CONCLUSÕES E TRABALHOS FUTUROS . . . 79

REFERÊNCIAS . . . 81

1 Introdução

1.1 Contextualização

A análise espacial de dados é fundamental em aplicações cujo objetivo é mapear uma superfície contínua, levando em consideração um conjunto discreto de localizações dentro de uma área de estudo. Tais aplicações são encontradas em várias ciências (sobretudo as ciências da terra) tais como geologia, para estimação de depósitos minerais; agronomia, para o zoneamento agrícola; construção civil, para o monitoramento de obras como barragens, entre outras.

O desafio comum dessas aplicações é que as variáveis geralmente não podem ser completamente caracterizadas apenas por um padrão de distribuição requerido pela estatística clássica como normalidade e independência de dados. Isto porque a maioria dos dados são altamente distorcidos e/ou possuem correlação espacial, ou seja, valores de dados de locais mais próximos tendem a ser mais semelhantes do que valores de dados de locais mais afastados.

Comparada às estatísticas clássicas que examinam a distribuição estatística de um conjunto de dados, a geoestatística incorpora, além dessa distribuição, a correlação espacial entre os dados. Devido à essa diferença, a maioria das aplicações, que apresentam componente espacial em seus dados, usam métodos da geoestatística para gerar resultados mais eficazes.

O principal objetivo da geoestatística é realizar a predição da distribuição de uma propriedade. Esta predição frequentemente toma a forma de um mapa ou uma série de mapas. Neste cenário, dentre as formas de predição, tem-se a chamada estimativa. Na estimativa um único mapa do fenômeno espacial é produzido, com base na amostra (dados) e em um modelo (variograma) representando a correlação espacial da amostra. Este estimativa (mapa) é produzido a partir de um processo denominado de Krigagem (HENGL, 2009).

A krigagem é uma técnica de interpolação da geoestatística que faz a predição de valores de medições em locais desconhecidos baseados em dados previamente coletados (HENGL, 2009). O erro da krigagem ou erro da interpolação é minimizado pelo estudo e modelagem da distribuição e variabilidade espacial das amostras coletadas. Essa distribuição ou variabilidade espacial é expressa na forma de variogramas experimentais.

O variograma experimental pode ser considerado como uma representação gráfica da distribuição de dados, além de mostrar a variância com o incremento da distância das amostras coletadas. O variograma é a base para a aplicação do método de Krigagem.

Assim, pode-se dividir o processo de krigagem em três etapas fundamentais (ZHANG, 2011): (1) Examinar a similaridade entre um conjunto de amostras coletadas via análise do variograma experimental; (2) Ajustar uma função matemática (modelo) ao variograma experimental, chamado de variograma teórico; (3) Aplicar a interpolação por krigagem com base na função definida no passo 2. Este fluxo do processo da krigagem é ilustrado na Figura 1.

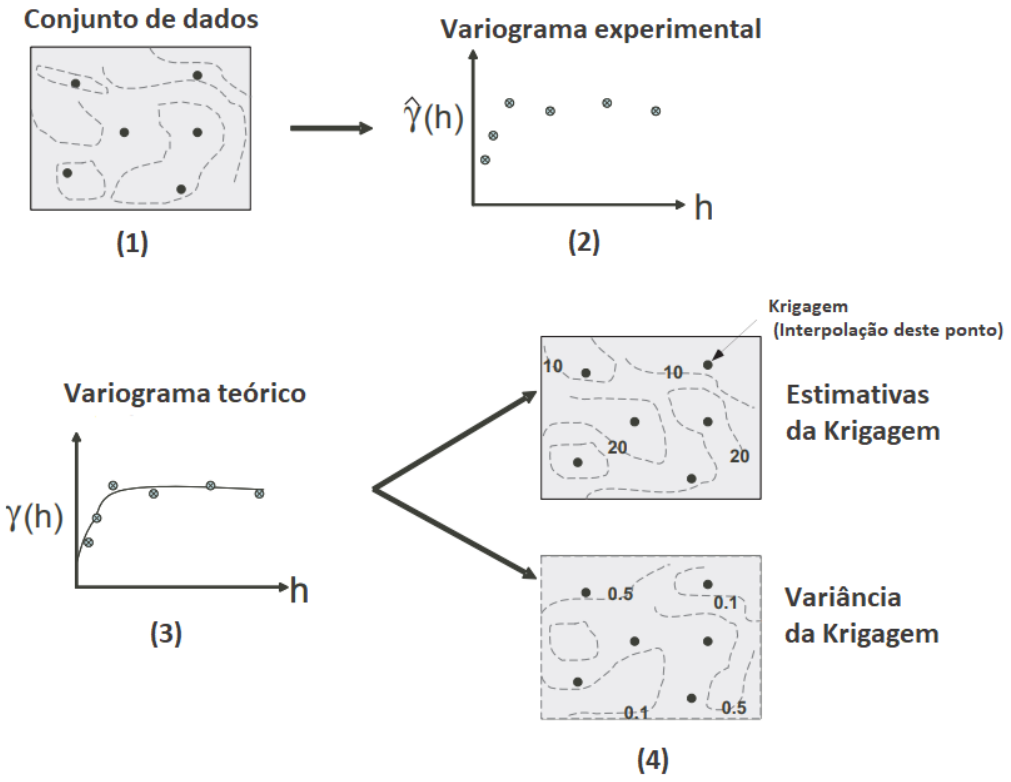

Figura 1. Fluxo do processo de krigagem. Adaptado de (JAKOB; YOUNG, 2016)

Apesar de avanços na área da geoestatística, especificamente no processo de krigagem, a modelagem do variograma teórico ainda permanece um desafio, já que está diretamente ligado a acurácia da interpolação do método. Para a modelagem do variograma teórico, é necessário a estimativa dos parâmetros relacionados à ele, sendo que este processo pode ser caracterizado como um problema de otimização.

Para ilustrar o desafio que é a modelagem do variograma teórico, cita-se o experimento conduzido pela *U.S. Environmental Protection Agency* (ZHANG, 2011). No experimento, 12 geoestatísticos foram convocados para realizar a modelagem do variograma teórico no processo da krigagem de uma mesma base de dados. Os resultados obtidos pelos 12 geoestatísticos foram totalmente divergentes tanto na escolha dos modelos do variograma quanto na seleção dos parâmetros relacionados à ele. A pesquisa conclui que se desconhece um algoritmo universal aceito para determinar o modelo do variograma, que a validação cruzada não é garantia que as estimativas produzirão valores satisfatórios em

pontos desconhecidos ou não coletados, e que todas as decisões tomadas na maioria dos casos, a partir de um estudo geoestatístico, são realizados com base em estudo exploratório dos dados.

Recentemente, técnicas de inteligência artificial têm sido aplicadas para aprimorar o processo de krigagem como em (WEI; LIU; CHEN, 2010; XIALIN et al., 2010; GONÇALVES; KUMAIRA; GUADAGNIN, 2017; LI et al., 2018; ABEDINI; NASSERI; BURN, 2012; HUIZAN et al., 2008; WANG et al., 2017b). Como afirmado em Gonçalves, Kumaira e Guadagnin (2017) e aplicado em Li et al. (2018), algoritmos bio-inspirados, como algoritmos genéticos (AG), e enxame de partículas (PSO - *Particle Swarm Optimization*) (WANG et al., 2017b) são adequados para auxiliar no processo de definição dos parâmetros do variograma teórico.

De acordo com Gonçalves, Kumaira e Guadagnin (2017), se há necessidade de muitos parâmetros e trabalhar com a existência de anisotropia, fenômeno em que a variabilidade espacial é distinto para mais de uma direção, deve-se optar por métodos complexos de otimização, por exemplo, algoritmos genéticos. Além disso, estes algoritmos não necessitam de "*seed*" inicial como entrada, mas sim intervalos de mínimos e máximos, diferente de métodos clássicos como Gauss-Newton (DESASSIS; RENARD, 2013) e Levenberg-Marquadt (MARQUARDT, 1963).

Considerando um processo automatizado, esses "*seeds*" e intervalos precisam ser definidos com base nos dados de estudo, no entanto, isso se torna uma tarefa difícil, já que não existem heurísticas bem definidas para este propósito. Em Larrondo, Neufeld e Deutsch (2003), os autores aplicaram uma técnica de minimização numérica para otimização dos parâmetros do variograma teórico e propuseram uma heurística para definir o "*seed*" inicial na estrutura automatizada do método. A pesquisa de Pesquer, Cortés e Pons (2011) apresentou a aplicação da técnica de otimização Levenberg-Marquadt com função de custo de mínimos quadrados. Nenhuma heurística foi especificada, porém os autores testaram diferentes valores iniciais de *seeds* na etapa de otimização. O trabalho de Li e Lu (2010) também utilizou técnicas de inteligência artificial (IA), como programação aprimorada linear, em conjunto com a Krigagem, utilizando parâmetros com base em tentativa e erro. Neste sentido, torna-se interessante a melhoria na definição de intervalos (mínimo e máximo) em um processo automatizado, consirando somente a base de dados de estudo e os algoritmos bioinspirados como algoritmos genéticos utilizados nesta tese.

Em Abedini, Nasseri e Ansari (2008), os autores propuseram uma nova metodologia para a aplicação da krigagem utilizando a técnica de clusterização K-means. A proposta consiste na criação de subgrupos da base de dados e realização da predição de pontos utilizando apenas informação do subgrupo que este ponto pertence. O método proposto apresentou melhores resultados que a krigagem convencional. Contudo, os autores utilizaram o mesmo modelo e parâmetros do variograma teórico para todos os subgrupos

de dados, o que pode ser aprimorado, já que as características de cada subgrupo podem ser diferentes. Em (WANG et al., 2017a), cada conjunto de dados possui seu próprio variograma teórico utilizado na etapa de interpolação/predição dos pontos. Contudo, foi aplicado apenas à dados não-espaciais e também não foi detalhado a forma de definição dos variogramas teóricos.

A clusterização de dados apresenta alguns problemas quando aplicados à dados espaciais, como a sobreposição de *clusters*. Como exposto em Fouedjio (2017), este comportamento, de criação de *clusters* espacialmente dispersos (sobreposição de *clusters*) não é desejável em aplicações da geoestatística, já que impacta nas propriedades da dependência espacial sobre a área de estudo, e é importante manter as características originais da base de dados. Recentemente, pesquisadores buscaram solucionar o problema da sobreposição de *clusters* por meio do desenvolvimento de métodos que garantam a continuidade espacial, onde os subgrupos criados são uniformes. Neste sentido, em Abedini, Nasseri e Ansari (2008), um fator de normalização foi utilizado para minimizar este problema. O método proposto em Chavent et al. (2018) apresentou um novo algoritmo de clusterização hierárquica atribuindo pesos tanto para as coordenadas espaciais quanto para a variável em estudo. Além disso, Fouedjio (2017) propôs a clusterização espectral para manter as propriedades de continuidade espacial. Para este problema, é proposto nesta tese um novo modelo através da aplicação do algoritmo K-means aprimorado pelo classificador K-Vizinhos mais próximos (K-Nearest Neighbors).

Outro ponto a ser ponderando, quando são incluídas as coordenadas espaciais como variáveis no processo de clusterização, a hipótese de estacionariedade não poder ser garantida (FOUEDJIO, 2017). A krigagem requer que esta hipótese seja satisfeita, ou seja, que algumas propriedades como média, variância, entre outras, sejam constantes ao longo do domínio espacial, caso contrário, tem-se o fenômeno chamado de *trend*. O processo de remoção de *trends*, também chamado de *detrending*, garante a hipótese de estacionariedade dos dados (VIEIRA et al., 2010). Assim, para avaliar os impactos do modelo proposto nesta tese, os experimentos foram realizados com e sem o processo de *detrending*.

Uma questão importante que não foi explorada nesses trabalhos relacionados é quando um novo ponto no domínio espacial precisa ser interpolado. Como ele deve pertencer a um subgrupo (*cluster*), é necessário um mecanismo para alocá-lo em um dos subgrupos definidos na etapa de clusterização. Então, para esta tarefa de alocação, este trabalho propõe a aplicação do algoritmo KNN.

Com base nos problemas expostos relacionados à todo o processo de krigagem, este trabalho propõe um modelo para melhorar as estimativas de krigagem, usando conceitos de pré-processamento, clusterização de dados e métodos de otimização. Na abordagem, para fins de comparação, os procedimentos de agrupamento K-means e clusterização

Hierárquica (CHAVENT et al., 2018) foram utilizados para separar o conjunto de dados em grupos por similaridade. O algoritmo KNN executa a tarefa de alocação de novos dados em seu respectivo *cluster* e resolve alguns problemas ocorridos na etapa de clusterização do método K-means quando aplicados à dados espaciais, como a sobreposição de *clusters*. Para analisar os impactos na hipótese estacionária, foram aplicados teste não paramétrico de Mann-Kendall (POHLERT, 2016). Por fim, algoritmos bio-inspirados foram utilizados para modelar um variograma teórico específico para cada grupo encontrado na etapa de agrupamento, com definição dos limites (mínimo e máximo) dos parâmetros com base nos dados de cada grupo. AG e PSO foram selecionados para servir de *baseline* para o modelo proposto e também porque foram aplicadas em metodologias de krigagem semelhantes, como pode ser visto em (XIALIN et al., 2010; LI et al., 2018; ABEDINI; NASSERI; BURN, 2012; YASOJIMA et al., 2019a). O modelo proposto foi avaliado usando duas bases de dados publicamente disponíveis e os resultados comparados com outras técnicas de otimização apresentadas na literatura.

1.2 Motivação e Justificativa

Nesta tese, a hipótese é que o *cluster kriging* juntamente com variogramas específicos para cada subgrupo de dados produza melhores resultados. A criação de um modelo automatizado e bem estruturado garante a redução da dependência de conhecimento especialista. Como dito anteriormente, a aplicação da interpolação de dados espaciais, mais especificamente a krigagem, abrange diversas áreas do conhecimento e torna-se custosa quando necessita-se de profundo estudo e análise dos dados espaciais. Remover esta etapa de análise profunda e de conhecimento especialista possibilita a utilização de um mecanismo de entrada e saída de dados transparante para o usuário, gerando modelos e estimativas confiáveis independente das características das bases.

Outro ponto de destaque é uma maior confiabilidade no resultado obtido, apesar de ser diretamente dependente da robustez e da boa estrutura do modelo automático, reduz a possibilidade de ruído inerente a intervenção humana no processo.

Além do aspecto técnico, de um novo modelo, este trabalho ainda contribui para a literatura no estudo do processo de *detrending* e o impacto que este fenômeno têm na clusterização de dados espaciais. Apesar de ser ainda uma análise inicial, esta pesquisa provê um bom subsídio para trabalhos futuros e estudos mais aprofundados no tema.

Como será demonstrado no Capítulo 3 posteriormente, dentro das pesquisas de trabalhos relacionados nos últimos 12 anos (2008-2019), existe uma lacuna que pode ser explorada ao analisar as técnicas aplicadas no aprimoramento da krigagem. Diversos métodos foram utilizados de forma individual (isoladamente) e seria de suma importância aplicá-las, de forma conjunta, neste único modelo.

1.3 Objetivos

Como objetivo geral, este trabalho propõe um novo modelo para melhorar as estimativas de krigagem, usando conceitos de pré-processamento, clusterização de dados e métodos de otimização.

Para atingir o objetivo geral desta tese, são propostos os seguintes objetivos específicos:

- Propor uma nova técnica para a clusterização de dados espaciais, minimizando a sobreposição de *clusters*, e comparando à outras técnicas da literatura.
- Propor um modelo automático para a definição dos variogramas teóricos para cada sub-grupo criado na etapa de clusterização em um base de dados espaciais.
- Propor uma heurística para definição automática dos limites mínimos e máximos das variáveis utilizadas nos algoritmos bio-inspirados para otimização dos parâmetros do variograma teórico.
- Realizar estudo comparativo do modelo proposto utilizando, ou não, técnicas de *detrending*.
- Avaliar a performance do algoritmo genético em relação a técnicas bem consolidadas na literatura na otimização de parâmetros do variograma teórico.
- Realizar um estudo bibliográfico acerca do tema Krigagem, verificando o estado da arte.

1.4 Publicações

O desenvolvimento das atividades desta tese produziram as seguintes publicações:

- *A Comparison of Genetic Algorithms and Particle Swarm Optimization to Estimate Cluster-Based Kriging Parameters* na *EPIA Conference on Artificial Intelligence* em 2019. Qualis B1.
- *Evaluation of Bio-Inspired Algorithms in Cluster-Based Kriging Optimization* na *International Conference on Computational Science and Its Applications* em 2019. Qualis B1.
- *A New Methodology for Automatic Cluster-Based Kriging Using K-Nearest Neighbor and Genetic Algorithms* na *Information Journal* em 2019. Qualis B2.

- Publicação do código desenvolvido nesta tese no repositório *GitHub* https://github.com/LABVIS-UFPA/GAClusterKriging. Incluindo manual de utilização, bases e bibliotecas utilizadas. Última atualização Fev 2019.

1.5 Organização da tese

Esta tese está dividida em 6 capítulos:

- O Capítulo 1 contextualiza o estudo realizado, apresentando a motivação, justificativa e os objetivos a serem alcançados.
- O Capítulo 2 apresenta o referencial teórico das técnicas utilizadas no modelo proposto.
- O Capítulo 3 faz um levantamento dos trabalhos relacionados e aborda os principais métodos e teorias, da inteligência artificial, utilizados na aprimoramento do processo de krigagem.
- O Capítulo 4 apresenta o modelo proposto, detalhando as etapas que compõe o processo como um todo.
- No Capítulo 5 são apresentados os resultados obtidos por meio da aplicação do modelo construído além de comparativos com outras técnicas disponíveis na literatura.
- Por fim, no Capítulo 6, são discutidas as conclusões desta tese, entre elas, as principais contribuições, as publicações obtidas, as limitações e propostas de trabalhos futuros.

2 Referencial Teórico

Este Capítulo tem como objetivo descrever a fundamentação teórica das técnicas abordadas nesta tese, além de métricas utilizadas para realizar o comparativo entre os diversos métodos testados para aprimoramento da krigagem.

2.1 Krigagem

Krigagem é uma técnica de interpolação amplamente utilizada na geoestatística para predição de dados espaciais. Este método leva em consideração as características da autocorrelação de variáveis regionalizadas, ou seja, que possuem dependência espacial como coordenadas X e Y. Essas variáveis têm alguma continuidade ou variância espacial, o que permite que os dados obtidos por amostragem de pontos específicos sejam utilizados para parametrizar a estimativa de pontos, onde o valor da variável é desconhecida (HENGL, 2009).

Seja Z um conjunto de observações de uma variável objetivo (variável de resposta) denotada por $\{z(s_1),\ z(s_2),\ \ldots,\ z(s_N)\}$, onde $s_i = (x_i, y_i)$ é um ponto em um espaço geográfico bidimensional; x_i e y_i são suas coordenadas (localizações principais); e N é o número de observações.

Os valores da variável objetivo em algum novo local s_0 podem ser derivados usando um modelo de predição espacial. A versão padrão da krigagem é chamada de krigagem ordinária (OK), onde as estimativas são baseadas no modelo:

$$\hat{z}_{OK}(s_0) = \sum_{i=1}^{N} w_i(s_0).z(s_i) = \lambda_0^T.\mathbf{z} \tag{2.1}$$

onde λ_0 é um vetor de pesos da krigagem (w_i), e $\mathbf{z}$ é o vetor de N observações na localização primária.

Portanto, para realizar a estimativa dos pesos, é necessário calcular as semivariâncias $\gamma(h)$ baseadas nas diferenças entre os valores da vizinhança (pontos vizinhos).

$$\gamma(h) = \frac{1}{2}E[(z(s_i) - z(s_i + h))^2] \tag{2.2}$$

onde $z(s_i)$ é a observação da variável objetivo em uma determinada localização, e $z(s_i + h)$ é a observação do ponto vizinho na distância $s_i + h$.

Supondo que haja N observações em pontos, isso produz pares de $N \times (N-1)/2$ para os quais uma semivariância pode ser calculada. Se traçarmos todas as semivariâncias

versus suas distâncias de separação, uma nuvem de variograma será produzida. Para uma visualização mais fácil dessa nuvem de variograma, os valores são geralmente calculados para uma distância padrão chamada *lag*. Se exibirmos esses dados médios, obtemos o variograma experimental padrão, que pode ser visto na Figura 2.

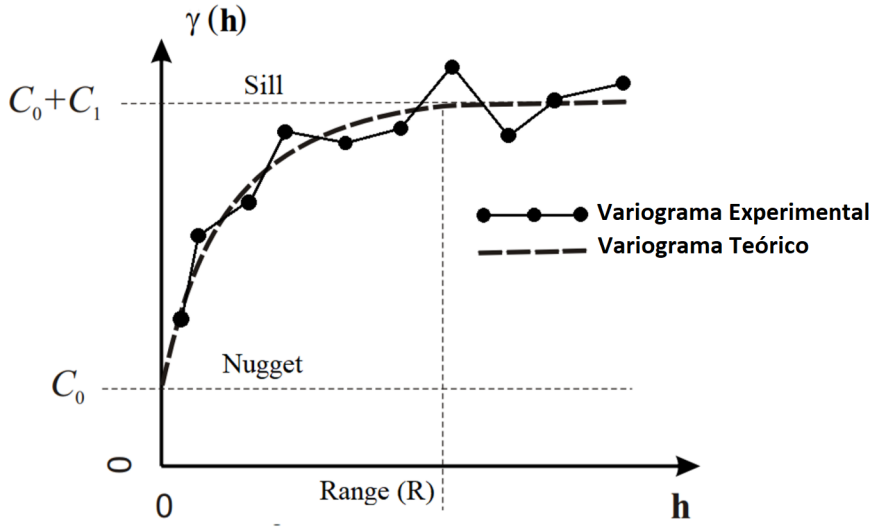

Figura 2. Exemplo de um modelo final de variograma.

Uma vez calculado o variograma experimental, devemos ajustá-lo usando um modelo do variograma teórico como linear, esférico, exponencial, gaussiano, entre outros. Os variogramas são geralmente ajustados usando uma função de custo como os mínimos quadrados ponderados (CRESSIE, 1985) e por 1-*fold cross validation* (LI et al., 2018). Portanto, o objetivo principal é minimizar essa função de custo. Nesta tese, para simplificar os experimentos, foi utilizado o modelo teórico *matern*, que é dado por

$$\gamma(h) = C_0 + C_1 \left\{ 1 - \frac{1}{2^{v-1}\Gamma(v)} \left(\frac{h}{R} \right)^v K_v \left(\frac{h}{R} \right) \right\} \tag{2.3}$$

onde R é o *practical range* ou simplesmente *range* e é igual à distância pela qual $\gamma(h) = 0,95(C_0 + C_1)$ (OLEA, 2012). Os locais de amostragem separados por distâncias mais próximas do que o *range* são automaticamente espacialmente correlacionados, caso contrário, não são; C_0 é o efeito pepita ou *nugget effect*, que pode ser atribuído a erros de medição ou fontes de variação espacial em distâncias menores que o intervalo de amostragem; $C_0 + C_1$ é o *sill*, que é o valor que o modelo atinge no *range* R; v é um parâmetro de suavidade chamado *kappa*; K_v é uma função de Bessel modificada (WEISSTEIN, 2002); e $\Gamma(v)$ é uma função fatorial de números complexos.

Depois de ajustar o modelo do variograma teórico, podemos usá-lo para derivar semivariâncias em todos os locais do espaço geográfico e resolver os pesos de krigagem. Os pesos da krigagem ordinária (OK) são resolvidos multiplicando as covariâncias:

$$\lambda_0 = \mathbf{C^{-1}}.\mathbf{c_0}; \qquad C(|h| = 0) = C_0 + C_1 \tag{2.4}$$

onde $\mathbf{C}$ é a matriz de covariância derivada para $N \times N$ observações e $\mathbf{c_0}$ é o vetor de covariâncias em um novo local. Observe que o $\mathbf{C}$ é de fato uma matriz $(N+1) \times (N+1)$ se for usada para derivar pesos de krigagem, uma vez que uma linha e coluna extra são usadas para garantir que o a soma dos pesos é igual a um:

$$\begin{bmatrix} C_{(s_1,s_1)} & \dots & C_{(s_1,s_N)} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ C_{(s_N,s_1)} & \dots & C_{(s_N,s_N)} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C_{(s_0,s_1)} \\ \vdots \\ C_{(s_0,s_N)} \\ 1 \end{bmatrix} = \begin{bmatrix} w_1(s_0) \\ \vdots \\ w_N(s_0) \\ \varphi \end{bmatrix} \tag{2.5}$$

onde φ é o *multiplicador Lagrange*. Após o cálculo dos pesos, a estimativa é dada pela Equação 2.1

Quando o variograma experimental é distinto para duas ou mais direções, temos um fenômeno anisotrópico (HENGL, 2009), como pode ser visto na elipse desenhada na Figura 3. A elipse representa a área a partir da qual os dados seriam considerados no processo de krigagem. A anisotropia é calculada considerando um certo ângulo de 0 a 180 graus, que representa a direção azimutal no sentido horário na direção principal e um fator dado por

$$Fatordeanisotropia = \frac{a_2}{a_1} \tag{2.6}$$

onde a_1 e a_2 são os raios maior e menor da elipse, respectivamente. Esse fator varia entre 0 e 1, sendo 1 um modelo isotrópico. Assim, no caso da anisotropia, cinco parâmetros são usados para estimar o modelo teórico do variograma: *nugget*, *sill*, *range*, ângulo e fator de anisotropia.

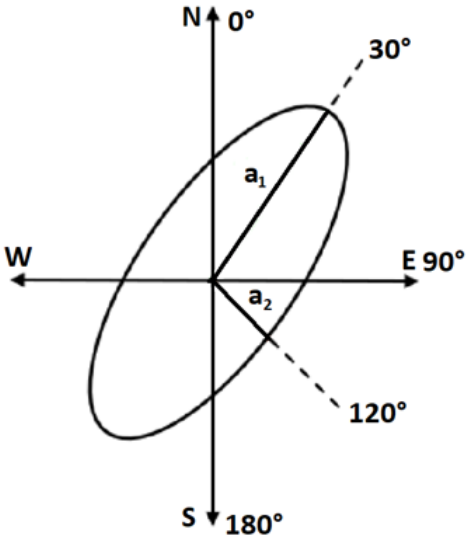

Figura 3. Exemplo de anisotropia.

2.2 Funções de custo

Nos métodos geostatísticos, o passo crítico para a estrutura espacial ou relações dos dados é a modelagem do variograma, o que vem sendo o foco de pesquisas na área por um longo período (LI et al., 2018). Comumente, o método de *Maximum likelihood* (ML) e mínimos quadrados são utilizados para realizar o ajuste do modelo do variograma teórico.

Funções de custo como *Maximum likelihood*, auxiliam na estimativa dos parâmetros do variograma teórico através da minimização de uma função de *log-likelihood* negativa. Uma característica importante desta função é que os parâmetros são diretamente calculados sem a necessidade de etapas intermediárias. No entanto, de acordo com alguns especialistas (LI et al., 2018), o ML é fortemente dependente do modelo escolhido. É assumido que os dados seguem uma distribuição gaussiana multivariada, o que é um requerimento difícil de se cumprir e é praticamente impossível de verificar (KERRY; OLIVER, 2007).

Em contraste, temos a função de custo de mínimos quadrados considerado como o meio mais comum para modelagem do variograma teórico. Dentre suas principais características destaca-se a sua simplicidade e a alta disponibilidade em diferentes plataformas computacionais. Além disso, a aplicação de mínimos quadrados, segundo Oliver e Webster (2014), é satisfatória em 90% dos casos de investigações geoestatísticas.

O processo principal da modelagem do variograma utilizando a função de mínimos quadrados é o ajuste de valores discretos do variograma experimental com função de modelo negativa definida mais aproximada. Para atingir este objetivo, é necessário selecionar um modelo de variograma e buscar seus parâmetros utilizando critérios dos mínimos quadrados.

No trabalho de Li et al. (2018), é demonstrado que é possível a utilização da função de custo por interpolação. Nesta função, é utilizado o método *1-Fold Cross Validation* na base de dados e calculado o erro total obtido a partir de um conjunto de parâmetros do variograma teórico. A partir deste cenário, busca-se minimizar o erro utilizando diferentes conjuntos de parâmetros. Nesta tese, foram comparados métodos utilizando a função de custo de mínimos quadrados e por interpolação.

2.3 Hipótese da estacionariedade

Dado um conjunto de N valores $z(s_i)$, a hipótese será intrínseca (ou estacionária) se ela seguir duas condições. A primeira condição requer que o valor esperado $E\{z(s_i)\}$ exista e não dependa da posição s_i. Essa condição pode ser matematicamente formulada por

$$E\{z(s_i)\} = m \tag{2.7}$$

Com relação a segunda condição, para a hipótese ser intrínseca, a variância do incremento $[z(s_i) - z(s_i + h)]$ é finito e não depende da posição s_i. Isto pode ser definido como

$$VAR[z(s_i) - z(s_i + h)] = E[z(s_i) - z(s_i + h)]^2 \tag{2.8}$$

Trends podem ser considerados como padrões ou ruídos nos dados, usualmente expressos como uma inclinação para cima ou para baixo dos valores dos dados através do tempo e do espaço. É importante ressaltar que as duas condições expressas nas Equações 2.7 e 2.8 não são satisfeitas na presença de *trends*. Isto se dá pelo fato de que o valor médio m dependeria da posição s_i e $VAR[z(s_i) - z(s_i + h)]$ seria infinito e também dependeria da posição no espaço (VIEIRA et al., 2010).

A tarefa de remoção de *trends*, também chamado de *detrending*, pode ser realizada através do ajuste de uma superfície de *trend* para os valores da base de dados, por exemplo, através do método de mínimos quadrados, e então subtrair um valor da função da superfície de *trend* ajustada, dos valores originais, obtendo-se uma nova variável residual (VIEIRA et al., 2010) que será utilizada nas fases posteriores da krigagem.

O processo de *detrending* das variáveis regionalizadas em estudo é um requerimento para que valores não-estacionários estejam sob a hipótese estacionária, isto é, quando aplicados a técnicas geoestatísticas estacionárias. Após o *detrending*, a análise segue seu percurso normal através do cálculo do variograma experimental, ajuste de um modelo do variograma teórico e a realização das estimativas da krigagem. Ao final do processo, é possível utilizar a mesma função de superfície de *trend* ajustada para reverter os valores para seu estado original.

2.4 K-vizinhos mais próximos

K-vizinhos mais próximos ou *K-Nearest Neighbor* (KNN) (WITTEN et al., 2016) é um algoritmo de aprendizado de máquina supervisionado utilizado tanto para problemas de classificação quanto para problemas de regressão. É um método não-paramétrico que utiliza diretamente os dados de treinamento para a classificação. Mais especificamente, o algoritmo KNN classifica um novo ponto baseado na sua vizinhança, ou seja, os pontos mais próximos à ele.

Dado um conjunto de treinamento $\{(\mathbf{s}_1, q_1), \ldots, (\mathbf{s}_N, q_N)\}$, com N pontos, cada ponto $(\mathbf{s}, q)$ consiste em um vetor $\mathbf{s} \in \mathbb{R}^L$ e um rótulo (ou classe) $q \in \{1,\ldots,Q\}$. Seja $\mathbf{s}_0 = (p_1,\ldots,p_L)$ um novo ponto ainda não classificado (sem rótulo ou classe), para classificar este novo ponto, o algoritmo KNN calcula a distância entre $\mathbf{s}_0$ e todos os outros pontos na base de treinamento através de uma medida de similaridade. Os K pontos mais próximos, isto é, que possuem a menor distância em relação $\mathbf{s}_0$ são então armazenadas. Na sequência,

é verificado qual o rótulo mais frequente entre os K vizinhos armazenados, e o rótulo eleito q é associado ao novo ponto em questão.

Uma das medidas de similaridade amplamente utilizadas na literatura, e que foi aplicada nos experimentos, é a distância euclidiana, que pode ser definida por:

$$d(\mathbf{s}, \hat{\mathbf{s}}) = \sqrt{\sum_{i=1}^{L}(p_i - \hat{p}_i)^2} \tag{2.9}$$

onde p_i e $\hat{p}_i$ são os elementos dos vetores $\mathbf{s}$ e $\hat{\mathbf{s}}$, respectivamente.

Abaixo podemos visualizar um algoritmo geral para a execução do KNN:

Algorithm 1: Algoritmo KNN

1 - Recebe um dado não classificado;
2 - Mede a distância (Euclidiana, Manhattan, Minkowski ou Ponderada) do novo dado com todos os outros dados que já estão classificados;
3 - Obtém as menores distâncias;
4 - Verifica a classe de cada da um dos dados que tiveram a menor distância e conta a quantidade de cada classe que aparece;
5 - Toma como resultado a classe que mais apareceu dentre os dados que tiveram as menores distâncias;
6 - Classifica o novo dado com a classe tomada como resultado da classificação;

Na Figura 4 pode-se observar um exemplo de classificação utilizando o algoritmo KNN. O círculo vermelho indica o dado a ser classificado. Supondo uma classificação considerando 3 vizinhos mais próximos, isto é, $k = 3$, a classe atribuída ao novo ponto (círculo) seria a classe B (quadrados). Caso fosse considerado 6 vizinhos mais próximos, isto é, $k = 6$, a classe atribuída ao novo ponto seria a classe A (triângulos).

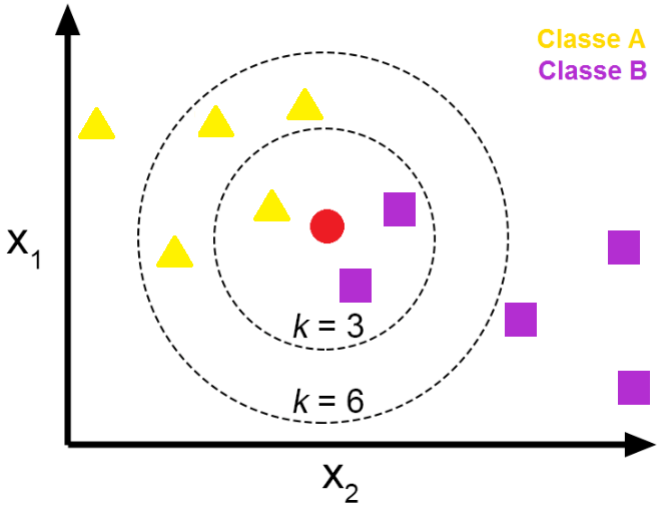

Figura 4. Exemplo de classificação utilizando o algoritmo KNN (Adaptado de JOSÉ, 2017).

2.5 k-means

O k-means é um algoritmo não supervisionado utilizado para problemas de agrupamento(clusterização). Esta técnica, bastante conhecida e utilizada pela sua simplicidade de implementação, realiza a partição de um conjunto de dados em U *clusters* (grupos), onde o valor de U é fornecido pelo usuário (WITTEN et al., 2016).

Dado um conjunto de observações $\{s_1, s_2, \ldots, s_N\}$, o algoritmo busca particionar as N observações em U grupos $\mathbf{u} = \{u_1, u_2, \ldots, u_N\}$ com o objetivo de minimizar a soma dos quadrados inter-clusters (variância).

Formalmente, o objetivo é encontrar:

$$\arg_u \min \sum_{i=1}^{U} \sum_{p \in u_i} \|p - \mu_i\|^2 = \arg_u \min \sum_{i=1}^{U} |u_i| \, Varu_i \tag{2.10}$$

onde μ_i é a média dos pontos em u_i.

O algoritmo k-means inicia seu processo pela inicialização de um conjunto de U centroides, um para cada *cluster*. Há diversas formas de inicialização para seleção dos centroides, umas das mais utilizadas é a seleção aleatória entre os pontos da base de dados. Após esta etapa, cada ponto é associado ao centroide mais próximo com base em uma medida de similaridade para formar os agrupamentos. Nesta tese foi utilizada a distância euclidiana(Equação 2.9). Posteriormente, os centroides são recalculados. O novo centroide é a média dos antigos pontos do *cluster*. Todo este processo é repetido até que nenhum centroide mude de posição ao final uma iteração, ou seja, que a função objetivo (Equação. 2.10) não se altere.

2.6 Algoritmos genéticos

Algoritmos genéticos (AG) são um grupo de métodos de otimização inspirados pela teoria da evolução natural biológica. Este algoritmo é geralmente composto por uma população de indivíduos que representam as soluções candidatas. Em outras palavras, cada indivíduo é uma possível solução para o problema em questão e através de mecanismos da natureza, evolui em busca da melhor solução (GOLDBERG; HOLLAND, 1988).

Um algoritmo genético é caracterizado por possuir as seguintes etapas fundamentais (Figura 5):

- Inicialização da população;
- Seleção;
- Cruzamento;

- Mutação;
- Reposição (Nova população);
- Avaliação dos Indivíduos.

Figura 5. Ciclo de execução de um algoritmo genético.

O processo evolucionário inicia com a etapa de *inicialização*, onde uma população inicial de candidatos é gerada. Existem vários métodos de inicialização de populações, mas em algoritmos genéticos é comum que ela seja gerada aleatoriamente. A partir do momento que a ela é gerada, dá início ao ciclo evolucionário.

Após a geração da população, cada indivíduo vai ser testado empiricamente em uma função e vai receber um valor de aptidão chamado *fitness*. Este valor geralmente é dado por uma função objetivo $f(x)$ que se deseja otimizar com o AG. Dependendo do problema a ser otimizado, o melhor valor pode ser o maior para casos de maximização, ou o menor para casos de minimização.

A cada geração (ou ciclo), um conjunto de pais é selecionado da população, por meio de um mecanismo de seleção, com base nos seus valores de *fitness*. Neste caso, indivíduos mais adaptados à solução do problema possuem maiores chances de serem selecionados para gerar os futuros indivíduos. Alguns dos principais métodos de seleção são:

- Roleta: Este método cria uma roleta em formato de "pizza", onde cada fatia representa a porcentagem de escolha de determinado indivíduo. Quanto melhor o *fitness*, maior a fatia correspondente na roleta. Este método não é muito efetivo quando a diferença de *fitness* dos indivíduos é muito alta, visto que *fitness* muito altos, tendem a pegar

uma fatia muito grande, praticamente anulando as chances de seleção das fatias menores. Este é método de seleção utilizado nos experimentos desta tese.

- *Torneio*: o método de torneio cria pequenos subgrupos de indivíduos selecionados aleatoriamente, dentro destes subgrupos é realizado uma disputa, onde indivíduos com melhor *fitness* tem mais chance de ganhar. A chance e o número de indivíduos no subgrupo são definidos na implementação do AG.

Após a seleção dos pais, temos a etapa de *crossover* ou cruzamento. Nesta etapa, dois pais (soluções) têm seus códigos genéticos combinados com o objetivo de criar novas soluções ou filhos com seu código genético, em busca de uma melhor solução para o problema. Há diversos métodos de cruzamento atualmente na literatura, como exemplo, pode-se citar os cruzamentos de 1 ponto de corte e 2 pontos de corte (GOLDBERG; HOLLAND, 1988).

Cada indivíduo gerado na etapa de cruzamento possui a chance de sofrer mutação. O processo de mutação se dá pela alteração de um valor do código genético daquele indivíduo.

Nesta tese, foi utilizado o processo de cruzamento e mutação de Laplace (DEEP; THAKUR, 2007b), que podem ser utilizados quando são utilizados valores reais nas variáveis do código genético.

Ao gerar toda a próxima geração, após as etapas de cruzamento e mutação, que também pode ser chamada de etapa de *replacement* ou substituição, os novos indivíduos são novamente avaliados através da função de *fitness* e uma novo geração é iniciada.

As condições de término do algoritmo geralmente são atingidas quando o AG apresenta uma das situações:

- Atinge o resultado ótimo do problema (Somente em casos onde se sabe previamente o ótimo para aquele problema);
- Executou o número máximo de gerações definidas;
- Ficou preso em algum ótimo local por número determinado de gerações (Ou seja, o melhor resultado encontrado não foi alterado por um determinado número de gerações).

2.7 Enxame de partículas

O Enxame de partículas ou *Particle swarm optimization* (PSO) usa uma estrutura semelhante ao dos algoritmos genéticos para avaliar cada indivíduo (também chamado de partículas). Todavia, em vez de usar métodos de cruzamento, mutação e seleção, novas

soluções são criadas ao "mover" a partícula pelo espaço de busca com base em dois fatores: i) A melhor solução alcançada pela partícula até agora; ii) A melhor solução alcançada por qualquer partícula na vizinhança. A estrutura de uma partícula consiste em quatro parâmetros, localização atual (idêntica ao cromossomo do AG), velocidade atual, melhor condicionamento local e melhor condicionamento global (MISHRA; TIWARI; MISRA, 2011).

De acordo com Poli, Kennedy e Blackwell (2007) no enxame de partículas, um número de entidades, chamadas de partículas, são inseridas no espaço de busca de um problema ou função, e cada uma dessas partículas avalia a função objetivo na sua atual localidade no espaço de busca. Cada partícula, então, determina seu movimento dentro do espaço de busca por meio da combinação de um histórico do seu atual e melhor *fitness* encontrado, com uma ou mais partículas dentro do enxame, com algumas perturbações aleatórias. A próxima iteração se dá após todas as partículas terem se movimentado.

Cada partícula do enxame é composta de três vetores L-dimensional, onde L é a dimensionalidade do espaço de busca. Os vetores são a posição atual $\overrightarrow{x_i}$, a melhor posição anterior $\overrightarrow{p_i}$, e a velocidade $\overrightarrow{v_i}$ (POLI; KENNEDY; BLACKWELL, 2007).

O posição atual $\overrightarrow{x_i}$, pode ser considerada com um conjunto de coordenadas descrevendo um ponto no espaço. A cada iteração do Algoritmo 2 (PSO), a posição atual é avaliada com uma solução do problema. Se esta posição é melhor que qualquer uma encontrada até o momento, então estas coordenadas são armazenadas do segundo vetor $\overrightarrow{p_i}$. O valor do melhor resultado geral é armazenado em uma variável que podemos chamar de $pbest_i$, para comparação em iterações posteriores. O objetivo é a contínua procura por melhores posições e atualizando $\overrightarrow{p_i}$ e $pbest_i$. Novos pontos ou movimentos são realizados através da adição das coordenadas de $\overrightarrow{v_i}$ para $\overrightarrow{x_i}$ (POLI; KENNEDY; BLACKWELL, 2007).

Algorithm 2: PSO Original

Inicializar uma população de partículas com posições e velocidades aleatórias em L dimensões no espaço de busca;

for *1 to Nº Iterações ou Critério Atingido* **do**

1 - Para cada partícula, calcular a função fitness com base nas L variáveis;

2 - Comparar a avaliação do fitness de cada partícula com o seu melhor pessoal $pbest_i$. Se o valor calculado for maior, então atualizar $pbest_i$ e $\overrightarrow{p_i}$ com o valor atual referente a localização $\overrightarrow{x_i}$ no espaço de busca;

3 - Identificar a partícula na vizinhança com o melhor resultado até o momento, e atribuir o seu índice a variável g;

4 - Atualizar a velocidade e a posição de cada partícula;

end

.

Considerando um enxame com H partículas, temos o vetor de posições $X_i^t = (x_{i1}, x_{i2}, \ldots, x_{iL})^{ITE}$ e um vetor de velocidades $V_i^t = (v_{i1}, v_{i2}, \ldots, v_{iL})^{ITE}$ na iteração *ite* para cada partícula i que a compõe. Os vetores são atualizados através da dimensão j de acordo com A atualização de velocidade de se dá pela equação:

$$V_{ij}^{t+1} = \omega V_{ij}^t + c_1 r_1^t (pbest_{ij} - X_{ij}^t) + c_2 r_2^t (gbest_j - X_{ij}^t) \tag{2.11}$$

e

$$X_{ij}^{t+1} = X_{ij}^t + V_{ij}^{t+1} \tag{2.12}$$

onde $i = 1, 2, \ldots, H$ e $j = 1, 2, \ldots, L$. ω é a constante de peso da inércia, r_1 e r_2 são valores aleatoriamente escolhidos entre 0 e 1 e c_1 e c_2 são os pesos cognitivos e social respectivamente.

2.8 Métrica de avaliação

A função de *fitness* utilizada no algoritmo genético foi obtida através da aplicação da krigagem em cada ponto da base de dados de treinamento (cerca de 90% da base de dados), no que é chamado de validação cruzada *leave-one-out*.

Em relação à métrica utilizada para comparar as técnicas avaliadas, foi utilizada a validação cruzada *10-fold* para obtenção do erro médio quadrático normalizado, ou *normalized mean squared error (NMSE)*. Para ambos os casos, cálculo do *fitness* e avaliação de cada técnica, foi utilizada a função de custo por interpolação (Equação 2.13) (LI et al., 2018). O NMSE é calculado através da equação

$$NMSE_u = \frac{1}{\sigma^2.n} \sum_{i=1}^{n} [z_{\hat{OK}}(s_i) - z_{OK}(s_i)]^2 \tag{2.13}$$

onde $z_{\hat{OK}}(s_i)$ é o valor estimado da variável objetivo obtido pela krigagem ordinária no ponto s_i; n é o número total de pontos no *cluster* u; e σ^2 é a variância da variável objetivo considerando os dados do *cluster* u. Quanto menor o valor NMSE melhores são os valores estimados pelo modelo do variograma teórico ajustado. Portanto, o índice NMSE da base de dados como um todo é dado por

$$NMSE = \sum_{u=1}^{U} NMSE_u \tag{2.14}$$

onde U é o número total de clusters, argumento este fornecido pelo usuário.

É importante ressaltar que a validação cruzada *leave-one-out* foi utilizada somente para calcular a função de *fitness* do algoritmo genético e do enxame de partículas. Para medir a acurácia, foi aplicado a validação cruzada *10-fold*. Diante disso, as bases de dados estudadas foram particionadas aleatoriamente em 90% para treinamento e 10% para teste

em cada iteração. No total, 10 iterações com diferentes particionamentos foram obtidos para cada quantidade de *clusters*. A média dessas 10 iterações foi calculada no final e comparada entre as técnicas.

2.9 Índice de diversidade da população

A diversidade padrão da população, também chamada de *standard population diversity (SPD)*, descreve o nível de variação alcançado por uma determinada população. Um grande índice de diversidade SPD indica, geralmente, que há maior variabilidade dentro da população (YASOJIMA et al., 2019b). Esta diversidade pode ser importante para que mais regiões do espaço de busca possam ser varridas, no entanto, é importante verificar o comportamento para cada problema específico.

Considerando uma população com P indivíduos (G_1, G_2,...,G_p), dado que cada indivíduo, ou partícula no caso do enxame de partículas, possuem T parâmetros, pode-se definir que $G_i = (G_{i,1}, G_{i,2}, \ldots, G_{i,T})$. Assim, a média geral de cada gene/variável do indivíduo T é dado por

$$G_T^{ave} = \frac{1}{P}\sum_{i=1}^{P} G_{i,T} \tag{2.15}$$

Na etapa de normalização, o desvio padrão de cada gene T em relação à população P é calculado por

$$\sigma(G_T^{ave}) = \sqrt{\frac{1}{P}\sum_{i=1}^{P}(G_{i,T} - G_T^{ave})^2} \tag{2.16}$$

Por fim, a variabilidade da população P em cada geração do algoritmo bioinspirado é dado por

$$SPD = \frac{1}{T}\sum_{j=1}^{T}(\frac{\sigma(G_j^{ave})}{G_j^{ave}}) \tag{2.17}$$

2.10 Testes Estatísticos

2.10.1 Teste de Friedman

Para o teste de Friedman, podemos considerar uma tabela onde os dados se dispõem em uma tabela com g linhas e k colunas, como pode ser observado na Tabela 1. As linhas representam os vários indivíduos/classes, e as colunas representam as diversas condições. Se estão sendo estudados os *scores* de indivíduos observados sob todas as condições, então cada linha dá os *scores* de um indivíduo sob as k condições (VIALI, 2017).

Tabela 1. *Scores* de três grupos correspondentes sob quatro condições.

	Condições			
	I	**II**	**III**	**IV**
Grupo A	**9**	4	1	7
Grupo B	**6**	5	2	8
Grupo C	**9**	1	2	6

Se a hipótese de nulidade é verdadeira, então a distribuição de postos em cada coluna será aleatória, aparecendo nas colunas com frequências aproximadamente iguais. Portanto, para qualquer grupo, é uma questão de acaso sob que condição ocorre o menor *score*, o que seria o caso se as condições realmente não diferissem entre si. Se H_0 fosse falsa, então os totais de condições iriam variar de uma coluna para outra significativamente (VIALI, 2017).

O teste de Friedman determina se as condições diferem significativamente. Obtemos o valor do teste pela equação.

$$\chi^2 = \frac{12}{gk(k+1)} \sum_{i=1}^{2} R_j^2 - 3g(k+1) \tag{2.18}$$

onde,

- g é o número de linhas;
- k número de colunas;
- R_j é a soma dos postos da coluna j.

2.10.2 Teste de Mann-Kendall

O teste não-paramétrico de Mann-Kendall é comumente utilizado para a detecção de trends. A hipótese nula, H_0, é que os dados são originados de uma distribuição de realizações independentes e são distribuídas identicamente, já a hipótese alternativa, H_1, é que os dados seguem um trend monotônico (MARÔCO, 2018). O teste estatístico de Mann-Kendall é calculado de acordo com a equação (MARÔCO, 2018):

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} sgn(x_j - x_e) \tag{2.19}$$

com

$$sgn(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases} \tag{2.20}$$

A média de S é $E[S] = 0$ e a variância σ^2 é

$$\sigma^2 = \left\{ n(n-1)(2n+5) - \sum_{j=1}^{o} t_j(t_j-1)(2t_j+5) \right\} /18 \tag{2.21}$$

onde o é o número de agrupamentos no conjunto de dados e t_j é o número de pontos no jth agrupamento.

A estatística S é relacionada ao T de Kendall e é dado por:

$$T = \frac{S}{D} \tag{2.22}$$

onde

$$D = \left[\frac{1}{2}n(n-1) - \frac{1}{2}\sum_{j=1}^{o} t_j(t_j-1) \right]^{1/2} \left[\frac{1}{2}n(n-1) \right]^{1/2} \tag{2.23}$$

2.10.3 Teste de Shapiro Wilk

O teste de Shapiro-Wilk verifica a normalidade de uma população de dados (ACTION, 2017).

O teste é aplicado através da Equação:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.24}$$

em que x_i são os valores da amostra ordenados (x_1 é o menor). A constante b é determinada da seguinte forma:

$$b = \begin{cases} \sum_{i=1}^{\frac{n}{2}} a_{n-i+1} x (x_{n-1+1} - x_i) & \text{se} \quad n \quad \text{é par} \\ \sum_{i=1}^{\frac{n+1}{2}} a_{n-i+1} x (x_{n-1+1} - x_i) & \text{se} \quad n \quad \text{é ímpar} \end{cases} \tag{2.25}$$

em que a_{n-i+1} são constantes geradas pelas médias, variâncias e covariâncias das estatísticas de ordem de uma amostra de tamanho n (ACTION, 2017).

Para a aplicação do teste Shapiro-Wilk, temos as seguintes prerrogativas (MARÔCO, 2018):

1. A hipótese nula H_0 de que a amostra provem de uma população normal e a hipótese H_1 de que a amostra não provém de uma população normal;
2. Estabelecer o nível de significância do teste α, normalmente 0,05;
3. (I). Ordenar as n observações da amostra; (II). Calcular $\sum_{i=1}^{n}(x_i - \bar{x})^2$; (III). Calcular b; (IV). Calcular W;
4. Rejeitar H_0 ao nível de significância α se $W_{\text{calculado}} < W_\alpha$

2.10.4 Teste T Pareado e One-Way Anova para medidas Repetidas

O teste T pareado realiza a comparação de médias de duas populações, envolvendo a coleta de observações em pares, de modo que os dois elementos de cada par sejam homogêneos em todos os sentidos, exceto em relação ao fator que se deseja comparar (MORETTIN; BUSSAB, 2017)

Pode-se também realizar o estudo no mesmo indivíduo para as duas amostras, ou seja, medir a característica do indivíduo antes e depois dele ser submetido a um tratamento. Assim, se houver diferença no estudo, há indícios de que tal diferença seja resultante do tratamento utilizado (MORETTIN; BUSSAB, 2017).

Nesta situação, podemos considerar duas amostras $x_1, x_2, \ldots, x_n$ e $y_1, y_2, \ldots, y_n$ onde x_i e y_i são pareadas, isto é, pode-se considerar uma amostra de pares $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Ao definir a variável aleatória $d = x - y$, tem-se a amostra $d_1, d_2, \ldots, d_n$, resultante das diferenças entre os valores de cada par.

O teste t pareado supõe que $d \sim N(\mu_d, \sigma_d^2)$ As hipóteses do teste são:

$$\begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \end{cases} \tag{2.26}$$

A estatística do teste t pareado pode ser descrita por

$$T_{\text{pareado}} = \frac{\bar{d}}{\frac{S_d}{\sqrt{n}}} \sim t_{n-1}, \tag{2.27}$$

que segue uma distribuição t de Student com $(n-1)$ graus de liberdade. Neste cenário, $S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$ é o estimador da variância da variável aleatória d e $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(x_i, y_i) = \bar{x} - \bar{y}$ é o estimador da média da diferença das duas populações.

O valor-p ou *p-value* é definido como a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, assumindo-se como verdadeira a hipótese nula. O valor-p é dado por

$$\text{valor} - p = 2P\left[t_{n-1} > |T_{\text{pareado}}|\right]. \tag{2.28}$$

Se o valor-p for menor que o nível de significância (α) estabelecido, rejeita-se a hipótese nula. Caso contrário, não rejeita-se H_0, não sendo possível afirmar que as duas médias sejam diferentes.

A técnica *One-Way* Anova para medidas repetidas é equivalente ao teste T pareado, no entanto, para três ou mais grupos. Podemos definir os passos desta técnica da seguinte forma:

1. Primeiro, a ANOVA calcula a média para cada um dos grupos;

2. Então ela calcula a média geral (as médias são somadas e divididas pelo número de grupos, nesse caso três);

3. Para cada grupo separadamente, a variação total de cada participante em relação à média do grupo é calculada. Essa é a variância dentro dos grupos (*within-group*). Esse cálculo é feito a partir da soma dos quadrados;

4. A variação da média de cada grupo em relação à média geral é calculada. Essa é a variância entre os grupos (*between-groups*).

A rejeição da hipótese nula em favor da hipótese alternativa indica que essas diferenças se devem ao fator (ou fatores), que surtiram efeitos estatisticamente significativos sobre os resultados.

No entanto, esta técnica identifica que há diferenças estatisticamente significativas entre as populações. Para identificar quais populações são diferentes, é necessário a aplicação de uma técnica post-hoc como o teste de Bonferroni.

2.10.5 Teste de Bonferroni

O teste F (Bonferroni) é usado para comparar variâncias que é a base para a ANOVA que, por sua vez, é a técnica usada para comparar se as médias de 3 ou mais grupos são diferentes. A ideia do teste F é usar o resultado para determinar se rejeita ou não a hipótese nula com base na distribuição F, em outras palavras, indicar se as diferenças entre os pares de grupos são significantes (ACTION, 2017).

Esta técnica é utilizada em conjunto com o *One-Way* Anova para medidas repetidas como técnica post-hoc para dados paramétricos, ou seja, que possuem dados em distribuição normal.

3 Trabalhos relacionados

Esta seção detalha a revisão bibliográfica acerca do tema de pesquisa: Krigagem. Muitos trabalhos tem procurado aperfeiçoar a técnica de Krigagem, utilizando ferramentas e conceitos de diversas áreas do conhecimento. Desta forma, faz-se necessário analisar os resultados destes trabalhos, verificando suas metodologias e práticas, expondo vantagens e desvantagens, e identificando possíveis lacunas, onde há a possibilidade de propor novas melhorias, contribuindo para o estado da arte da área.

3.1 Revisão sistemática

Uma revisão sistemática é uma revisão orientada por um protocolo que sintetiza estudos concentrado-se em um tema ou questões chaves dadas (RUSSELL et al., 2009). O objetivo principal é, através de palavras chaves, identificar temas mais estudados, desafios, lacunas, metodologias, técnicas e resultados relevantes acerca do tema abordado. O guia de revisão sistemática de Kitchenham (KITCHENHAM, 2004), em uma versão simplificada, foi utilizada como metodologia de revisão da literatura nesta tese. Mais especificamente, a Figura 6 apresenta o fluxo das atividades da revisão da literatura realizada.

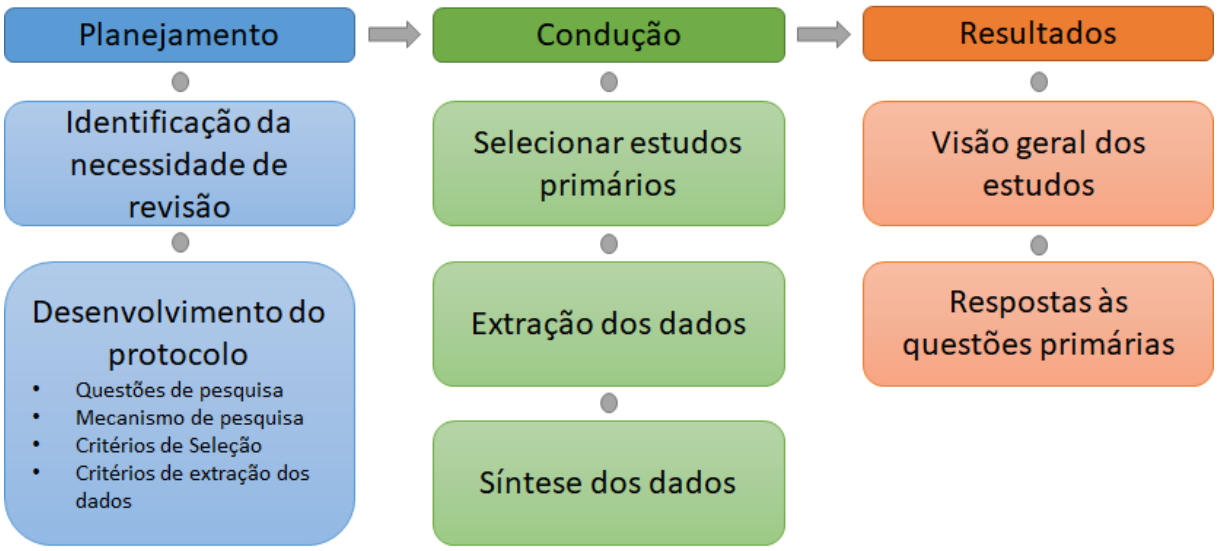

Figura 6. Protocolo da revisão sistemática adaptada de Kitchenham, 2004.

3.2 Atividade: Planejamento

Engloba atividades para identificar qual a necessidade de revisão, e o desenvolvimento do protocolo de revisão.

3.2.1 Identificação da necessidade de revisão

O processo de krigagem é muito utilizado na área da geoestatística. Mais especificamente, na áreas de mineração, estudo de solo, geologia e ciências ambientais em geral. No entanto, nestes exemplos, o conhecimento especialista é indispensável para a modelagem de parâmetros importantes da técnica. Neste sentido, busca-se comprovar a hipótese da importância de aprimoramentos do processo de krigagem, aperfeiçoando e automatizando etapas da sua aplicação, utilizando principalmente técnicas de inteligência artificial.

3.2.2 Desenvolvimento do protocolo

As atividades executadas no desenvolvimento do protocolo da revisão da literatura deste trabalho estão descritas a seguir:

1. **Questões de Pesquisa** - Estas questões deverão guiar a leitura dos artigos e deverão ser respondidas ao final da revisão.

Questão principal: Qual o estado da arte no aprimoramento do processo de krigagem utilizando técnicas de inteligência artificial?

Questão Secundária 1 (QS1): Quais as metodologias e técnicas utilizadas na otimização/seleção de parâmetros do variograma?

Questão Secundária 2 (QS2): Em que etapa da krigagem estão sendo utilizadas técnicas de inteligência artificial para aprimoramento do processo?

Questão Secundária 3 (QS3): Qual função de custo está sendo utilizada para a otimização dos parâmetros do semivariograma?

Questão Secundária 4 (QS4): Qual a natureza dos dados em que está se aplicando a krigagem? Dados espaciais ou não espaciais?

2. **Mecanismo de pesquisa**: Inicialmente as palavras usadas na busca de trabalhos foram: “Kriging”. De forma à aprofundar a pesquisa, foram utilizados sub-filtros para melhor seleção dos artigos científicos como:“variogram”, “Optimization” e “Genetic Algorithms”. A busca procurou qualquer parte do trabalho que contivesse as palavras usando as ferramentas dos indexadores de trabalhos científicos on-line, *Science Direct* [1] e *IEEExplore* [2].

[1] <http://sciencedirect.com/>
[2] <http://ieeexplore.ieee.org/>

3. **Critérios de seleção**: Para a seleção de trabalhos mais atuais e relevantes para o tema, foram selecionados trabalhos até 10 anos para trás da data de aplicação da revisão bibliográfica, portanto, entre 2008 e 2019.

4. **Critérios de extração dos dados**: Após a aplicação da busca nos indexadores de artigos científicos especificados no item 2, é realizada uma verificação inicial de cada trabalho consistindo na análise do título e resumo. Com essa premissa, filtra-se trabalhos mais alinhados com o objetivo da pesquisa, e estes, são analisados mais profundamente, observando a metodologia e modelos propostos, área de concentração, técnicas, resultados obtidos e conclusões.

3.2.3 Condução

A aplicação da revisão bibliográfica foi realizado durante o curso da pesquisa, iniciando em 2016 e perdurando até o fim de 2019 com revisões periódicas a cada 4 meses.

3.2.4 Selecionar estudos primários

Para a seleção dos artigos, foram realizados pesquisas com palavras-chave selecionadas, nos repositórios de artigos da *Science Direct* e *IEEExplore*. Do total de artigos obtidos, foi realizada uma segunda filtragem através da leitura dos títulos e resumos. Desse total foram retirados estudos não alinhados com o tema de Krigagem. Os resultados desta etapa podem ser observados na seção 3.3 de *Resultados* da pesquisa bibliográfica.

3.2.5 Extração dos dados

Na etapa de extração dos dados, é realizado uma leitura mais criteriosa dos artigos selecionados na etapa anterior. Em seguida, é preenchido um formulário com as seguintes informações: Autores, ano de publicação, técnicas, modelos automáticos de otimização do variograma, estrutura da base de dados, função de custo utilizada, validação dos resultados.

3.2.6 Síntese dos dados

Neste seção, são montados gráficos e tabelas para a apresentação final do dados extraídos, discutindo problemas comuns encontrados, desafios em aberto e resultados importantes e relevantes.

3.3 Resultados

Esta seção apresenta os resultados de forma geral, responde às questões principal e secundárias e discute os trabalhos relacionados, identificando avanços, lacunas e desafios em geral.

3.3.1 Visão geral dos estudos

Inicialmente, foi aplicado somente a palavra chave “Kriging” para verificação do quantitativo de pesquisas publicadas na grande área do tema.

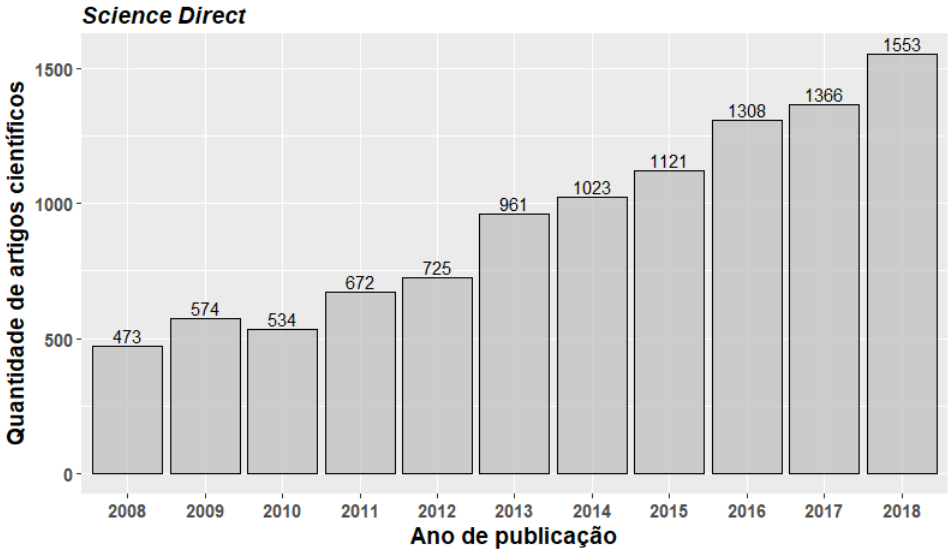

Figura 7. Artigos científicos publicados no *Science Direct* - Palavra chave: Kriging.

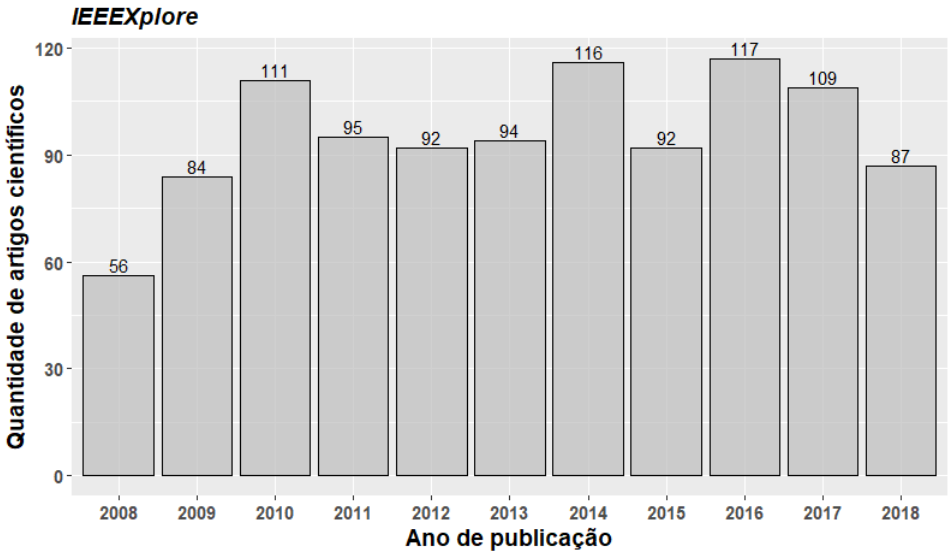

Figura 8. Artigos científicos publicados no IEEExplore - Palavra chave: Kriging.

O total de pesquisas relacionadas, encontrado no repositório de artigos *Science Direct*, foi 15901 entre 2008 e 2019. É possível perceber, na Figura 7, que há um crescimento sistemático no número de publicações ao longo dos anos, demonstrando a importância e relevância do assunto.

No repositório do IEEExplore, o total foi de 1014 artigos publicados no período, e o quantitativo de publicações anuais permaneceu praticamente constante como pode ser observado na Figura 8.

Para se obter um crivo mais alinhado ao objetivo dessa pesquisa, foram utilizados os sub-filtros *Semivariogram*, *Variogram* e *Optimization*, e *Semivariogram*, *Variogram* e *Artifical Intelligence*.

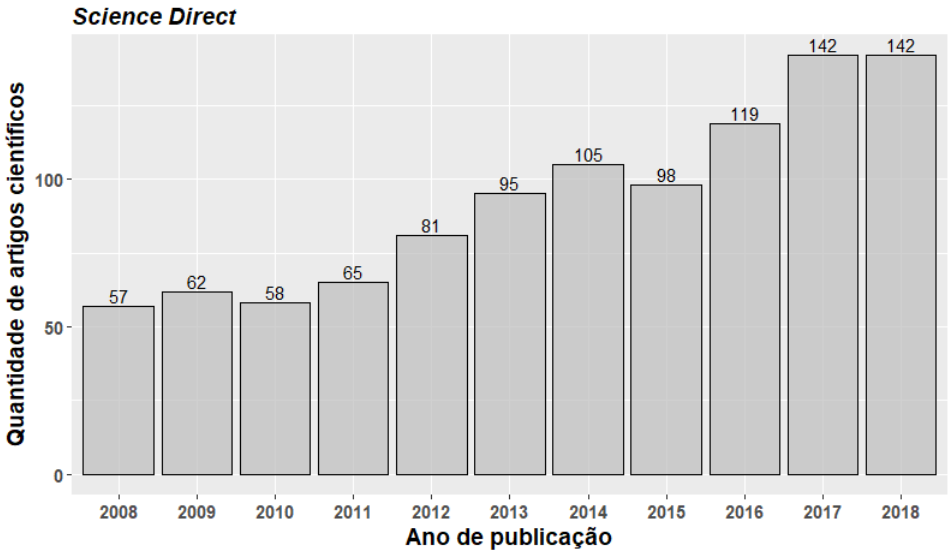

Figura 9. Artigos científicos publicados no Science Direct - Palavra chave: Kriging + Variogram + Optimization.

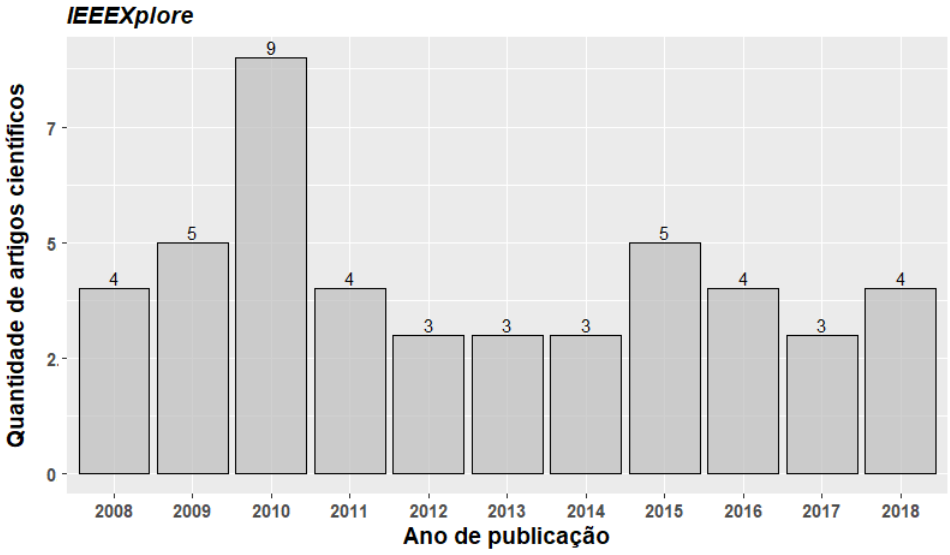

Figura 10. Artigos científicos publicados no IEEExplore - Palavra chave: Kriging + Variogram + Optimization.

Aplicado-se os filtros, foram obtidos resultados mais alinhados com o objetivo da pesquisa. No repositório do sciencedirect foram encontrados um total de 1024 artigos científicos. No ieeexplore foram encontrados 63 artigos. O resultado de cada ano entre 2008 e 2018 com os filtros podem ser visualizados nas Figuras 9 e 10.

Após a leitura dos títulos e resumos de cada trabalho resultante da segunda filtragem exibida nas Figuras 9 e 10, foram retirados estudos repetidos, temas não alinhados, estudos em áreas muito distintas da área computacional e geoestatística, e dado preferência para modelos que utilizaram técnicas de inteligência artificial e algoritmos bioinspirados. No total, passaram para a etapa de extração de dados, 16 artigos, sendo 3 do repositório do *IEEEplore* e 13 do repositório *ScienceDirect*. Destes artigos, 10 foram publicados em periódicos e 6 em conferências.

A leitura dos artigos possibilitou extrair a área de pesquisa em que os 16 artigos levantados atuou. Na Figura 11, é possível perceber que a maioria dos artigos teve uma abrangência geral, sem uma área de atuação definida. Destacam-se a hidrologia e a engenharia como áreas com maior número de artigos publicados após a aplicação dos filtros e seleção de artigos.

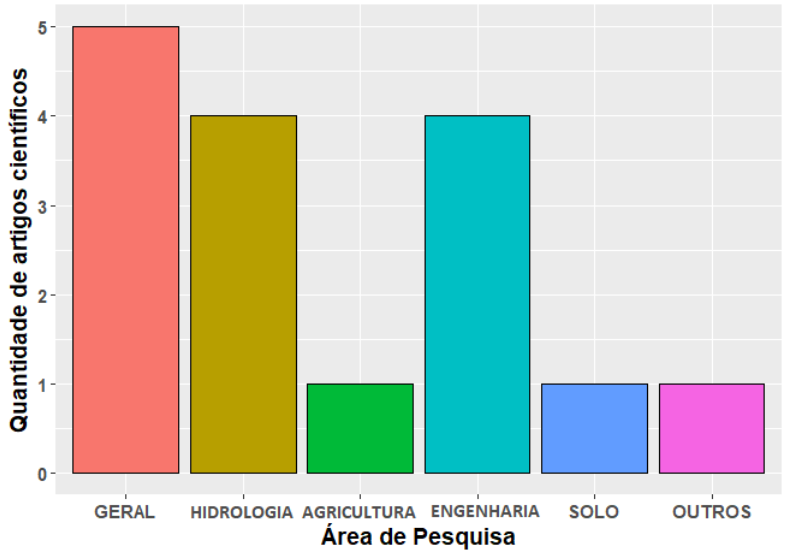

Figura 11. Áreas de pesquisa objetivo de cada publicação.

Na tabela 2 é demonstrado um comparativo entre pontos chave sobre as características de cada modelo apresentado. Há 6 colunas identificando os seguintes pontos:

- **Krigagem**: Indica qual o tipo de krigagem aplicada na pesquisa;
- **Clustering**: Se há aplicação de técnicas de agrupamento;
- **Vario. Param.**: Se há otimização das variáveis do variograma teórico/experimental ou se foi definido manualmente;

- **Dados Espaciais**: Se foram utilizados dados espaciais, ou seja, possuem coordenadas geográficas;
- **Anisotropia**: Se foram considerados parâmetros de anisotropia;
- **Automático**: Se as definições dos variogramas teóricos/experimentais, como ranges de variáveis e hiper-parâmetros foram definidos através dos dados.

É possível perceber que há uma lacuna com relação a pesquisas, com dados espaciais, utilizando a técnica de cluster-kriging. Além disso, como mencionado em diversos trabalhos (ABEDINI; NASSERI; ANSARI, 2008; ABEDINI; NASSERI; BURN, 2012; LI et al., 2018), é importante a utilização de variáveis anisotrópicas na otimização dos parâmetros do variograma teórico. Por fim, só há um trabalho em que podemos considerar que há um processo automático (PESQUER; CORTÉS; PONS, 2011).

Tabela 2. Comparativo entre os modelos dos artigos levantados.

Artigo	**Krigagem**	**Clustering**	**Variog. Param.**	**Dados Espaciais**	**Anisotropia**	**Automático**
(ABEDINI; NASSERI; ANSARI, 2008)	Ordinária	Sim	Manual	Sim	Sim	Não
(BARGAOUI; CHEBBI, 2009)	Ordinária	Não	Manual	Sim	Sim	Não
(XIALIN et al., 2010)	Ordinária	Não	Manual	Sim	Sim	Não
(WEI; LIU; CHEN, 2010)	Ordinária	Não	Otimizados	Sim	Não	Não
(VASAT; HEUVELINK; BORVKA, 2010)	Ordinária	Sim	Otimizados	Sim	Não	Não
(MASOOMI; MESGARI; MENHAJ, 2011)	Ordinária	Não	Otimizados	Sim	Não	Não
(PESQUER; CORTÉS; PONS, 2011)	Ordinária	Não	Otimizados	Sim	Não	Sim
(BALU; ULAGANATHAN; ASPROULIS, 2012)	Ordinária	Não	Manual	Não	Não	Não
(ABEDINI; NASSERI; BURN, 2012)	Ordinária	Não	Manual	Sim	Sim	Não
(BAAR; DWIGHT; BIJL, 2013)	Ordinária	Não	Otimizados	Não	Não	Não
(WANG et al., 2017b)	Ordinária	Não	Manual	Não	Não	Não
(WANG et al., 2017a)	Ordinária	Sim	Otimizados	Não	Não	Não
(GONÇALVES; KUMAIRA; GUADAGNIN, 2017)	Cokrigagem	Não	Otimizados	Não	Não	Não
(ARCIDIACONO et al., 2018)	Ordinária	Não	Otimizados	Não	Não	Não
(LI et al., 2018)	Ordinária	Não	Otimizados	Sim	Sim	Não
(NOURI; MOHAMMADI; ZAREZADEH, 2018)	Ordinária	Não	Otimizados	Não	Não	Não
Proposta da Tese	**Ordinária**	**Sim**	**Otimizados**	**Sim**	**Sim**	**Sim**

3.3.2 Discussão geral das pesquisas

A pergunta *Qual o estado da arte no aprimoramento no processo de krigagem utilizando técnicas de inteligência artificial?*, norteou muitas análises em cima dos 16 trabalhos selecionados. Inicialmente, é possível perceber que a área da Krigagem é ampla e diversos segmentos de seu processo possui espaço e lacunas para aprimoramentos. Dentre os artigos selecionados, podemos identificar pesquisas que buscam melhorar etapas da seleção dos parâmetros do variograma experimental e teórico (ABEDINI; NASSERI; BURN, 2012; VASAT; HEUVELINK; BORVKA, 2010; MASOOMI; MESGARI; MENHAJ, 2011; PESQUER; CORTÉS; PONS, 2011; BALU; ULAGANATHAN; ASPROULIS, 2012; BAAR; DWIGHT; BIJL, 2013), outras que buscam identificar a melhor vizinhança para a aplicação da krigagem através da seleção dos pontos ideais na etapa de interpolação (ABEDINI; NASSERI; BURN, 2012; GONÇALVES; KUMAIRA; GUADAGNIN, 2017),

aprimoramento da função de custo utilizada para avaliar a qualidade dos parâmetros do variograma teórico (LI et al., 2018), utilização de técnicas bioinspiradas na otimização de diversos parâmetros das etapas do processo de krigagem(ABEDINI; NASSERI; BURN, 2012; LI et al., 2018), propostas de krigagem automática realizando a estimativa dos parâmetros do variograma experimental e teórico com base somente nos dados de entrada (PESQUER; CORTÉS; PONS, 2011), trabalhos buscando reduzir a complexidade computacional do processo de krigagem (WANG et al., 2017a; PESQUER; CORTÉS; PONS, 2011) e utilização de algoritmos de agrupamento para seleção de sub-grupos no processo de krigagem (ABEDINI; NASSERI; ANSARI, 2008; WANG et al., 2017a).

3.3.3 Discussão das pesquisas: Questão secundária 1

Questão secundária 1 (QS1): Quais as metodologias, modelos e técnicas utilizadas na otimização/seleção de parâmetros do variograma?

A priori, é possível realizar a distinção da seleção dos parâmetros do variograma, tanto experimental quanto teórico, em duas vertentes: Seleção manual (tentativa e erro), que consiste em utilização de conhecimento especialista sobre o problema estudado; Seleção através de métodos de otimização, que consiste em vasculhar a melhor solução dentro de um espaço de busca. Nos 16 artigos levantados, é possível identificar através da Tabela 2, que 10 artigos utilizaram métodos de otimização na seleção dos parâmetros do variograma, e 6 utilizaram métodos manuais ou parâmetros pré-determinados.

Com relação à parâmetros de anisotropia, este conceito não foi mencionado ou não foi identificado em algumas pesquisas, mais precisamente 5 pesquisas explicitaram a utilização de parâmetros anisotrópicos e 11 não.

Na pesquisa de (PESQUER; CORTÉS; PONS, 2011) é apresentado uma metodologia de seleção e otimização automática dos parâmetros do variograma. A autora menciona a necessidade de gerar resultados com tempo de execução reduzidos de forma a auxiliar a rápida tomada de decisão, e pra isso, se torna necessário a eliminação da necessidade de conhecimento especialista e consequentemente uma metodologia para selecionar parâmetros do variograma experimental e teórico com base somente nos dados. Na etapa de otimização são utilizadas as funções de custo por mínimos quadrados (CRESSIE, 1985) e Levenberg-Marquadt (MARQUARDT, 1963). No entanto, não são especificados parâmetros de anisotropia apesar de serem usados dados espaciais.

Técnicas de agrupamento, como K-means, foram aplicadas nos trabalhos de (ABEDINI; NASSERI; ANSARI, 2008; VASAT; HEUVELINK; BORVKA, 2010; WANG et al., 2017a) para realizar o agrupamento dos dados espaciais antes da aplicação da krigagem, chamados de cluster-kriging. Em (ABEDINI; NASSERI; ANSARI, 2008), os dados espaciais foram divididos em sub-grupos, de forma que a interpolação de um ponto só era

realizado com a vizinhança do mesmo cluster, conforme pode ser observado na Figura 12.

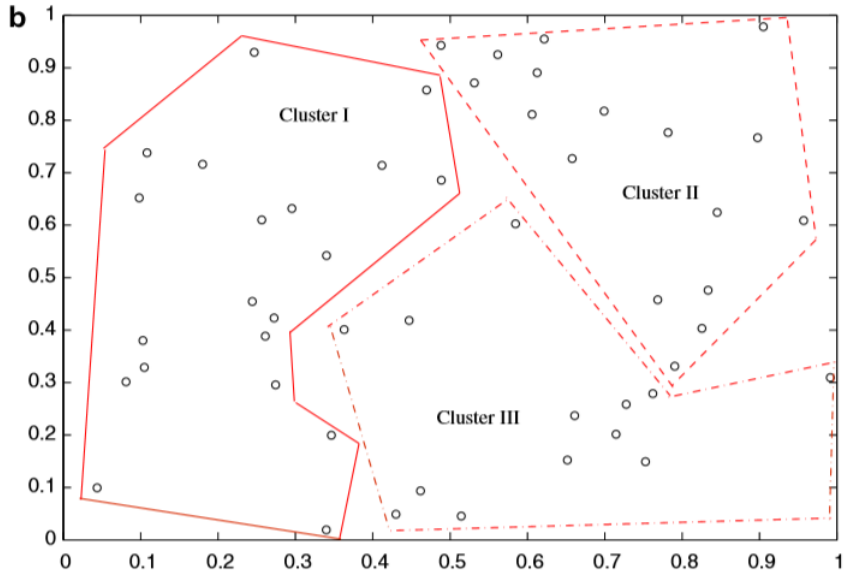

Figura 12. Cluster-Kriging (ABEDINI; NASSERI; ANSARI, 2008).

Uma lacuna deixada pelos autores é um mecanismo confiável para atribuição de um ponto desconhecido à um dos *clusters*. Este mecanismo foi desenvolvido em (WANG et al., 2017a) utilizando árvore de decisão, no entanto, os dados trabalhados eram não espaciais, o que invalida a aplicação em problemas da geoestatística já que não há variáveis georeferenciadas. Atualmente, há algoritmos que foram construídos para tratar de dados espaciais, como o trabalho de Chavent et al. (2018), e que podem ser úteis na tratativa de agrupamento de dados espaciais.

Diferentes algoritmos de otimização foram utilizadas nas 16 pesquisas. Entre eles, pode-se citar Nelder-Mead (BAAR; DWIGHT; BIJL, 2013), Levenberg-Marquadt (ARCIDIACONO et al., 2018), Processos Gaussianos (GONÇALVES; KUMAIRA; GUADAGNIN, 2017), Enxame de Partículas (WANG et al., 2017b) e algoritmos genéticos(MASOOMI; MESGARI; MENHAJ, 2011; LI et al., 2018).

3.3.4 Discussão das pesquisas: Questão secundária 2

Questão secundária 2 (QS2): Em que etapa da krigagem estão sendo utilizadas técnicas de inteligência artificial para aprimoramento do processo?

Nos artigos levantados, a grande totalidade das pesquisas busca aprimorar a estimativa do modelo (variograma teórico). No entanto, esta mesma etapa pode ser divida em algumas sub-tarefas, mais especificamente em:

1. **Seleção dos parâmetros do variograma**: Nos trabalhos (BARGAOUI; CHEBBI, 2009; WEI; LIU; CHEN, 2010; MASOOMI; MESGARI; MENHAJ, 2011; WANG et al., 2017b; LI et al., 2018; ARCIDIACONO et al., 2018; GONÇALVES; KUMAIRA; GUADAGNIN, 2017; NOURI; MOHAMMADI; ZAREZADEH, 2018), os autores tiveram como objetivo aprimorar a seleção dos parâmetros do variograma teórico.

2. **Seleção da vizinhança para interpolação (krigagem)**: Nos trabalhos (XIALIN et al., 2010; ABEDINI; NASSERI; BURN, 2012), buscou-se aprimorar a seleção dos pontos em que participariam do processo de interpolação através de técnicas de inteligência artificial.

3. **Pré-processamento dos dados**: Nos trabalhos de (ABEDINI; NASSERI; ANSARI, 2008; VASAT; HEUVELINK; BORVKA, 2010; WANG et al., 2017a) foram utilizados métodos de agrupamento de dados em etapas antes da aplicação da krigagem e modelagem do variograma teórico.

4. **Redução da complexidade computacional**: Nos trabalhos de (PESQUER; CORTÉS; PONS, 2011; BAAR; DWIGHT; BIJL, 2013) buscou-se reduzir o custo computacional da aplicação da krigagem.

3.3.5 Discussão das pesquisas: Questão secundária 3

Questão secundária 3 (QS3): Qual função de custo está sendo utilizada para a otimização dos parâmetros do semivariograma?

Na Figura 13 é demonstrado a utilização das funções de custo nos trabalhos levantados. É importante mencionar que a função de custo por interpolação está diretamente ligada a utilização de algoritmos bioinspirados, como algoritmos genéticos e enxame de partículas. Nos trabalhos de (XIALIN et al., 2010; WEI; LIU; CHEN, 2010; MASOOMI; MESGARI; MENHAJ, 2011; ABEDINI; NASSERI; BURN, 2012; LI et al., 2018), a função de custo de mínimos quadrados, formulada por (CRESSIE, 1985), é a mais utilizada dentre os 16 artigos selecionados. Como mencionado por (LI et al., 2018), a facilidade de uso e a ampla variedade de implementações na literatura desta função de custo impulsiona a sua alta utilização, como pode ser visto em (ABEDINI; NASSERI; ANSARI, 2008; BARGAOUI; CHEBBI, 2009; VASAT; HEUVELINK; BORVKA, 2010; PESQUER; CORTÉS; PONS, 2011; BALU; ULAGANATHAN; ASPROULIS, 2012; NOURI; MOHAMMADI; ZAREZADEH, 2018). A técnica de minimização/otimização geralmente utilizada junto com a função de custo de mínimos quadrados é a Nelder-Mead (OLSSON; NELSON, 1975) e Levenberg-Marquadt (MARQUARDT, 1963). Já a função de custo *Maximum Likelihood* é a menos utilizada, até pela sua complexidade de implementação e a necessidade de alguns pressupostos para a sua aplicação.

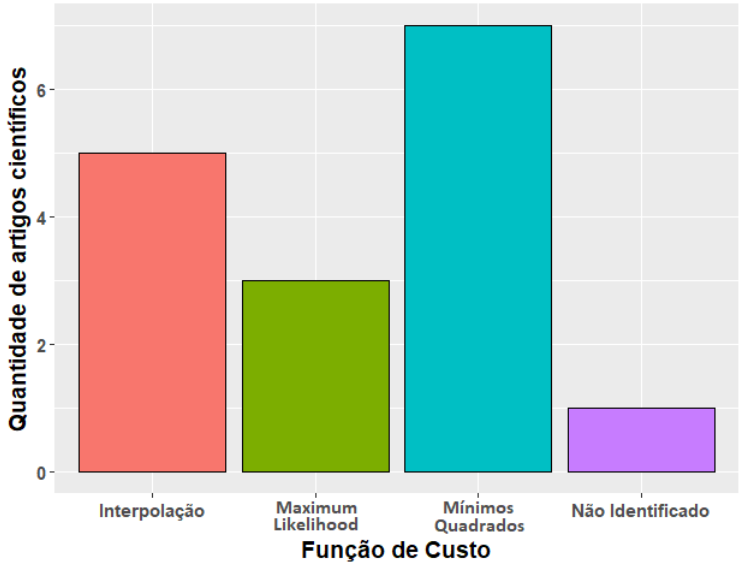

Figura 13. Funções de custo utilizadas nas pesquisas.

3.3.6 Discussão das pesquisas: Questão secundária 4

Questão secundária 4 (QS4): Qual a natureza dos dados em que está se aplicando a krigagem? Dados espaciais ou não espaciais?

A aplicação da krigagem não se limita a dados espaciais apesar de utilizarem o mesmo conceito de modelagem do variograma. A diferença dos dados espaciais é a utilização de coordenadas que identificam a geolocalização da coleta/medição. Foi realizado o levantamento dos tipos de dados que os autores dos artigos utilizaram em suas pesquisas. Como pode ser observado na Figura 14, 11 artigos utilizaram dados espaciais (sejam 2D ou 3D) e 5 artigos utilizaram dados não-espaciais.

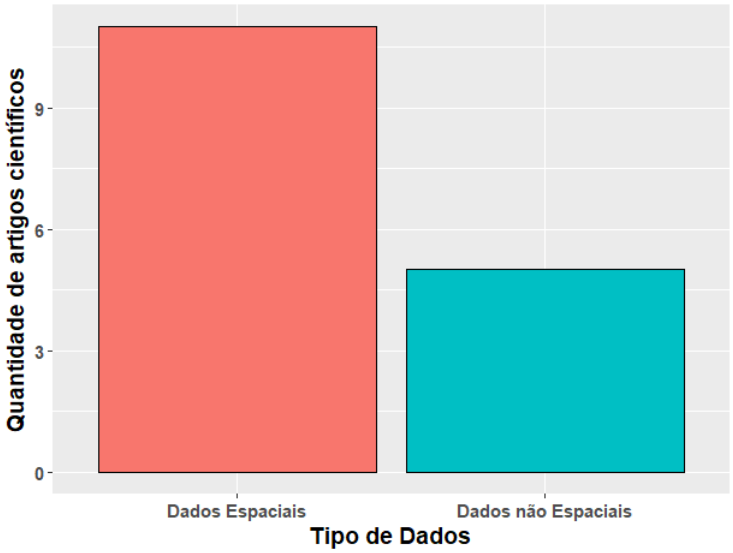

Figura 14. Tipos de dados utilizados nas pesquisas.

4 Modelo Proposto

O modelo proposto consiste na estruturação de um processo automático para definir os parâmetros do variograma teórico, com o intuito de aprimorar o processo da krigagem como um todo e reduzir a dependência que existe do conhecimento especialista. A Figura 15 apresenta um fluxograma do modelo proposto, que está dividido em 4 etapas principais:

1. Pré-processamento dos dados;
2. Clusterização dos dados;
3. Ajuste de um modelo para cada grupo obtido na etapa de clusterização;
4. Associação de *cluster* de novos dados para a realização da interpolação/krigagem.

Para cada número de *clusters*, cada etapa do modelo foi repetido 10 vezes (validação cruzada *10-fold*), exceto na etapa de pré-processamento de dados. Para evitar qualquer tipo de enviesamento do modelo, cada iteração foi realizada utilizando diferentes conjuntos de treinamento e teste.

Na primeira etapa, o pré-processamento dos dados é iniciado através da aplicação de algoritmos de normalização, tratamento de *outliers* e *detrending* dos dados. Em seguida, U conjuntos dos dados são gerados baseados na técnica de clusterização selecionada. Finalmente, um algoritmo de otimização é utilizado para buscar os melhores valores dos parâmetros que definem o modelo do variograma teórico. Este conjunto de parâmetros, representados pelo vetor θ^*, é calculado pela seguinte função objetivo:

$$\theta^* = \arg\min_{\theta \in \Theta} M(\theta), \tag{4.1}$$

onde θ é o conjunto de parâmetros do variograma teórico, obtido automaticamente pelos dados do variograma experimental, e $M(\theta)$ representa o resultado da função de custo para θ. É importante ressaltar que um modelo (e conjunto de parâmetros) é ajustado para cada *cluster*.

Em relação à classificação de novos pontos, um modelo de aprendizado de máquina supervisionado (ou classificador) é utilizado para selecionar o melhor *cluster* para este novo e desconhecido ponto. Assim, os dados que pertencem ao *cluster* alocado são posteriormente utilizados na interpolação deste novo ponto.

É importante observar que qualquer algoritmo do modelo proposto pode ser modificado de acordo com o problema a ser tratado e de acordo com o conhecimento do

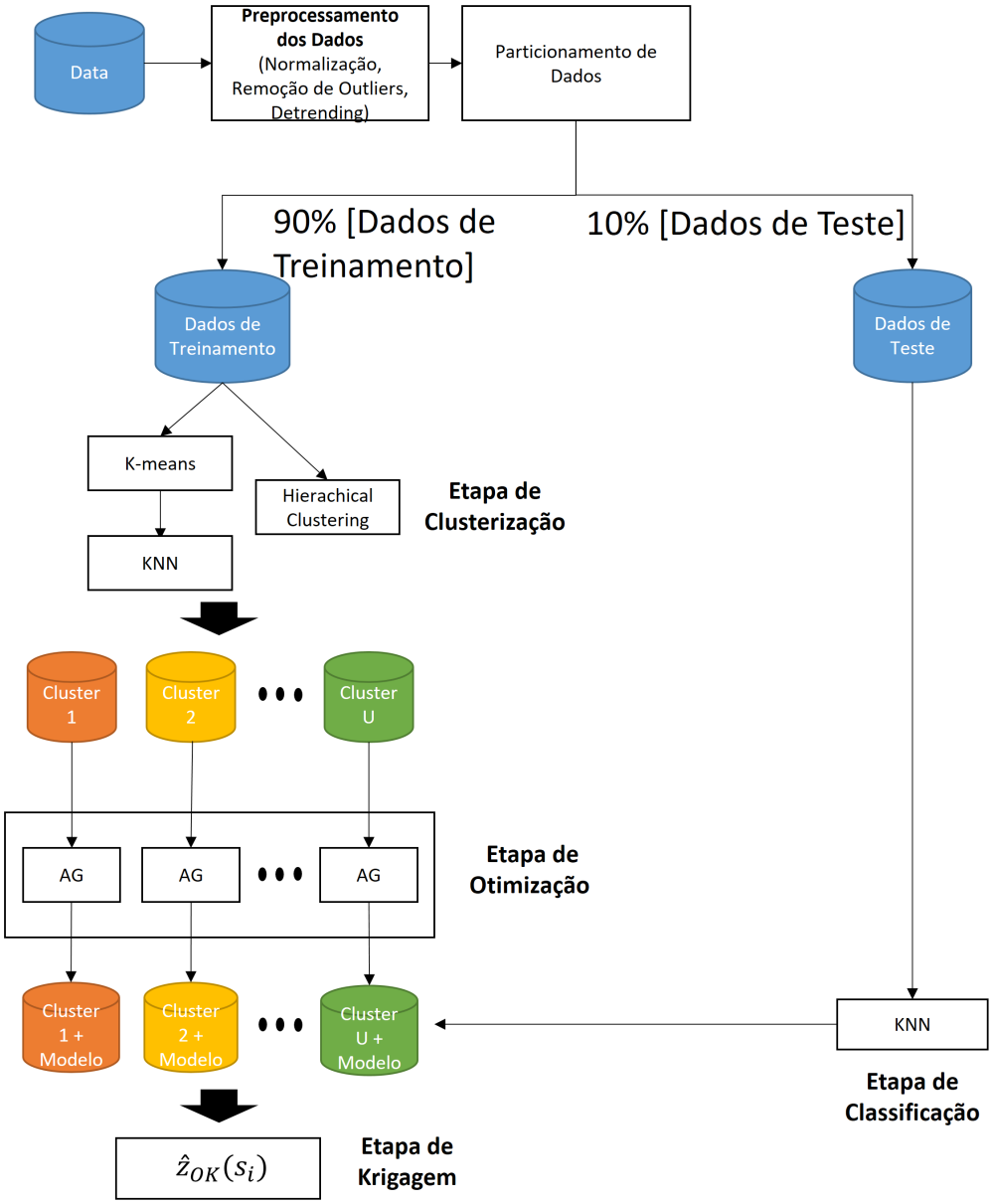

Figura 15. Esquema geral das etapas de pré-processamento, clusterização e ajuste de modelos para cada *cluster* separadamente.

usuário. Em outras palavras, a grande contribuição deste trabalho é estruturar um *baseline* para ser seguido, levando em consideração que diferentes técnicas de pré-processamento, clusterização, classificação e otimização podem ser aplicadas. Além disso, é proposto a normalização dos *clusters* via algoritmo do KNN juntamente com o k-means, uma heurística para definir os limites das variáveis quando utilizados algoritmos bioinspirados e a classificação de pontos desconhecidos a um *cluster* por meio do KNN.

4.1 Pré-processamento dos dados

Como *outliers* podem ter um grande impacto, em alguns casos negativo, na clusterização e no processo de krigagem (APARNA; NAIR, 2016), pode ser utilizado um mecanismo para tratar estes elementos (AMRI; JEMAIN; HASSAN, 2014). Portanto, inicialmente, os vetores com as coordenadas primárias $\mathbf{x}$ e $\mathbf{y}$, assim como o conjunto de observações da variável objetivo $\mathbf{z}$, foram normalizados para o intervalo entre 0 e 1. Este procedimento é importante para assegurar que todas as variáveis possuem o mesmo peso, tanto no processo de clusterização quanto no processo de krigagem, evitando assim problemas como a sobreposição de *clusters*. Na sequência, foi utilizado o teste *Z-score* (BOSLAUGH, 2012) com 99% de confiança para a remoção de *outliers*. Por fim, é aplicado o processo de *detrending* para garantir, pelo menos inicialmente, a hipótese de estacionariedade nos dados originais. Nos experimentos, uma função polinomial de segunda ordem foi utilizada para ajustar a superfície de *trend* (VIEIRA et al., 2010). É importante mencionar que, dependendo da natureza do problema, manter os outliers podem indicar uma importante fonte de informação, nestes casos, fica a necessidade ainda de conhecimento especialista para averiguar a necessidade de manter ou não estes.

4.2 Clusterização dos dados

Na etapa de clusterização, os dados de treinamento são divididos em U *clusters* utilizando o algoritmo k-means com as informações espaciais, as coordenadas x e y, e a variável objetivo z. Como demonstrado em (ABEDINI; NASSERI; ANSARI, 2008), o processo de clusterização, dependendo do número de *clusters* selecionado, pode resultar em *clusters* sobrepostos, ou seja, sem uma uniformidade espacial bem definida. Neste sentido, de forma a minimizar este fenômeno, toda a base de dados foi normalizada entre 0 e 1 e o algoritmo KNN foi aplicado para aprimorar o agrupamento através da alocação do ponto atual com base nos vizinhos mais próximos. Por exemplo, os círculos vermelhos destacados na Figura 16(a) demostram a sobreposição de dados, a qual foi reduzida com a aplicação do modelo proposto, como pode ser observado na Figura 16(b). O tamanho dos círculos indica o valor da variável objetivo.

Os resultados na seção de experimentos foram obtidos utilizando 3 vizinhos mais próximos na classificação e normalização dos dados, ou seja, a configuração com 3-NN. O número de vizinhos foi testado com 1, 3 e 5, sendo que a configuração com 3-NN foi responsável por 70.83% dos melhores resultados, além de apresentar maior estabilidade (menor variância entre os resultados). Como pode ser observado na Figura 17, algumas bases encontradas na literatura foram testadas utilizando diferentes configurações (1, 3 e 5 vizinhos) com algoritmos genéticos. GA1 representa o algoritmo genético utilizando parâmetros de anisotropia e GA2 o algoritmo genético sem a utilização de parâmetros de

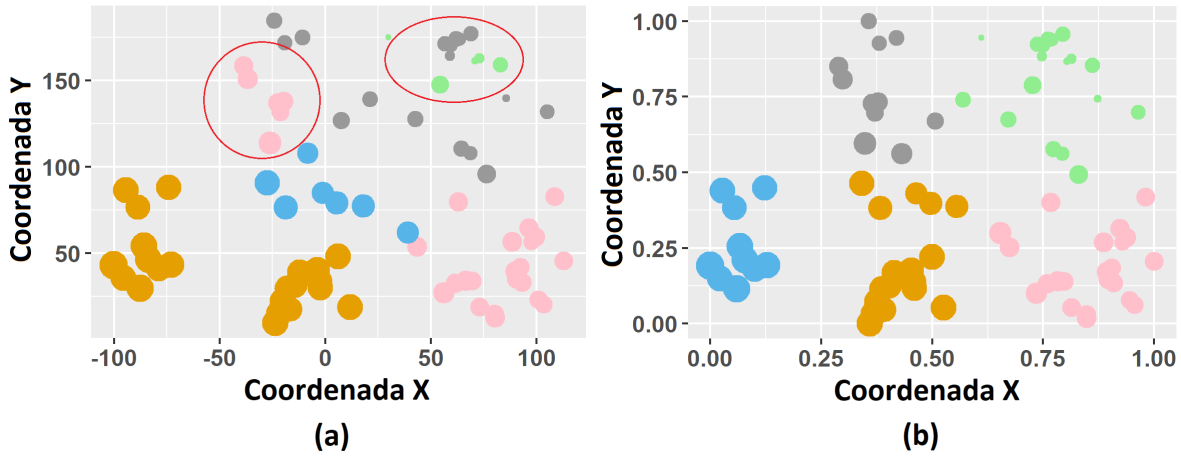

Figura 16. Exemplo da aplicação do algoritmo KNN para minimização da sobreposição de clusters. (a) Antes da normalização + KNN. (b) Após normalização + KNN.

anisotropia.

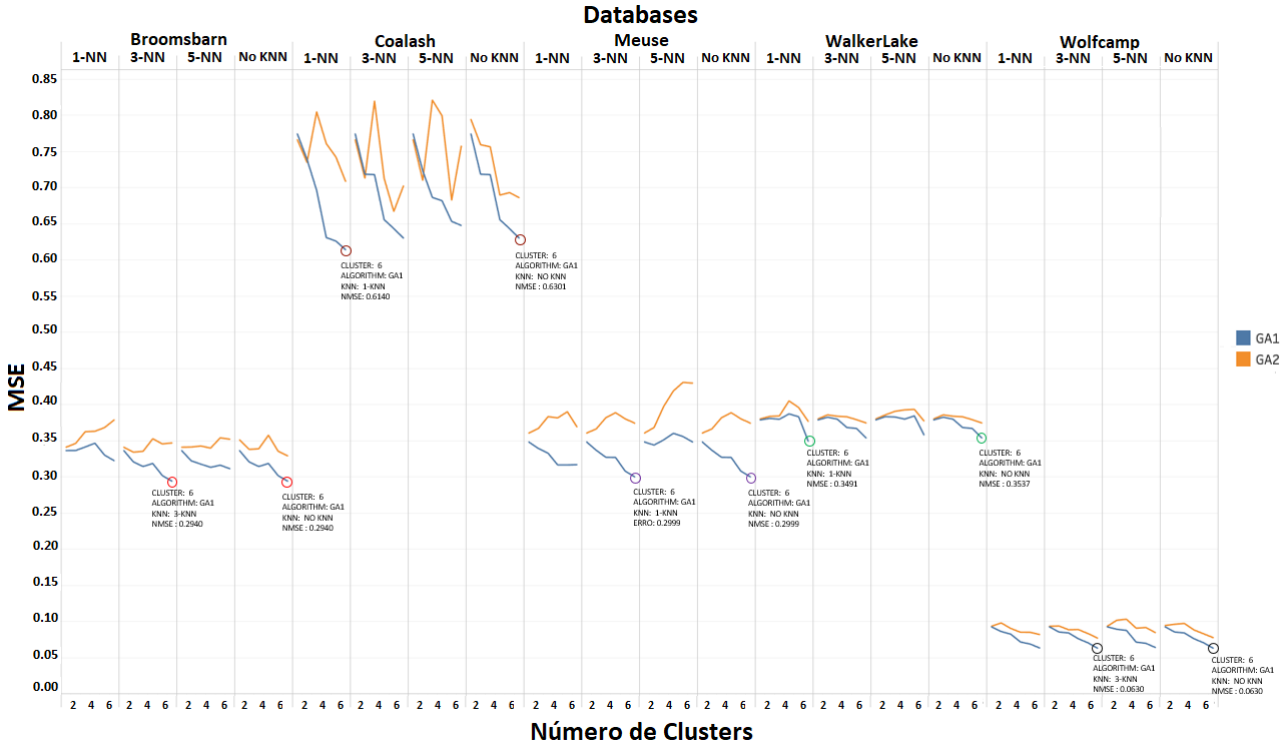

Figura 17. Testes realizados com diferentes vizinhos utilizando o algoritmo genético para otimização dos parâmetros do variograma teórico.

Uma implementação do algoritmo de Clusterização hierárquica chamada ClustGeo (CHAVENT et al., 2018), foi utilizada na etapa de clusterização como uma alternativa à solução proposta com k-means + KNN. Em outras palavras, a ideia é utilizar o algoritmo

ClustGeo como um *baseline*. Este algoritmo aplica um peso para as coordenadas espaciais e a variável objetivo, este peso é chamado de parâmetro *alpha*. Este parâmetro pode ser definido manualmente para identificar os melhores grupos formados. Com base em alguns testes iniciais entre 0.1 e 0.9, este parâmetro foi definido em 0.4.

4.3 Otimização

Na etapa de otimização, algoritmos bioinspirados, como AG e PSO, foram utilizados para otimizar (ou estimar) os melhores valores para os parâmetros que definem o modelo do variograma teórico. Assim, os cromossomos do AG e as partículas do PSO foram estruturadas com as seguintes variáveis: sill, range, kappa, ângulo de anisotropia e fator de anisotropia.

Na Figura 18 está ilustrado como o cromossomo está organizado no algoritmo genético, esta estrutura também foi aplicada para o PSO. Os blocos em verde representam os parâmetros ativos utilizados no processo de otimização e os blocos em vermelho ilustram parâmetros que não foram otimizados, ou seja, seus valores foram definidos manualmente. O efeito pepita, ou *nugget effect*, foi definido com valor padrão 0 para todos os testes e o número de *lags* com valor 10. Ainda, foi utilizada a implementação do AG proposto em Scrucca et al. (2013), que utiliza o cruzamento de Laplace (DEEP; THAKUR, 2007a), e *power mutation*(DEEP; THAKUR, 2007b). Os ranges utilizados para cada variável estão detalhados nas seções de configuração dos experimentos.

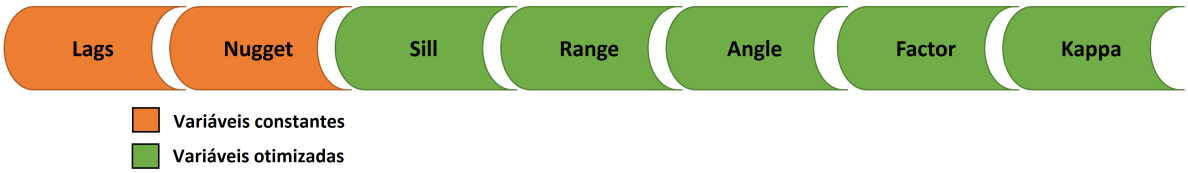

Figura 18. Esquema do cromossomo utilizado na configuração do AG.

4.4 Classificação

O algoritmo KNN foi aplicado para classificar os dados de teste, considerados como novos pontos, em um dos *clusters* definidos na etapa de clusterização através das variáveis de coordenadas x e y. Então, aplicou-se a krigagem e os erros foram calculados para comparação entre as técnicas estudadas.

4.5 Passo a passo

Nesta seção, será demonstrada a aplicação do modelo proposto em uma das bases de dados utilizada nos experimentos. Cada etapa do modelo será apresentado com ilustrações da configuração dos dados e o impacto da aplicação das técnicas. É importante ressaltar que o objetivo nesta seção não é apresentar os resultados, mas sim elucidar as etapas do modelo.

4.5.1 Dados originais

Na Figura 19 podemos observar a configuração espacial dos dados da base *Meuse*. O gradiente em verde indica o valor de zinco resultante da medição naquele ponto.

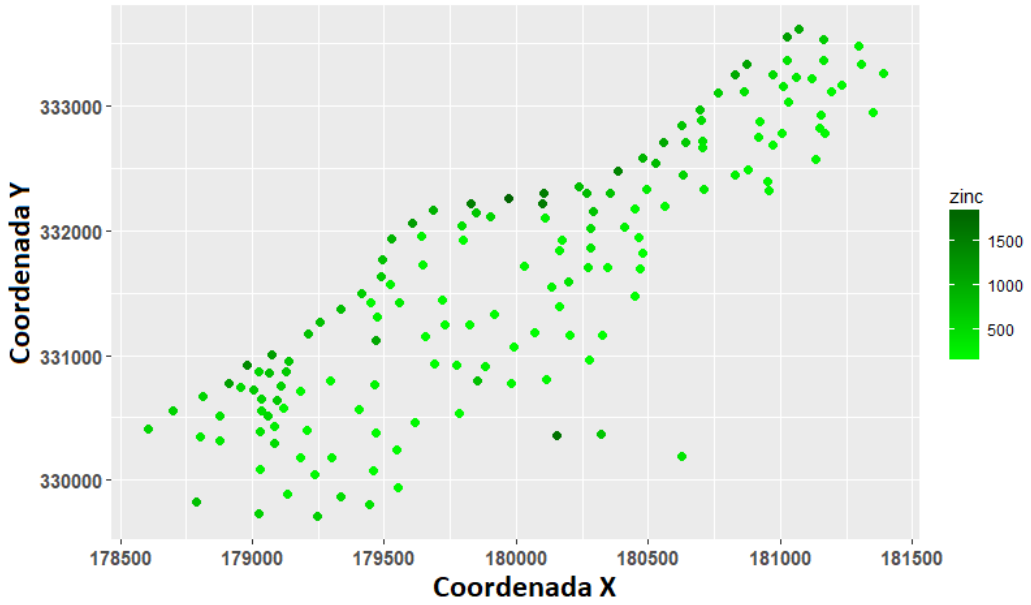

Figura 19. Estrutura espacial original da base de dados *Meuse*.

4.5.2 Remoção de outliers, normalização e detrending

A etapa da inicial do modelo, consiste, se desejado pelo usuário, na remoção de outliers que podem existir na base de dados. Utilizando um intervalo de confiança de 99%, os outliers podem ser visualizado na Tabela 3 e na Figura 20.

Como mencionado anteriormente, *outliers* podem conter informações relevantes acerca da área de estudo, no entanto, quando não é o caso, pode influenciar diretamente e negativamente na próximas etapas do modelo. A etapa de clusterização e da krigagem são sensíveis aos *outliers*, com isso, se desejado pelo usuário, a etapa de identificação e retirada de *outliers* pode ser desconsiderada.

Tabela 3. Coordenadas X e Y e valores dos outliers identificados.

X	Y	Zinco
180383	332476	1454
180103	332297	1548
179973	332255	1839
179826	332217	1528
180100	332213	1571
178981	330924	1383
180151	330353	1672

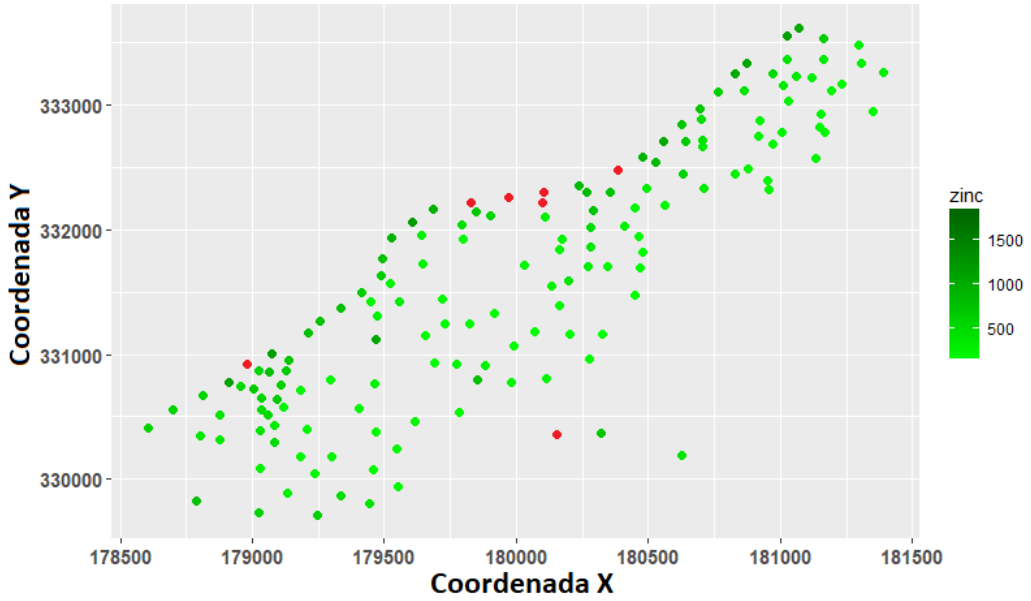

Figura 20. Estrutura espacial da base de dados *Meuse* indicando, em vermelho, os *outliers* identificados.

Por fim, temos a etapa de normalização dos dados em 0 e 1 e a aplicação de métodos de *detrending*. A estrutura espacial resultante destes processos pode ser visualizada na Figura 21. É importante lembrar que tantos as coordenadas como o valor da variável de estudo estão sujeitas ao processo de normalização. A alteração de valores ao longo dos pontos é resultado do *detrending*.

4.5.3 Particionamento

Após os procedimentos realizados no pré-processamento, particiona-se os dados em 90% para treinamento e 10% para teste. Esta etapa tem como objetivo realizar as validações dos métodos. Na Figura 22 pode ser observado a configuração dos dados de uma execução/teste do modelo proposto na etapa de particionamento.

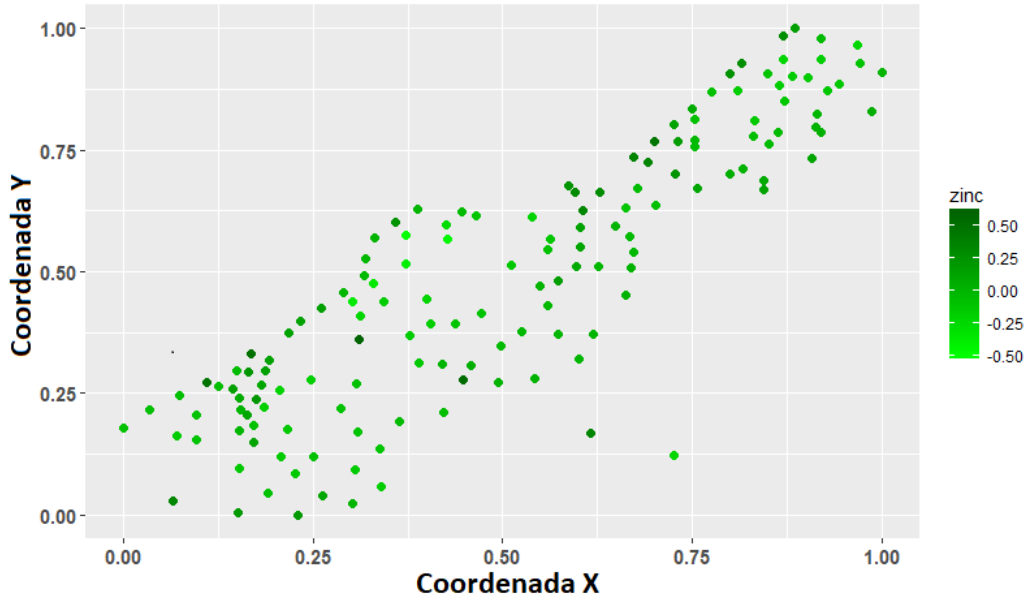

Figura 21. Estrutura espacial da base de dados *Meuse* final após etapa de normalização e detrending.

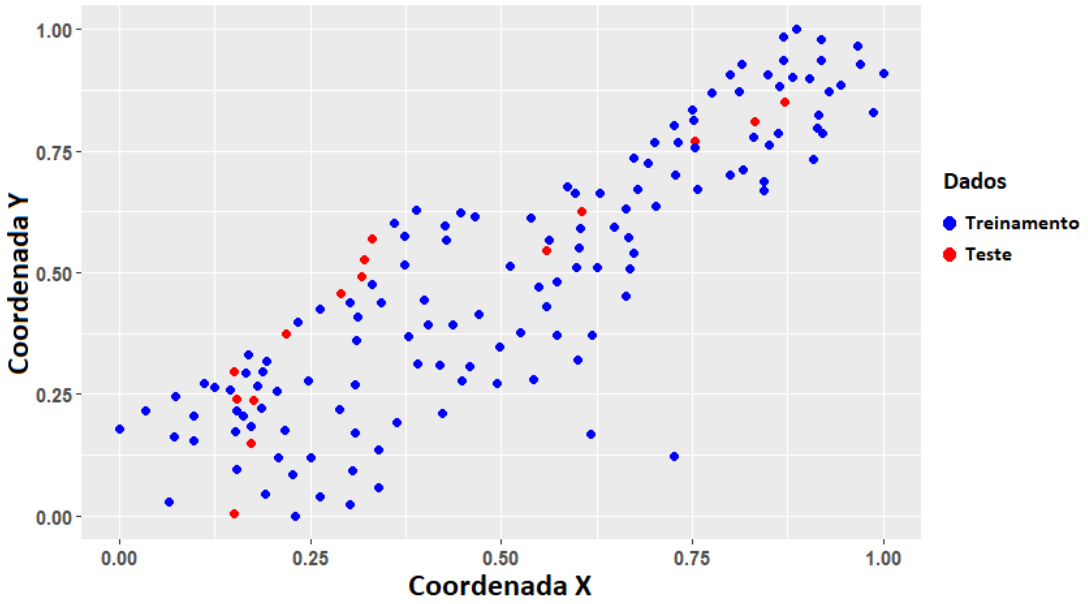

Figura 22. Pontos selecionados para treino (azul) e teste (vermelho) para uma iteração do modelo.

4.5.4 Clusterização

Após a etapa de particionamento, a base de treinamento é submetida a um processo de clusterização. Nas Figuras 23 e 24 podem ser observados o agrupamentos dos dados para 2 exemplos: 2 e 3 clusters utilizando K-Means + KNN.

É importante mencionar que devido ao processo de normalização dos dados, os

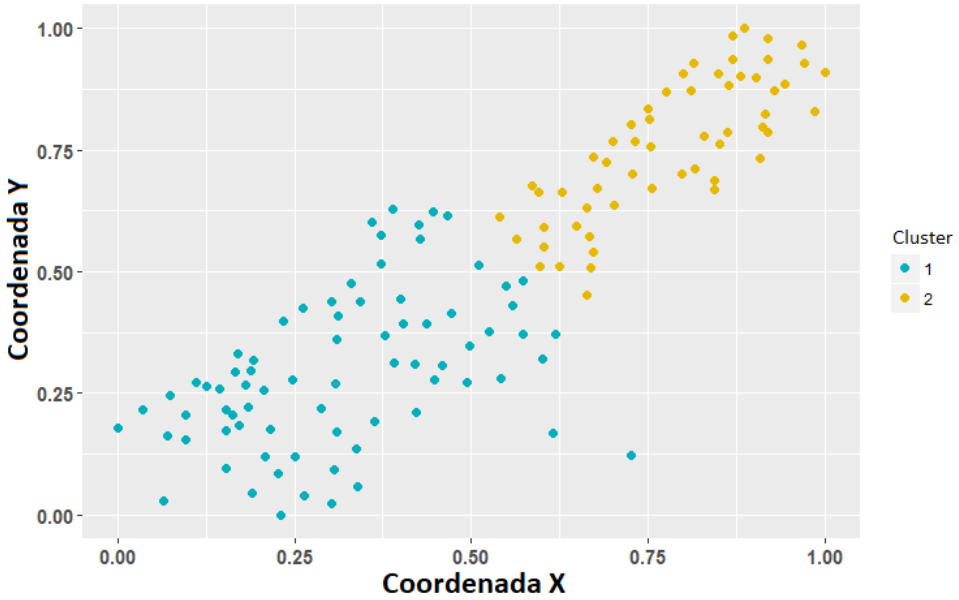

Figura 23. Clusterização da base de dados com 2 clusters utilizando a técnica de agrupamento K-Means + KNN.

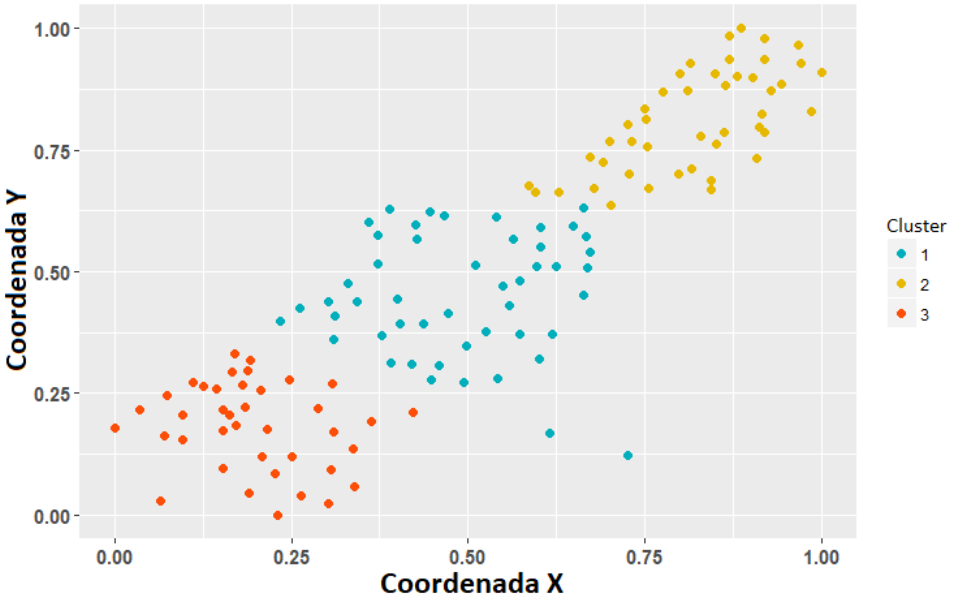

Figura 24. Clusterização da base de dados com 3 clusters utilizando a técnica de agrupamento K-Means + KNN.

grupos formados permaneceram uniformes espacialmente, ou seja, sem a sobreposição de clusters.

4.5.5 Otimização

Nesta etapa, cada grupo formado na etapa de clusterização é submetido à um dos algoritmos de otimização (AG, PSO, Levenberg-Marquardt, etc) para seleção dos parâmetros do variograma teórico. Consequentemente, serão formados n conjuntos de parâmetros de acordo com o número de grupos formados.

4.5.6 Alocação de pontos desconhecidos

Após a seleção dos parâmetros na etapa de otimização dos n grupos formados na etapa de clusterização, os pontos da base de teste deverão ser submetidos à um método de classificação para alocá-los em um dos *clusters*. Nos exemplos das Figuras 25 e 26, os pontos *desconhecidos* foram alocados utilizando o método de KNN com 3 vizinhos. Os pontos foram destacados com a borda preta.

Figura 25. Pontos desconhecidos (destacados com a borda preta) alocados via KNN com 3 vizinhos mais próximos para a base com 2 clusters.

4.5.7 Krigagem e geração de mapas da krigagem

As etapas anteriores são necessárias basicamente para a realização da krigagem de fato, que é a interpolação dos pontos da base de dados, utilizando:

1. Os parâmetros do variograma teórico estimados na etapa de otimização;
2. A vizinhança do ponto a ser interpolado pela krigagem, que deverá ser aqueles pertencentes ao mesmo grupo classificado na etapa de *Alocação de pontos desconhecidos*;

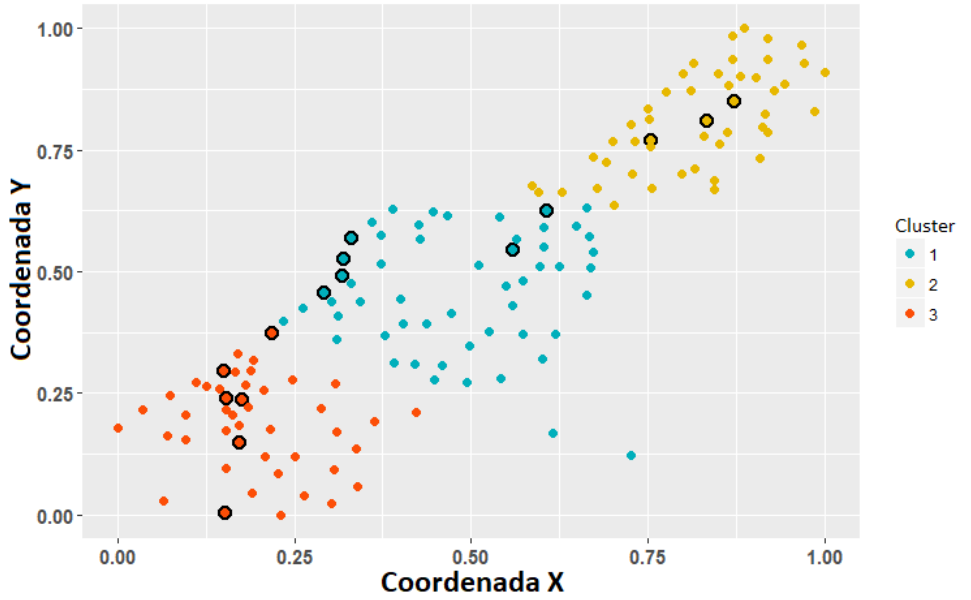

Figura 26. Pontos desconhecidos (destacados com a borda preta) alocados via KNN com 3 vizinhos mais próximos para a base com 3 clusters.

3. Por fim, se desejável, gerar os mapas da krigagem com a estimativa dos pontos e a variância destas estimativas.

Uma análise e demonstração mais aprofundada desta etapa pode ser vista na seção de *Experimentos e Resultados*.

5 Experimentos e Resultados

Este Capítulo descreve os experimentos realizados da aplicação do modelo em bases de dados disponíveis na literatura.

Primeiro, as bases utilizadas nesta tese são apresentadas na Seção 5.1. Em seguida, um estudo comparativo entre as técnicas bioinspiradas, algoritmos genéticos e enxame de partículas, é apresentado. Este estudo serviu para a seleção do algoritmo que servirá de baseline para o restante dos experimentos, já que a utilização dos dois algoritmos seria computacionalmente custoso, e levaria um longo tempo para processamento de todos os testes.

Por fim, um estudo mais aprofundado da aplicação do modelo proposto é apresentado, discutindo os impactos da clusterização e da hipótese da estacionariedade no modelo, além de um comparativo completo com técnicas de otimização consolidadas na área da krigagem. Todos os testes foram executados em um computador com processador Intel Core i5-4570 3.20GHz e 8 GB de RAM.

5.1 Bases de dados

Duas bases de dados disponíveis na literatura foram utilizadas para avaliar o modelo proposto: Meuse (CLARK, 1979) e Wolfcamp (MASOOMI; MESGARI; MENHAJ, 2011). O número de medições e a variável objetivo em estudo para cada base de dados podem ser observadas na Tabela 4. Essas bases de dados foram selecionadas levando em consideração o comportamento distintos entre elas em relação aos níveis de *trends*.

Tabela 4. Informações das bases de dados.

Base de dados	**Medições**	**Variável Objetivo**
Meuse	155	Zinc
Wolfcamp	85	Piezometric Level

A base de dados Wolfcamp é consequência de uma pesquisa realizada na área montanhosa de *West Texas* no Novo México, que tem uma área total é de 96,560 km^2. Esta área tem sido extensivamente objeto de estudo por ser local em potencial para despejo de lixo nuclear (CRESSIE, 1985). Nesta área de estudo, foram realizadas medições em 85 pontos de poços piezométricos. Poço Piezométrico é um equipamento para medir pressões estáticas ou a compressibilidade dos líquidos. Usam-se em furos que servem para monitoração de níveis da água nos aquíferos. As distâncias entre os pontos de medição

foram distintas, ou seja, irregulares. A seleção desta área de dados é devido a sua natureza anisotrópica (ABEDINI; NASSERI; ANSARI, 2008). A configuração espacial desta base de dados pode ser observada na Figura 27.

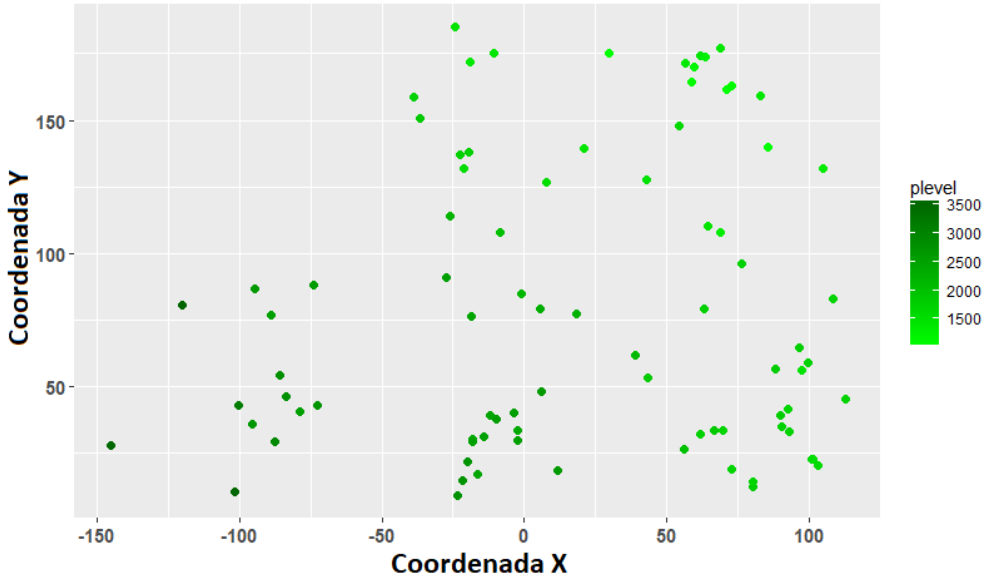

Figura 27. Estrutura espacial da base de dados Wolfcamp.

A base de dados Meuse refere-se a coleta de informações de metais pesados no solo. As amostras foram coletadas na planície de inundação do rio Meuse, próximo a vila Stein na Holanda. No total foram realizadas 155 medições e a organização espacial pode ser observada na Figura 28.

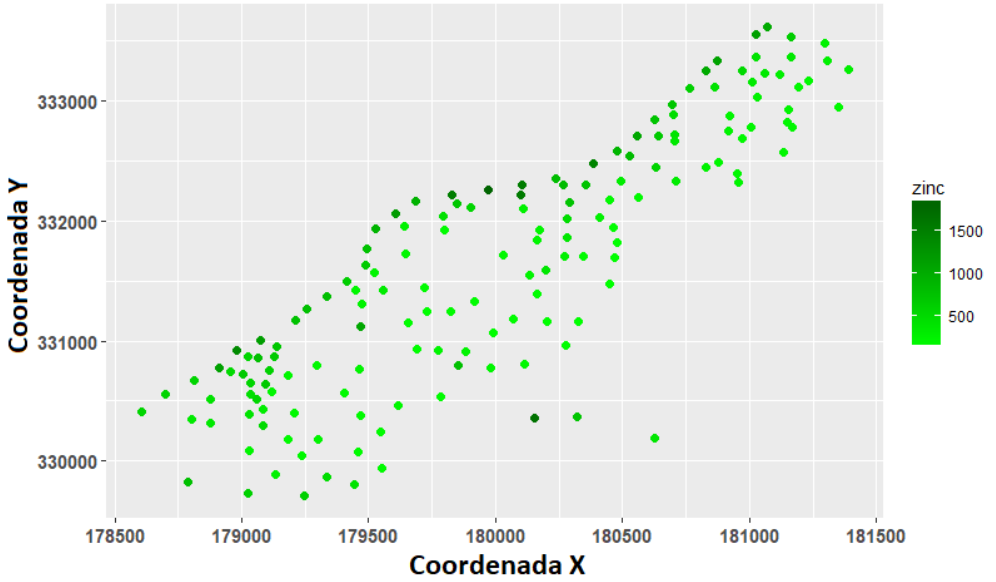

Figura 28. Estrutura espacial da base de dados Meuse.

5.2 Comparativo das técnicas bioinspiradas: AG e PSO

Com o objetivo de comparar e avaliar os resultados dos dois algoritmos bioinspirados, AG e PSO, um *tuning* manual foi realizado com base em dois fatores: tempo de execução (custo computacional) e convergência dos algoritmos. O tamanho da população e o número de iterações foi igual para ambos os algoritmos, e outros parâmetros como cruzamento, seleção e mutação do AG, foram padrão da implementação utilizada. Os valores finais dos parâmetros dos algoritmos podem ser observados na Tabela 5.

Tabela 5. Parâmetros do AG e PSO.

Algoritmo	**Parâmetro**	**Valor**
AG	Tamanho da População	100
	Gerações	50
	Probabilidade de Cruzamento	0.8
	Probabilidade de Mutação	0.1
	Elitismo	5%
	Método de Seleção	Roleta
	Método de Cruzamento	Números Reais (DEEP; THAKUR, 2007a)
	Método de Mutação	Números Reais (DEEP; THAKUR, 2007b)
PSO	Partículas	100
	Iteração	50
	Constante Social	2
	Constante Cognitiva	2
	Range da Inércia	0.4 até 0.8

Para simplificação dos testes realizados, alguns parâmetros do variograma experimental foram fixados, como o número de *lags* em 10, o tipo do modelo do variograma teórico em exponencial e o efeito pepita em 0. Assim, o cromossomo (AG) e a partícula (PSO) tiveram as seguintes variáveis: sill, range, ângulo de anisotropia e fator de anisotropia. Os intervalos das variáveis, após a etapa de pré-processamento dos dados, foram definidos como: range (= 0 até d); sill (= 0 até σ^2); ângulo (= 0 até 180°); e fator (= 0 até 1); onde d é a distância máxima entre dois pontos na base de dados e σ^2 a variância da variável objetivo.

Os *fitness* do AG e PSO considerando de 1 até 5 *clusters* no processo de otimização estão descritos na Tabela 6 para a base Wolfcamp. Para configurações com mais de 1 *cluster*, os resultados foram obtidos através da soma dos erros médios quadráticos. Foram realizadas 10 execuções para cada número de *clusters*. É possível notar na Tabela 6 que os melhores resultados (ou menor error) foram obtidos a medida que o número de *clusters* aumenta para ambos os algoritmos. O desvio padrão e a média dos erros médios quadráticos também tende a cair com o aumento de *clusters*. Apesar do PSO ter encontrado melhores valores, pode-se inferir que o AG é mais estável, já que o desvio padrão apresentado pelo

PSO foi bem mais alto que o AG.

Tabela 6. Melhor Fitness/MSE, média dos fitness e desvio padrão dos fitness.

		Número de clusters				
		1	**2**	**3**	**4**	**5**
AG	Melhor Fitness	0.357	0.341	0.353	0.310	**0.287**
	Média	0.458	0.457	0.441	0.407	**0.381**
	Des. Pad.	0.098	0.051	0.036	0.038	0.044
PSO	Melhor Fitness	0.354	0.299	0.314	0.269	**0.222**
	Média	0.739	0.625	0.537	0.513	**0.498**
	Des. Pad.	0.712	0.300	0.114	0.111	0.110

Na Figura 29 são apresentas as curvas de convergências dos algoritmos AG e PSO, considerando de 1 à 5 clusters na etapa de otimização para a base de dados Wolfcamp. Os resultados apontaram que o PSO converge mais rápido que o AG. Em outras palavras, o PSO atinge o melhor resultado encontrado pelo AG antes da décima geração na maioria dos casos. Essa melhor convergência do PSO é provavelmente explicada pelo fato do PSO ter apresentado populações com nível de diversidade elevado, como será discutido nas seções posteriores. Outro ponto importante, pensando em testes futuros, é que a convergência de ambos os algoritmos (PSO e AG) é mais acentuada entre as gerações 10 e 20, o que implica na possibilidade de diminuir o número de gerações utilizadas, e consequentemente o custo computacional, sem comprometer a qualidade dos resultados.

Na Figura 30 pode ser visualizado o comportamento da diversidade padrão da população, ou SPD, calculada para ambos os algoritmos com a configuração de 1 *cluster*, já que o comportamento foi idêntico para as outras configurações. É possível perceber que até a geração 50, o PSO mantém uma diversidade maior que o AG. Apesar de não garantir que um melhor resultado seja alcançado, pode-se inferir que as soluções no PSO estão mais "espalhadas" que no AG considerando todo o espaço de busca.

Na Figura 31 são apresentados os resultados da etapa de classificação, levando em consideração a validação cruzada *10-fold*. Mais especificamente, para cada iteração (= 10 para cada número de clusters) foram utilizadas 10% de amostras da base de dados de treino para testar os parâmetros do variograma teórico definidos com os 90% das amostras restantes. O *p-value* obtido pelo teste de Friedman (GIBBONS; FIELDEN, 1993) foi 0.17, indicando que não há diferença estatística entre os valores encontrados pelo AG e PSO.

Devido ao AG apresentar maior estabilidade (menor variância) e acurácia estatisticamente equivalente ao PSO, no prosseguimento dos testes mais detalhados, o AG foi selecionado para comparativo com outras técnicas consolidadas na literatura.

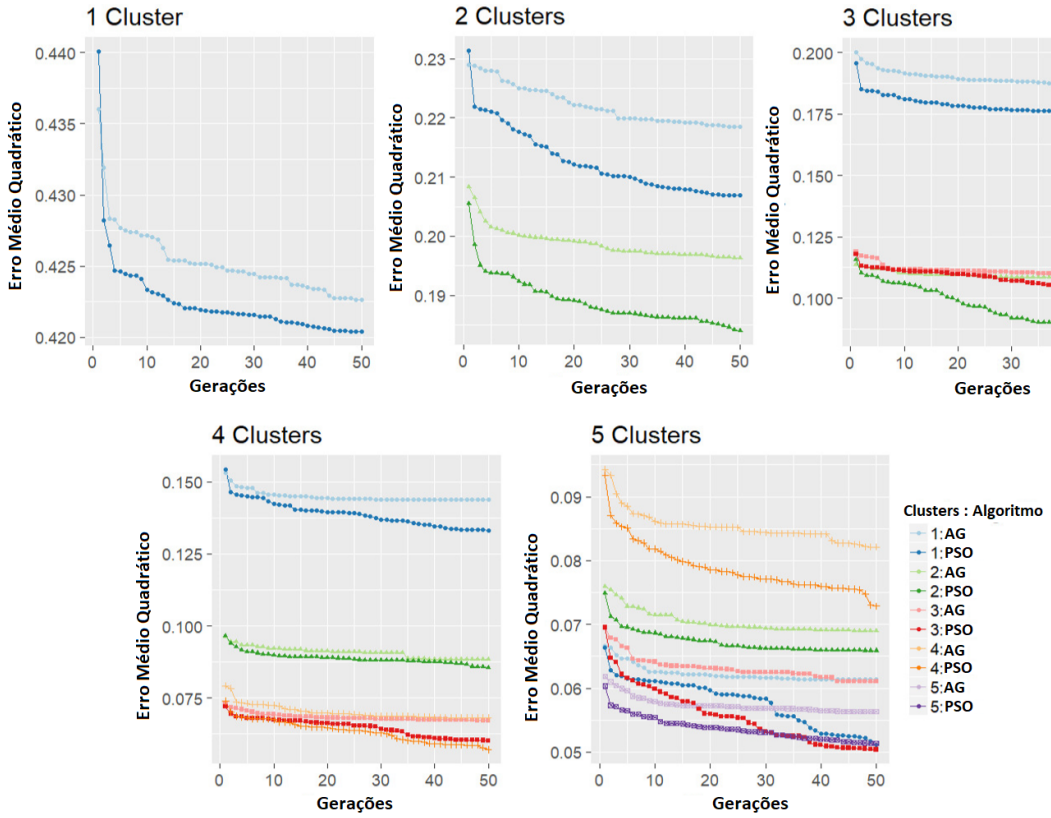

Figura 29. Curvas de convergência para 1 até 5 clusters. Cada linha representa a média de 10 execuções do AG e PSO.

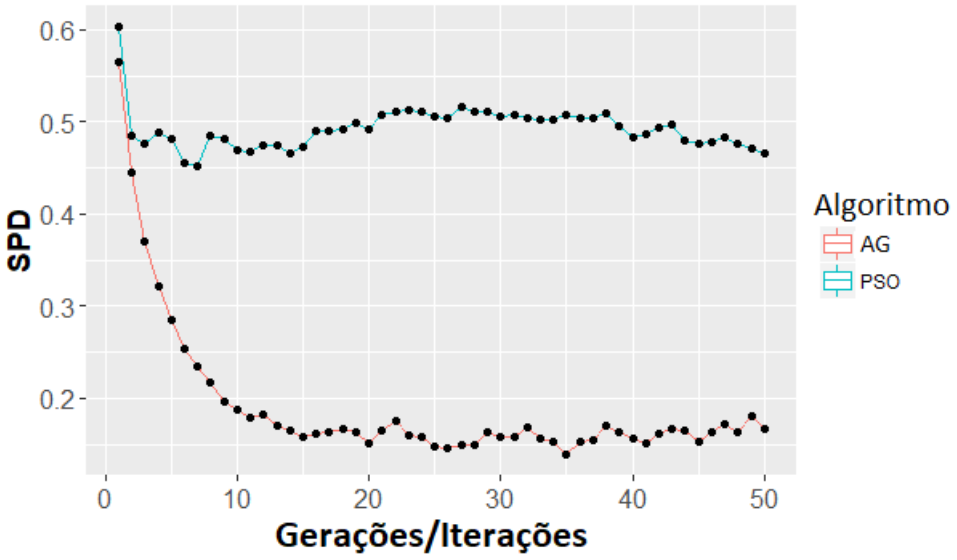

Figura 30. SPD para AG e PSO. Média de 10 execuções para configuração com 1 *cluster*.

5.3 Organização dos experimentos

Os algoritmos aplicados para otimização dos parâmetros do variograma teórico estão descritos na Tabela 7. O objetivo principal é utilizar um algoritmo bioinspirado, neste

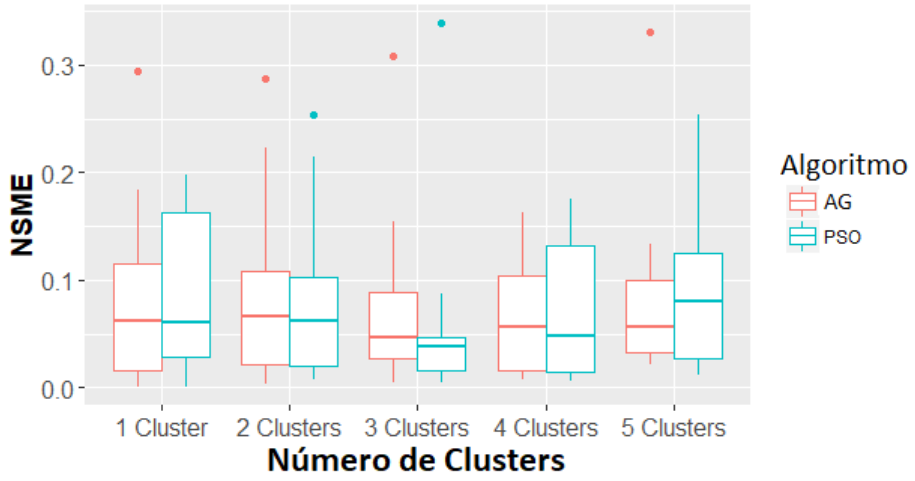

Figura 31. Boxplot dos MSE considerando a etapa de classificação para o AG e PSO. Média de 10 execuções para cada número de *cluster*.

caso o AG, para servir como base para o modelo proposto e comparar o seu desempenho com outros modelos largamente utilizados na literatura. É importante ressaltar que a grande maioria dos modelos utilizados para a krigagem, atualmente são determinísticos e o AG é considerado como um modelo estocástico, apesar de não ser o ideal, estes resultados servem como um norte para futuras pesquisas e aprimoramentos nesta área do conhecimento. Como pode ser observado também na Tabela 7, os algoritmos utilizam diferentes funções de custo, além de variações com e sem a utilização de parâmetros de anisotropia para comparações.

Tabela 7. Lista dos algoritmos de otimização e funções de custo utilizados nos experimentos.

Algoritmo	**Função de Custo**	**Anisotropia**	**Abreviação**
AG Proposto	Interpolação (LI et al., 2018)	Sim	AG
Gauss-Newton	Mínimos Quadrados Iterativo	Sim	GN-ILS1 (DESASSIS; RENARD, 2013)
L-Marquadt (MARQUARDT, 1963)	Mínimos Quadrados com Pesos	Não	LM-WLS (PEBESMA, 2004)
Gauss-Newton	Mínimos Quadrados Iterativo	Não	GN-ILS2 (DESASSIS; RENARD, 2013)

Entre as funções de custo que avaliam a qualidade dos parâmetros selecionados, temos a função de mínimos quadrados com peso ou *weighted least squares (WLS)* (CRESSIE, 1985; MARQUARDT, 1963), que de acordo com Li et al. (2018) apresenta bons resultados no ajuste do modelo do variograma teórico com relação a outras funções de custo. Mais recentemente, em (DESASSIS; RENARD, 2013), os mínimos quadrados iterativos ou *iterative least squares (ILS)* foi proposto como um aprimoramento da função de custo WLS. Já para algoritmos bioinspirados, muitos trabalhos utilizaram a função de custo por interpolação (Equation 2.13), como visto em (XIALIN et al., 2010; WEI; LIU; CHEN, 2010; MASOOMI; MESGARI; MENHAJ, 2011; ABEDINI; NASSERI; BURN, 2012; LI et

al., 2018). Assim, esta função foi empregada no algoritmo AG testado juntamente com as outras técnicas.

GN-ILS1, GN-ILS2 e LM-WLS são algoritmos determinísticos que requerem uma semente inicial para iniciar o processo de otimização. Os valores apresentados na Tabela 8 foram definidas de acordo com os trabalhos de (DESASSIS; RENARD, 2013) e (LARRONDO; NEUFELD; DEUTSCH, 2003), onde σ^2 é a variância da variável objetivo e d é a maior distância entre dois pontos considerando dados de cada *cluster* separadamente, O valor inicial de Kappa foi fixado em 0.5.

Tabela 8. *Seed* inicial para os algoritmos GN-ILS1, GN-ILS2 e LM-WLS.

Algoritmo	**Sill**	**Range**	**Ãngulo**	**Fator**	**Kappa**
GN-ILS1	σ^2	$d/2$	$0°$, $45°$, $90°$, and $135°$	1	0.5
GN-ILS2	σ^2	$d/2$	-	-	0.5
LM-WLS	σ^2	$d/2$	-	-	0.5

Com relação ao AG proposto, alguns testes foram realizados anteriormente para definição de um número adequado de gerações, levando em consideração a curva de convergência e o custo computacional. É importante ressaltar que técnicas de *tuning* não foram utilizadas, como em (VEČEK et al., 2016; BIRATTARI et al., 2002; TRINDADE; CAMPELO, 2019), mas sim um processo manual de testes e verificações foi aplicado para esta tarefa. Neste caso, apesar dos testes preliminares serem realizados com 50 gerações, no caso do comparativo entre AG e PSO, este número foi reduzido para 20 já que o custo computacional mostrou-se elevado para os experimentos subsequentes. No que se refere as taxas de cruzamento e mutação seus valores foram definidos em 0.9 e 0.1, respectivamente.

Além disso, um intervalo de possíveis valores precisa ser definido para cada variável a ser otimizada. Para isso, uma heurística foi criada para automaticamente definir os limites mínimos e máximos, com o objetivo de cobrir todo o espaço de busca da melhor solução. Estes limites estão definidos na Tabela 9.

Tabela 9. Heurística aplicada ao AG proposto.

Algoritmo	**Limite**	**Sill**	**Range**	**Ãngulo**	**Fator**	**Kappa**
AG	Mínimo	0	0	0	0	0
	Máximo	$\sigma^2.5$	d	$180°$	1	1

Em relação ao parâmetro sill, diversos valores de limites máximos foram testados. Os resultados desse teste podem ser observados na Figura 32. O índice NMSE médio após 5 iterações foi avaliada nas bases Meuse e Wolfcamp utilizando AG (sem clusterização),

apenas para efeito de testes. Todas as configurações alcançaram resultados equivalentes, e, para os experimentos subsequentes, o valor máximo do sill foi fixado em $\sigma^2.5$ dado que foi observada uma estabilidade com esse parâmetro.

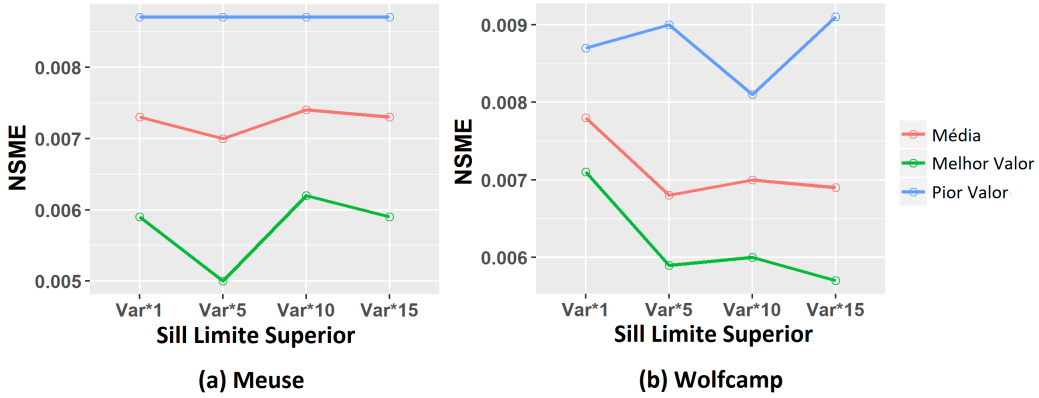

Figura 32. Testes com diferentes valores de limite máximo para o parâmetro sill para a otimização do AG. O eixo Y representa o erro e o eito X representa 4 valores de limites máximos testados: A variância da variável objetivo multiplicado por 1, 5, 10 e 15.

5.4 Discussão

A remoção dos trends é uma das tarefas realizadas durante o bloco de pré-processamento dos dados. Portanto, esta primeira discussão investiga o impacto do processo de *detrending* na hipótese de estacionariedade. Os variogramas das bases Meuse e Wolfcamp, antes e depois do processo de *detrending*, são exibidos na Figura 33. É possível identificar que os dados sem *trends*, ou seja, sujeitos ao *detrending*, apresentaram um comportamento mais estável que os variogramas originais, especialmente na base de dados Wolfcamp, onde a variância estava crescendo exponencialmente.

A ocorrência de *trends* é mais visível no gráfico de dispersão da base Meuse, ilustrado na Figura 34, e o gráfico de dispersão da base Wolfcamp, ilustrado na Figura 35. A linha inclinada nas coordenadas X e Y em ambas as figuras indicam que a média varia de acordo com a distância. Esse pressuposto já é suficiente para identificarmos *trends* nesses dados. A intensidade da inclinação é maior na base de dados Wolfcamp, comportamento esperado de acordo com trabalhos já publicados (ABEDINI; NASSERI; ANSARI, 2008). Após o *detrending*, pode ser observado que a linha é reta, indicando média constante ao longo dos dados e a hipótese de dados estacionaria é alcançada.

Antes do início da etapa de clusterização, a método híbrido proposto, K-means + KNN, e o algoritmo ClustGeo podem manter a continuidade espacial através do ajuste de alguns parâmetros das técnicas. No entanto, é importante ressaltar que isso não garante a hipótese de estacionariedade dos agrupamentos de dados criados posteriormente. Os

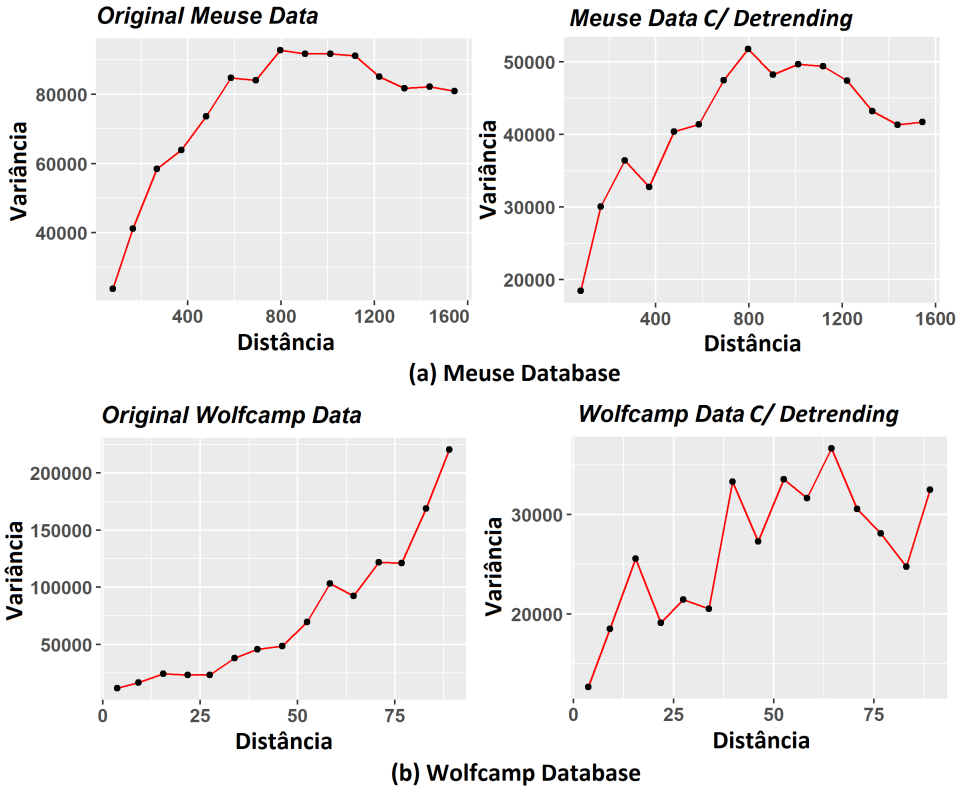

Figura 33. Variogramas experimentais das bases Meuse e Wolfcamp, antes e depois do processo de *detrending*.

resultados da ocorrência de *trends* nos grupos criados pela etapa de clusterização são exibidas na Tabela 10. Um total de 10 execuções foi realizada para cada configuração e o teste de Mann-Kendall (MCLEOD, 2005) foi aplicado para identificar a presença de *trends* em cada *cluster*. As porcentagens na Tabela 10 indicam o número de grupos com *trends* dentro das 10 execuções. Para ambos o algoritmos de clusterização, ClustGeo e K-Means+KNN, a etapa de *detrending* foi benéfica, reduzindo a ocorrência do fenômeno. Contudo, é importante ressaltar que alguns clusters, mesmo com o processo de *detrending*, apresentaram infrações na hipotese de estacionariedade, por exemplo, ocorrendo em média 50% na base de dados Meuse com a configuração com 3 clusters, ou seja, em 15 dos 30 grupos criados.

Após a análise da influência dos trends, é discutido os resultados principais obtidos com o modelo proposto. Para isso, a técnica de visualização da informação por treemap foi utilizada, principalmente devido ao grande número de variáveis e classes envolvidas. O treemap foi desenvolvido para a visualização de estruturas de árvores complexas. Esta técnica particiona horizontalmente e verticalmente o espaço disponível em uma hierarquia de quadrado não oclusivos de acordo com o número de ramificações da árvore. Valores maiores (quadrado maiores) são localizados no canto superior esquerdo e valores menores (quadrados menores) são localizados no canto inferior direito da estrutura. Cada nível

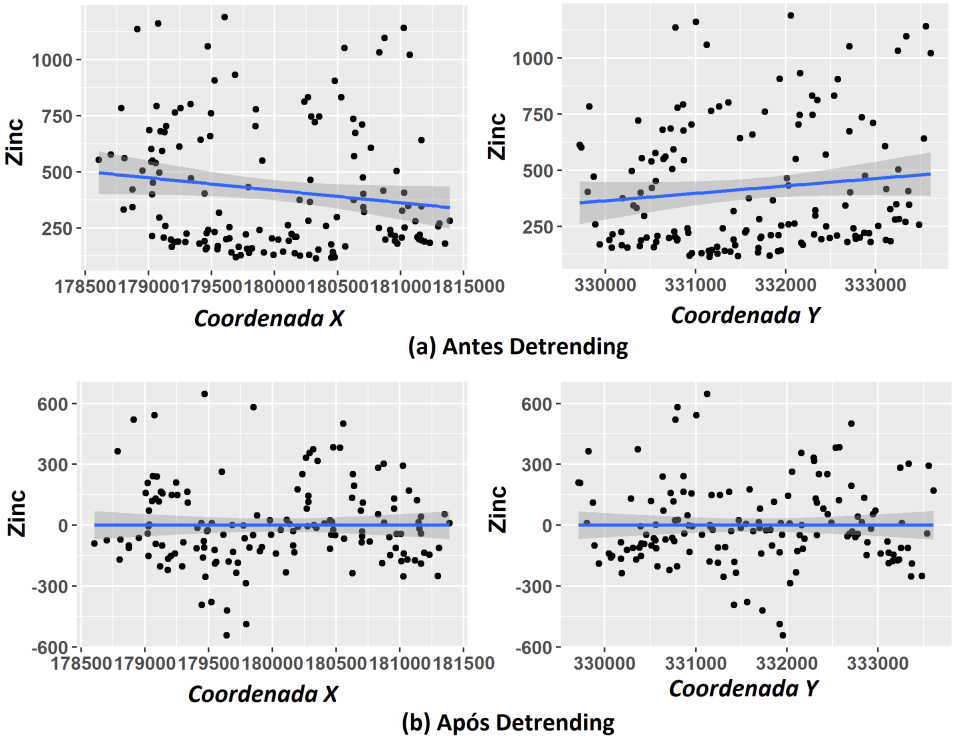

Figura 34. Base de dados Meuse. Valores de zinco nas coordenadas X e Y - (a) Antes do processo de *detrending* e (b) após o processo de *detrending*.

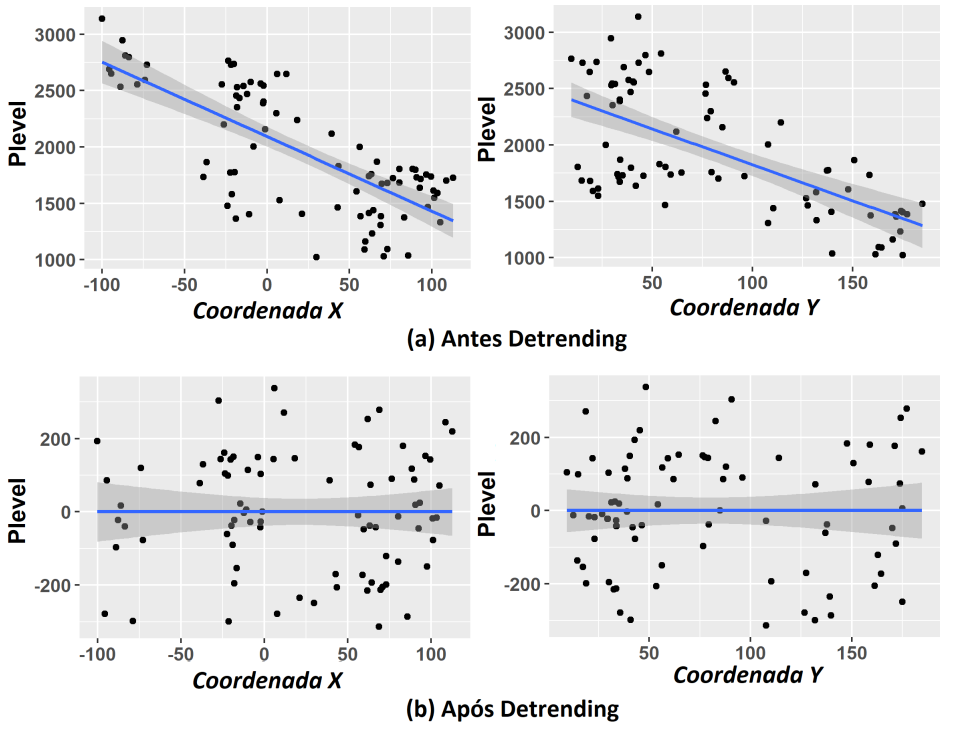

Figura 35. Base de dados Wolfcamp. Valores de nível piezométrico nas coordenadas X e Y - (a) Antes do processo de *detrending* e (b) após o processo de *detrending*.

Tabela 10. Porcentagem da ocorrência de *trends* nos grupos com base no índice de Mann-Kendall, considerando as configurações com 1 à 3 clusters.

		ClustGeo		
	Clusters	**1**	**2**	**3**
Sem Detrend	Meuse	30%	85%	70%
	Wolfcamp	100%	60%	84%
Com Detrend	Meuse	0%	50%	47%
	Wolfcamp	0%	0%	27%
		K-Means + KNN		
	Clusters	**1**	**2**	**3**
Sem Detrend	Meuse	30%	65%	44%
	Wolfcamp	100%	75%	100%
Com Detrend	Meuse	0%	25%	50%
	Wolfcamp	0%	10%	33%

hierárquico contém informações relevantes sobre uma variável, e este método utiliza diversas representações de dados como tamanho do quadrado, cor, rótulo entre outros para apresentação das informações (CARVALHO; MEIGUINS; MORAIS, 2016; SOARES et al., 2018).

A Figura 36 apresenta o treemap com os resultados dos experimentos realizados para as bases Meuse e Wolfcamp. Cada quadrado do treemap representa a média de 10 execuções, e seu tamanho reflete o índice NMSE. Os quatro melhores resultados de ambas as bases de dados foram destacados com uma borda vermelha no canto inferior direito do treemap. É possível inferir que os melhores resultados foram alcançados quando as etapas de *detrending* e clusterização são executadas em conjunto. Além disso, o AG proposto esteve presente nos melhores resultados dentre os algoritmos estudados. Em relação às técnicas de clusterização, ClustGeo e K-means+KNN, o desempenho do algoritmo híbrido foi compatível com os resultados do algoritmo consolidado na literatura ClustGeo, demonstrando a viabilidade da solução apresentada nesta tese.

As técnicas estatísticas, *One-Way Anova* para medidas repetidas e Teste T pareado (BOSLAUGH, 2012) foram aplicados com o objetivo de investigar quais configurações são estatisticamente melhores que as outras utilizadas nestes experimentos, considerando um intervalo de confiança de 95%. É conhecido que estes testes paramétricos requerem que os dados sejam modelados por uma distribuição normal, então, o teste de Shappiro-Wilk foi utilizado para este propósito (BOSLAUGH, 2012). Em relação ao teste *One-Way Anova* para medidas repetidas, se a análise de variância é significante baseada na correção de Greenhouse-Geisser (ABDI, 2010), é necessário utilizar um teste *Post-Hoc* para identificar quais amostras são diferentes, portanto, a correção de Bonferroni foi aplicada para este problema (BLAND; ALTMAN, 1995). Os resultados desta análise estatística são apresen-

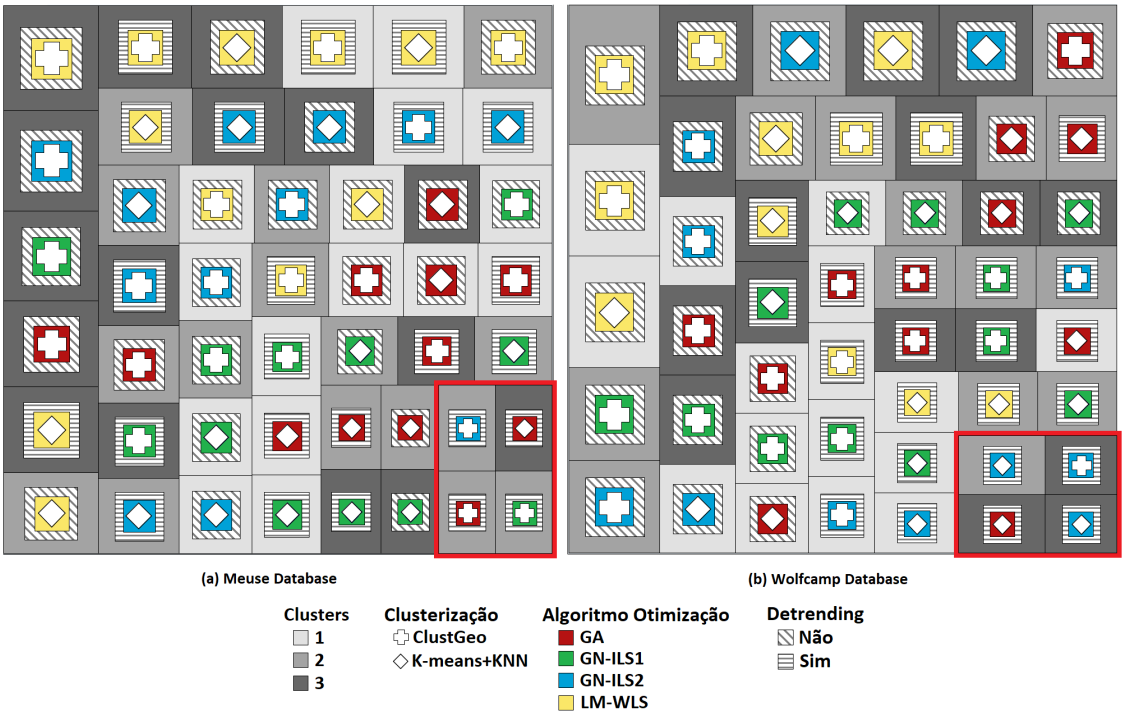

Figura 36. Representação treemap do índice NMSE nas diversas configurações testadas durante os experimentos utilizado as bases de dados Meuse e Wolfcamp.

tados e discutidos na próxima seção. Os testes de normalidade apresentaram distribuição normal para todas as variáveis.

5.5 Processo de *detrending*

O teste t pareado foi aplicado considerando a hipótese nula de que as médias obtidas com e sem o processo de detrending são iguais. Os resultados analisados demonstraramm que as médias diferem. Em outras palavras, a etapa de remoção de *trends* no pré-processamento dos dados produz um efeito positivo na acurácia do modelo, reduzindo o índice NMSE, e considerando as duas bases de dados: $t(479) = 2.658; p = 0.008 < 0.05$ para Meuse e $t(479) = 1.969; p = 0.049 < 0.05$ para Wolfcamp. As estatísticas descritivas são exibidas na Tabela 11, onde N indica o número de amostras, a Média dos valores NMSE obtidos e seus respectivos desvios padrão.

5.6 Clusterização dos dados

Para os dois algoritmos de clusterização avaliados, Kmeans+KNN e ClustGeo, o teste t pareado indicou que as médias não diferem significativamente, considerando ambas as bases: $t(239) = 0.607; p = 0.544 > 0.05$ para Meuse e $t(239) = 1.241; p = 0.216 > 0.05$ para Wolfcamp. Assim, pode-se concluir que a estratégia proposta de clusterização é

Tabela 11. Estatísticas descritivas com e sem a etapa de *detrending*.

		Meuse	
Algoritmo	**N**	**Média**	**Desv. Padrão**
Sem Detrend	240	0.055	0.074
Com Detrend	240	0.050	0.062
		Wolfcamp	
Algoritmo	**N**	**Média**	**Desv. Padrão**
Sem Detrend	240	0.011	0.013
Com Detrend	240	0.009	0.018

equivalente quando comparada à outra estratégia disponível na literatura. As estatísticas descritivas podem ser observadas na Tabela 12.

Tabela 12. Estatísticas descritivas dos algoritmos de clusterização.

		Meuse	
Algoritmo	**N**	**Média**	**Desv. Padrão**
ClustGeo	240	0.054	0.066
K-Means+KNN	240	0.051	0.070
		Wolfcamp	
Algoritmo	**N**	**Média**	**Desv. Padrão**
ClustGeo	240	0.011	0.021
K-Means+KNN	240	0.009	0.007

O teste One-Way Anova para medidas repetidas indicou que as médias diferem significativamente quando o número de *clusters* aumenta. Considerando ambas as bases: $F(1.712, 272.219) = 5.695; p = 0.006 < 0.05$ para Meuse e $F(1.179, 187.442) = 3.822; p = 0.045 < 0.05$ para Wolfcamp. As estatísticas descritivas e os resultados dos testes Post-Hoc estão descritos na Tabela 13 e na Tabela 14. De acordo com a significância calculada pelo teste de Bonferroni, pode-se concluir que, em média, as configurações do modelo com a etapa de clusterização obtiveram erros maiores que as configurações com 1 *cluster* (ou sem a etapa de clusterização), como pode ser visto na Tabela 10. Apesar deste resultado não ter sido esperado, devido ao alto desvio padrão nos resultados com clusterização, há a hipótese de que a presença de *trends* nos grupos criados pela etapa de clusterização, mesmo após os dados serem submetidos ao processo de *detrending*, aumentam o erro de forma significativa e possivelmente mascarando os melhores resultados obtidos em algumas iterações.

Tabela 13. Estatísticas descritivas considerando o número de clusters.

		Meuse	
#Cluster	**N**	**Média**	**Desv. Padrão**
1	160	0.040	0.034
2	160	0.056	0.080
3	160	0.063	0.079
		Wolfcamp	
#Cluster	**N**	**Média**	**Desv. Padrão**
1	160	0.007	0.004
2	160	0.010	0.026
3	160	0.012	0.009

Tabela 14. Resultados do teste Bonferroni considerando diferente número de clusters.

		Meuse	
#Cluster	**Diferença das Médias**	**Desv. Padrão**	**Sig.**
1-2	0.016	0.006	0.026
1-3	0.023	0.007	0.002
2-3	0.007	0.008	1.000
		Wolfcamp	
#Cluster	**Diferença das Médias**	**Desv. Padrão**	**Sig.**
1-2	0.003	0.002	0.373
1-3	0.005	0.001	0.000
2-3	0.002	0.002	1.000

5.7 Algoritmos de otimização

Estatisticamente, a utilização de diferentes algoritmos de otimização tem impacto no índice NMSE, corroborado pelo teste One-Way Anova para medidas repetidas, considerando ambas as bases de dados: $F(1.041, 123.907) = 43.374; p = 0.000 < 0.05$ para Meuse e $F(1.032, 122.840) = 3.956; p = 0.048 < 0.05$ para Wolfcamp. As estatísticas descritivas e os resultados do teste Post-Hoc são apresentados na Tabela 15 e na Tabela 16. De acordo com o test Post-Hoc Bonferroni, em média, os algoritmos que não fazem uso de parâmetros de anisotropia para o ajuste do modelo do variograma teórico, GN-ILS2 e LM-WLS, obtiveram erros maiores que os algoritmos que utilizam informações de anisotropia, AG e GN-ILS1. Este fenômeno ainda é mais visível na base de dados Meuse. Os parâmetros adicionais, ângulo e fator, inclusos nos algoritmos AG e GN-ILS1 são essenciais para a captura de anisotropia zonal, que pode ser ainda mais intensificada pela etapa de clusterização.

Tabela 15. Estatísticas descritivas considerando os algoritmos de otimização.

		Meuse	
Algoritmo	**N**	**Média**	**Desv. Padrão**
AG	120	0.034	0.024
GN-ILS1	120	0.034	0.025
GN-ILS2	120	0.040	0.031
LM-WLS	120	0.103	0.114
		Wolfcamp	
Algoritmo	**N**	**Média**	**Desv. Padrão**
AG	120	0.008	0.005
GN-ILS1	120	0.008	0.004
GN-ILS2	120	0.009	0.005
LM-WLS	120	0.014	0.030

Tabela 16. Resultados do teste Bonferroni com diferentes algoritmos de otimização.

		Meuse	
Algoritmo	**Diferença das Médias**	**Desv. Padrão**	**Sig.**
GA - GN-ILS1	0.000	0.001	1.000
GA - GN-ILS2	0.006	0.002	0.001
GA - LM-WLS	0.070	0.010	0.000
GN-ILS1 - GN-ILS2	0.006	0.001	0.000
GN-ILS1 - LM-WLS	0.070	0.010	0.000
GN-ILS2 - LM-WLS	0.064	0.010	0.000
		Wolfcamp	
Algoritmo	**Diferença das Médias**	**Desv. Padrão**	**Sig.**
GA - GN-ILS1	0.001	0.000	0.007
GA - GN-ILS2	0.000	0.000	1.000
GA - LM-WLS	0.005	0.003	0.338
GN-ILS1 - GN-ILS2	0.001	0.000	0.004
GN-ILS1 - LM-WLS	0.006	0.003	0.183
GN-ILS2 - LM-WLS	0.005	0.003	0.435

5.8 Mapas de krigagem

As estimativas da krigagem e os mapas de variância foram gerados com e sem a etapa de detrending, para ambas as bases de dados com os parâmetros do melhor resultado obtido, considerando todas as iterações realizadas nos testes. Para a base de dados Meuse, o melhor resultado foi alcançado utilizando K-means+KNN com a configuração de 3 clusters e para a base Wolfcamp, o melhor resultado foi alcançado utilizando ClustGeo com 2 clusters, ambos com o algoritmo de otimização AG. Esta verificação é importante para analisar se a utilização de métodos como a clusterização teria impacto negativo na continuidade

espacial na base de dados. De acordo com a Figura 37 e Figura 38 a uniformidade espacial é mantida em ambos os mapas. É importante notar também que a utilização do modelo proposto, com parâmetros de anisotropia (ângulo e fator) definidos para cada *cluster* individualmente, consegue captar anisotropia zonal, modelando-a corretamente.

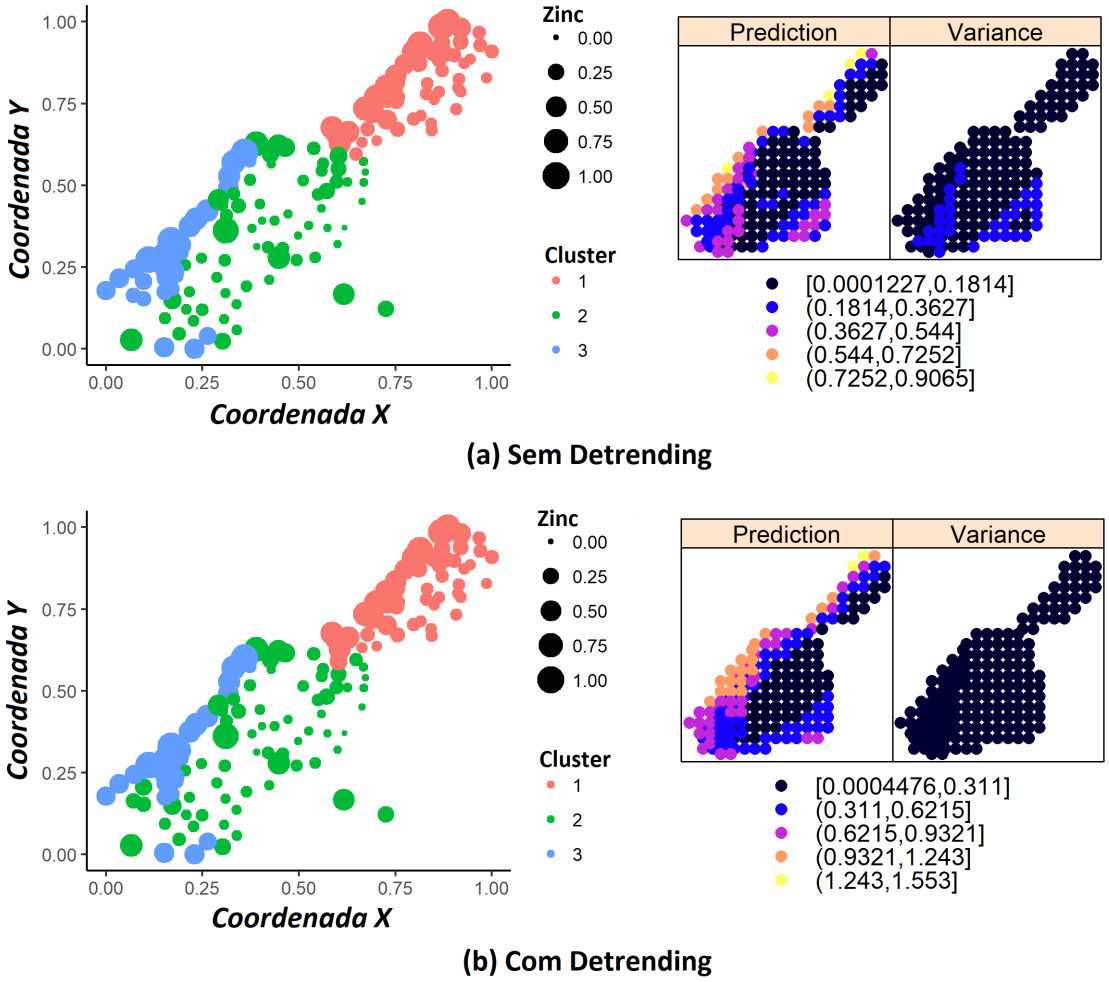

Figura 37. Mapas de krigagem, estimativas e variância, para a base de dados Meuse.

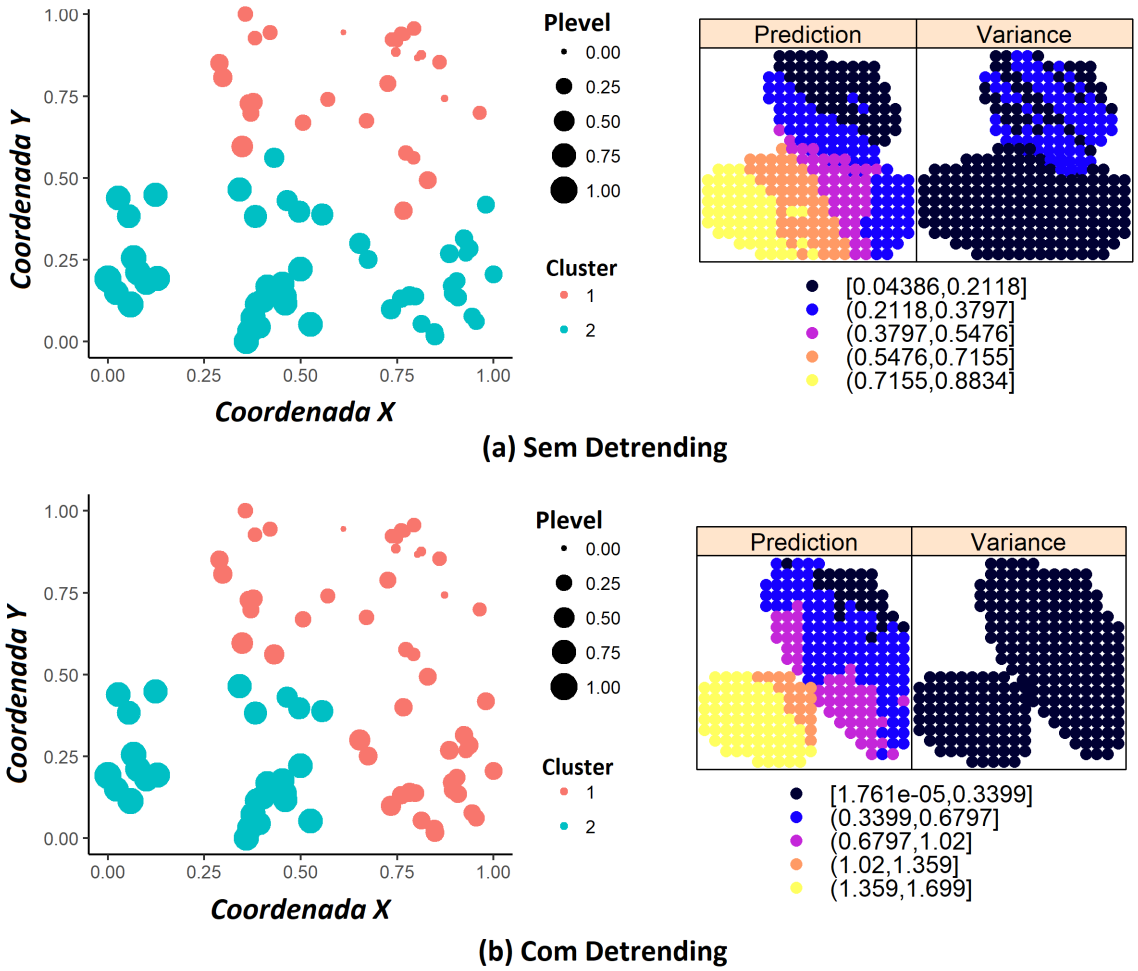

Figura 38. Mapas de krigagem, estimativas e variância, para a base de dados Wolfcamp.

6 Conclusões e trabalhos futuros

A modelagem e seleção dos parâmetros do variograma teórico ainda é considerado um desafio para aplicações da geostatística. Os resultados obtidos com a modelo proposto utilizando clusterização de dados, técnicas evolucionárias e parâmetros de anisotropia, atingiram taxas de erros menores em comparação com outras proposta tradicionais no processo de krigagem. Diante destes resultados, podemos inferir que as contribuições alcançadas por esta tese são:

- Uma configuração híbrida (K-means + KNN) foi aplicada com sucesso na diminuição do problema da sobreposição de clusters. Este algoritmo proposto atingiu resultados semelhantes ao ClustGeo, outro algoritmo de clusterização espacial. O primeiro algoritmo apresentou melhores resultados na base de dados Meuse, enquanto o segundo algoritmo de clusterização atingiu melhores resultados na base de dados Wolfcamp.
- Para cada cluster definido na etapa de clusterização foi definido um AG específico e consequentemente um conjunto de parâmetros do variograma teórico, diferentemente de outros trabalhos, como em (ABEDINI; NASSERI; ANSARI, 2008), que possuía um único modelo para todos os grupos.
- Estimativa automática dos parâmetros que definem o variograma teórico de acordo com as características da base de dados, que mitiga a necessidade de conhecimento especialista.
- A tarefa de clusterização de dados e a estimativa de parâmetros específicos do variograma para cada clusters, podem ser considerados procedimentos que melhoram a acurácia do processo de krigagem.
- Uma discussão inicial sobre os impactos da clusterização na hipótese da estacionariedade também é apresentada nesta tese. A remoção de *trends* da base de dados original é benéfica, em alguns casos, em relação aos clusters criados posteriormente. Este tópico necessita de uma investigação mais aprofundada para a construção de técnicas de agrupamento que garantam a hipótese da estacionariedade de forma automatica.

Com base nos experimentos realizados e nos resultados apresentados, podemos concluir que o modelo proposto é válido para a produção de modelos e parâmetros do variograma teórico com qualidade satisfatória em comparação com outras abordagens disponíveis na literatura. Contudo, é imprescindível dizer que há espaço para mais aprimoramentos e pesquisas nessa área. Neste sentido, como sugestão para trabalhos futuros, podemos citar a avaliação de algoritmos em que a hipótese da estacionariedade é garantida,

como propõe (FOUEDJIO, 2017); utilizar métricas de avaliação da qualidade dos grupos formados, de forma a identificar a quantidade ideal de *clusters* para determinado problema; Aprimoramento do método de classificação de pontos desconhecidos utilizando informações de anisotropia no algoritmo KNN; Utilização de técnicas fuzzy como tentativa de evitar problemas de sobreposição de clusters.

Outro ponto de atenção é a avaliação de outros algoritmos bioinspirados, como *Differential Evolution* (PRICE, 2013), Colônia de abelhas (KARABOGA; BASTURK, 2007), entre outras. Por fim, é interessante a utilização de *tuners* de meta-heurísticas como CRS-Tuning (VEČEK et al., 2016), F-Race (BIRATTARI et al., 2002) e outros (TRINDADE; CAMPELO, 2019) para parâmetros de controle dos algoritmos bioinspirados.

Referências

ABDI, H. The greenhouse-geisser correction. *Encyclopedia of research design*, Sage Thousand Oaks, CA, v. 1, p. 544–548, 2010. Citado na página 72.

ABEDINI, M.; NASSERI, M.; ANSARI, A. Cluster-based ordinary kriging of piezometric head in west texas/new mexico–testing of hypothesis. *Journal of Hydrology*, Elsevier, v. 351, n. 3-4, p. 360–367, 2008. Citado 11 vezes nas páginas 9, 18, 19, 45, 46, 47, 48, 53, 63, 69 e 79.

ABEDINI, M.; NASSERI, M.; BURN, D. The use of a genetic algorithm-based search strategy in geostatistics: application to a set of anisotropic piezometric head data. *Computers & Geosciences*, Elsevier, v. 41, p. 136–146, 2012. Citado 7 vezes nas páginas 18, 20, 45, 46, 48, 67 e 68.

ACTION, P. *Teste de Shapiro-Wilk*. 2017. Citado 2 vezes nas páginas 36 e 38.

AMRI, N. A.; JEMAIN, A. A.; HASSAN, W. F. W. Kriging on comparison of original and outlier-free data. In: AIP. *AIP Conference Proceedings*. [S.l.], 2014. v. 1614, p. 929–935. Citado na página 53.

APARNA, K.; NAIR, M. K. Effect of outlier detection on clustering accuracy and computation time of chb k-means algorithm. In: *Computational Intelligence in Data Mining—Volume 2*. [S.l.]: Springer, 2016. p. 25–35. Citado na página 53.

ARCIDIACONO, G. et al. A kriging modeling approach applied to the railways case. *Procedia Structural Integrity*, Elsevier, v. 8, p. 163–167, 2018. Citado 3 vezes nas páginas 45, 47 e 48.

BAAR, J. D.; DWIGHT, R. P.; BIJL, H. Speeding up kriging through fast estimation of the hyperparameters in the frequency-domain. *Computers & geosciences*, Elsevier, v. 54, p. 99–106, 2013. Citado 3 vezes nas páginas 45, 47 e 48.

BALU, R.; ULAGANATHAN, S.; ASPROULIS, N. Effect of variogram types on surrogate model based optimisation of aircraft wing shapes. *Procedia engineering*, Elsevier, v. 38, p. 2713–2725, 2012. Citado 2 vezes nas páginas 45 e 48.

BARGAOUI, Z. K.; CHEBBI, A. Comparison of two kriging interpolation methods applied to spatiotemporal rainfall. *Journal of Hydrology*, Elsevier, v. 365, n. 1-2, p. 56–73, 2009. Citado 2 vezes nas páginas 45 e 48.

BIRATTARI, M. et al. A racing algorithm for configuring metaheuristics. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*. [S.l.], 2002. p. 11–18. Citado 2 vezes nas páginas 68 e 80.

BLAND, J. M.; ALTMAN, D. G. Multiple significance tests: the bonferroni method. *Bmj*, British Medical Journal Publishing Group, v. 310, n. 6973, p. 170, 1995. Citado na página 72.

BOSLAUGH, S. *Statistics in a nutshell: A desktop quick reference*. [S.l.]: "O'Reilly Media, Inc.", 2012. Citado 2 vezes nas páginas 53 e 72.

CARVALHO, M. B. de; MEIGUINS, B. S.; MORAIS, J. M. de. Temporal data visualization technique based on treemap. In: IEEE. *2016 20th International Conference Information Visualisation (IV)*. [S.l.], 2016. p. 399–403. Citado na página 72.

CHAVENT, M. et al. Clustgeo: an r package for hierarchical clustering with spatial constraints. *Computational Statistics*, Springer, v. 33, n. 4, p. 1799–1822, 2018. Citado 4 vezes nas páginas 19, 20, 47 e 54.

CLARK, I. *Practical geostatistics*. [S.l.]: Applied Science Publishers London, 1979. v. 3. Citado na página 62.

CRESSIE, N. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, Springer, v. 17, n. 5, p. 563–586, 1985. Citado 5 vezes nas páginas 24, 46, 48, 62 e 67.

DEEP, K.; THAKUR, M. A new crossover operator for real coded genetic algorithms. *Applied mathematics and computation*, Elsevier, v. 188, n. 1, p. 895–911, 2007. Citado 2 vezes nas páginas 55 e 64.

DEEP, K.; THAKUR, M. A new mutation operator for real coded genetic algorithms. *Applied mathematics and Computation*, Elsevier, v. 193, n. 1, p. 211–230, 2007. Citado 3 vezes nas páginas 31, 55 e 64.

DESASSIS, N.; RENARD, D. Automatic variogram modeling by iterative least squares: univariate and multivariate cases. *Mathematical Geosciences*, Springer, v. 45, n. 4, p. 453–470, 2013. Citado 3 vezes nas páginas 18, 67 e 68.

FOUEDJIO, F. A spectral clustering approach for multivariate geostatistical data. *International Journal of Data Science and Analytics*, Springer, v. 4, n. 4, p. 301–312, 2017. Citado 2 vezes nas páginas 19 e 80.

GIBBONS, J. D.; FIELDEN, J. D. G. *Nonparametric statistics: An introduction*. [S.l.]: Sage, 1993. Citado na página 65.

GOLDBERG, D. E.; HOLLAND, J. H. Genetic algorithms and machine learning. *Machine learning*, Springer, v. 3, n. 2, p. 95–99, 1988. Citado 2 vezes nas páginas 29 e 31.

GONÇALVES, Í. G.; KUMAIRA, S.; GUADAGNIN, F. A machine learning approach to the potential-field method for implicit modeling of geological structures. *Computers & Geosciences*, Elsevier, v. 103, p. 173–182, 2017. Citado 4 vezes nas páginas 18, 45, 47 e 48.

HENGL, T. *A practical guide to geostatistical mapping*. [S.l.]: Hengl, 2009. v. 52. Santa Clara, USA. ISBN 630813-054US. Citado 3 vezes nas páginas 16, 23 e 25.

HUIZAN, W. et al. Improved kriging interpolation based on support vector machine and its application in oceanic missing data recovery. In: IEEE. *Computer Science and Software Engineering, 2008 International Conference on*. [S.l.], 2008. v. 4, p. 726–729. Citado na página 18.

JAKOB, A. A. E.; YOUNG, A. F. O uso de métodos de interpolação espacial de dados nas análises sociodemográficas. *Anais*, p. 1–22, 2016. Citado 2 vezes nas páginas 9 e 17.

KARABOGA, D.; BASTURK, B. Artificial bee colony (abc) optimization algorithm for solving constrained optimization problems. In: SPRINGER. *International fuzzy systems association world congress*. [S.l.], 2007. p. 789–798. Citado na página 80.

KERRY, R.; OLIVER, M. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma*, Elsevier, v. 140, n. 4, p. 383–396, 2007. Citado na página 26.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004. Citado na página 39.

LARRONDO, P. F.; NEUFELD, C. T.; DEUTSCH, C. V. Varfit: A program for semiautomatic variogram modeling. In: . [S.l.: s.n.], 2003. Citado 2 vezes nas páginas 18 e 68.

LI, S.; LU, W. Automatic fit of the variogram. In: IEEE. *Information and Computing (ICIC), 2010 Third International Conference on*. [S.l.], 2010. v. 4, p. 129–132. Citado na página 18.

LI, Z. et al. An automatic variogram modeling method with high reliability fitness and estimates. *Computers & geosciences*, Elsevier, v. 120, p. 48–59, 2018. Citado 11 vezes nas páginas 18, 20, 24, 26, 33, 45, 46, 47, 48, 67 e 68.

MARÔCO, J. *Análise Estatística com o SPSS Statistics.: 7ª edição*. [S.l.]: ReportNumber, Lda, 2018. Citado 2 vezes nas páginas 35 e 36.

MARQUARDT, D. W. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, SIAM, v. 11, n. 2, p. 431–441, 1963. Citado 4 vezes nas páginas 18, 46, 48 e 67.

MASOOMI, Z.; MESGARI, M. S.; MENHAJ, M. B. Modeling uncertainties in sodium spatial dispersion using a computational intelligence-based kriging method. *Computers & geosciences*, Elsevier, v. 37, n. 10, p. 1545–1554, 2011. Citado 6 vezes nas páginas 45, 47, 48, 62, 67 e 68.

MCLEOD, A. I. Kendall rank correlation and mann-kendall trend test. *R Package Kendall*, Western Univ., 2005. Citado na página 70.

MISHRA, K.; TIWARI, S.; MISRA, A. A bio inspired algorithm for solving optimization problems. In: IEEE. *2011 2nd International Conference on Computer and Communication Technology (ICCCT-2011)*. [S.l.], 2011. p. 653–659. Citado na página 32.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. [S.l.]: Editora Saraiva, 2017. Citado na página 37.

NOURI, N. M.; MOHAMMADI, S.; ZAREZADEH, M. Optimization of a marine contra-rotating propellers set. *Ocean Engineering*, Elsevier, v. 167, p. 397–404, 2018. Citado 2 vezes nas páginas 45 e 48.

OLEA, R. A. *Geostatistics for engineers and earth scientists*. [S.l.]: Springer Science & Business Media, 2012. Citado na página 24.

OLIVER, M.; WEBSTER, R. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena*, Elsevier, v. 113, p. 56–69, 2014. Citado na página 26.

OLSSON, D. M.; NELSON, L. S. The nelder-mead simplex procedure for function minimization. *Technometrics*, Taylor & Francis Group, v. 17, n. 1, p. 45–51, 1975. Citado na página 48.

PEBESMA, E. J. Multivariable geostatistics in r: the gstat package. *Computers & Geosciences*, Elsevier, v. 30, n. 7, p. 683–691, 2004. Citado na página 67.

PESQUER, L.; CORTÉS, A.; PONS, X. Parallel ordinary kriging interpolation incorporating automatic variogram fitting. *Computers & Geosciences*, Elsevier, v. 37, n. 4, p. 464–473, 2011. Citado 4 vezes nas páginas 18, 45, 46 e 48.

POHLERT, T. Non-parametric trend tests and change-point detection. *CC BY-ND*, v. 4, 2016. Citado na página 20.

POLI, R.; KENNEDY, J.; BLACKWELL, T. Particle swarm optimization. *Swarm intelligence*, Springer, v. 1, n. 1, p. 33–57, 2007. Citado na página 32.

PRICE, K. V. Differential evolution. In: *Handbook of Optimization*. [S.l.]: Springer, 2013. p. 187–214. Citado na página 80.

RUSSELL, R. et al. Issues and challenges in conducting systematic reviews to support development of nutrient reference values: workshop summary: nutrition research series, vol. 2. Agency for Healthcare Research and Quality (US), Rockville (MD), 2009. Citado na página 39.

SCRUCCA, L. et al. Ga: a package for genetic algorithms in r. *Journal of Statistical Software*, Citeseer, v. 53, n. 4, p. 1–37, 2013. Citado na página 55.

SOARES, A. G. M. et al. Visualizing multidimensional data in treemaps with adaptive glyphs. In: IEEE. *2018 22nd International Conference Information Visualisation (IV)*. [S.l.], 2018. p. 58–63. Citado na página 72.

TRINDADE, Á. R.; CAMPELO, F. Tuning metaheuristics by sequential optimisation of regression models. *Applied Soft Computing*, Elsevier, v. 85, p. 105829, 2019. Citado 2 vezes nas páginas 68 e 80.

VASAT, R.; HEUVELINK, G.; BORVKA, L. Sampling design optimization for multivariate soil mapping. *Geoderma*, Elsevier, v. 155, n. 3-4, p. 147–153, 2010. Citado 3 vezes nas páginas 45, 46 e 48.

VEČEK, N. et al. Parameter tuning with chess rating system (crs-tuning) for meta-heuristic algorithms. *Information Sciences*, Elsevier, v. 372, p. 446–469, 2016. Citado 2 vezes nas páginas 68 e 80.

VIALI, L. Testes de hipóteses nao paramétricos. 2008. *Apostila-UFRGS. Disponıvel em:< http://www. mat. ufrgs. br/ viali/estatistica/mat2282/material/apostilas/Testes Nao Parametricos. pdf>. Acesso em*, v. 7, 2017. Citado 2 vezes nas páginas 34 e 35.

VIEIRA, S. R. et al. Detrending non stationary data for geostatistical applications. *Bragantia*, SciELO Brasil, v. 69, p. 01–08, 2010. Citado 3 vezes nas páginas 19, 27 e 53.

WANG, H. et al. Time complexity reduction in efficient global optimization using cluster kriging. In: ACM. *Proceedings of the Genetic and Evolutionary Computation Conference*. [S.l.], 2017. p. 889–896. Citado 5 vezes nas páginas 19, 45, 46, 47 e 48.

WANG, Z. et al. Optimization of riveting parameters using kriging and particle swarm optimization to improve deformation homogeneity in aircraft assembly. *Advances in Mechanical Engineering*, SAGE Publications Sage UK: London, England, v. 9, n. 8, p. 1687814017719003, 2017. Citado 4 vezes nas páginas 18, 45, 47 e 48.

WEI, Z.; LIU, Z.; CHEN, Q. Ga-based kriging for isoline drawing. In: IEEE. *Environmental Science and Information Application Technology (ESIAT), 2010 International Conference on*. [S.l.], 2010. v. 2, p. 170–173. Citado 5 vezes nas páginas 18, 45, 48, 67 e 68.

WEISSTEIN, E. W. Modified bessel function of the second kind. *From MathWorld–A Wolfram Web Resource.*, Wolfram Research, Inc., 2002. Citado na página 24.

WITTEN, I. H. et al. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016. Citado 2 vezes nas páginas 27 e 29.

XIALIN, Z. et al. An intelligent improvement on the reliability of ordinary kriging estimates by a ga. In: IEEE. *Intelligent Systems (GCIS), 2010 Second WRI Global Congress on*. [S.l.], 2010. v. 2, p. 61–64. Citado 6 vezes nas páginas 18, 20, 45, 48, 67 e 68.

YASOJIMA, C. et al. A comparison of genetic algorithms and particle swarm optimization to estimate cluster-based kriging parameters. In: SPRINGER. *EPIA Conference on Artificial Intelligence*. [S.l.], 2019. p. 750–761. Citado na página 20.

YASOJIMA, E. K. K. et al. Cam-adx: A new genetic algorithm with increased intensification and diversification for design optimization problems with real variables. *Robotica*, Cambridge University Press, p. 1–46, 2019. Citado na página 34.

ZHANG, Y. Introduction to geostatistics—course notes. *Dept. of Geology & Geophysics, University of Wyoming*, 2011. Citado na página 17.