ITALO OLIVEIRA FERREIRA

CONTROLE DE QUALIDADE EM LEVANTAMENTOS HIDROGRÁFICOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Civil, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
F383c
2018

Ferreira, Italo Oliveira, 1988-
Controle de qualidade em levantamentos hidrográficos / Italo Oliveira Ferreira. – Viçosa, MG, 2018.
xv, 216f. : il. (algumas color.) ; 29 cm.

Inclui apêndices.
Orientador: Afonso de Paula dos Santos.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Levantamentos hidrográficos. 2. Controle de qualidade. I. Universidade Federal de Viçosa. Departamento de Engenharia Civil. Programa de Pós-graduação em Engenharia Civil. II. Título.

CDD 22 ed. 526.99

ITALO OLIVEIRA FERREIRA

CONTROLE DE QUALIDADE EM LEVANTAMENTOS HIDROGRÁFICOS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Civil, para obtenção do título de *Doctor Scientiae*.

APROVADA: 02 de fevereiro de 2018.

João Marcos Louzada

Éder Teixeira Marques

Júlio César de Oliveira
(Coorientador)

Paulo César Emiliano
(Coorientador)

Afonso de Paula dos Santos
(Orientador)

"Os que se encantam com a prática sem a ciência são como os timoneiros que entram no navio sem timão nem bússola, nunca tendo certeza do seu destino".

Leonardo da Vinci

"Restará sempre muito o que fazer..."

AGRADECIMENTOS

Principalmente a Deus, por ter me concedido o dom da vida, inteligência, paciência, capacidade de crescimento e, sobretudo, persistência e coragem para sempre seguir em frente.

Aos meus pais, Geraldo Magela e Eliane, e aos meus irmãos, Danilo e Murilo, que nunca deixaram de me dar força, acreditando mais em mim do que eu mesmo e me incentivando sempre a trilhar o caminho do bem.

À minha namorada, Laís, por todo amor, carinho, cuidado, paciência, incentivo, cumplicidade e, acima de tudo, compreensão nos momentos de ausência ocasionados pela dedicação, quase que exclusiva, a esse trabalho de tese. De coração, muito obrigado!

A toda a minha família, principalmente aos meus avós, vó Nenê, vô Caetano *(in memoriam)*, vó Maria e vô Cici *(in memoriam).* Vocês sempre estarão comigo!

À Universidade Federal Viçosa, por todos os conhecimentos adquiridos e por toda a infraestrutura disponibilizada na minha graduação, mestrado, doutorado e em meu ambiente de trabalho.

A todos os professores do setor de Engenharia de Agrimensura e Cartográfica, pelos ensinamentos, pela ótima convivência durante todos esses anos e pela ajuda nesse período de treinamento.

Ao professor Afonso dos Santos, pelos ensinamentos, contribuições e orientação no desenvolvimento deste trabalho.

Ao professor Júlio César, pela amizade, dedicação ao trabalho e a grande ajuda na implementação dos algoritmos em ambiente R. Sem o senhor, não teríamos chegado tão longe!

Ao professor Paulo Emiliano, pelas contribuições e ensinamentos na área de estatística e pelo tempo disponibilizado, principalmente, nos momentos de reflexão sobre os resultados alcançados.

À professora Nilcilene Medeiros, pelas discussões e tempo dedicado a esta tese.

Ao professor Gérson Santos, pela ajuda nas análises geoestatísticas.

E a todos aqueles que, de forma direta ou indireta, contribuíram para a realização deste trabalho.

SUMÁRIO

LISTA DE FIGURAS vii

LISTA DE TABELAS x

LISTA DE ABREVIATURAS E SIGLAS xii

RESUMO xiv

ABSTRACT xv

INTRODUÇÃO GERAL 1

1. HIPÓTESES 18
2. OBJETIVOS 19
3. JUSTIFICATIVAS E IMPORTÂNCIA 20
4. ESTRUTURAÇÃO DO TRABALHO 21

REFERÊNCIAS BIBLIOGRÁFICAS 22

CAPÍTULO 1. METODOLOGIA ROBUSTA PARA DETECÇÃO DE *SPIKES* EM DADOS BATIMÉTRICOS 28

1. INTRODUÇÃO 28
2. PRINCIPAIS FONTES DE *SPIKES* EM SONDAGENS BATIMÉTRICAS 33
3. MÉTODOS PARA DETECÇÃO DE *OUTLIERS* EM CONJUNTOS DE DADOS UNIVARIADOS 35
4. PROPOSIÇÃO DO MÉTODO 38
5. EXPERIMENTOS E RESULTADOS 47
   - 5.1. Aplicação do método proposto em dados simulados 47
   - 5.2. Aplicação do método proposto em dados reais 53
6. CONCLUSÕES 64

REFERÊNCIAS BIBLIOGRÁFICAS 65

CAPÍTULO 2. PROPOSTA METODOLÓGICA PARA AVALIAÇÃO DA QUALIDADE VERTICAL DE SONDAGENS BATIMÉTRICAS

MONOFEIXE, COM ÊNFASE EM TESTES DE NORMALIDADE E INDEPENDÊNCIA 71

1. INTRODUÇÃO 71

2. PROPOSIÇÃO DO MÉTODO 77

2.1. Amostra independente e normal 80

2.2. Amostra independente e não-normal 85

2.3. Amostra dependente 88

2.4. Classificação do Levantamento Hidrográfico 90

3. EXPERIMENTOS E RESULTADOS 91

3.1. Reservatório 1 94

3.2. Reservatório 2 101

4. CONCLUSÕES 108

REFERÊNCIAS BIBLIOGRÁFICAS 110

CAPÍTULO 3. PROPOSTA METODOLÓGICA PARA AVALIAÇÃO DA QUALIDADE VERTICAL DE DADOS BATIMÉTRICOS COLETADOS A PARTIR DE SISTEMAS DE SONDAGEM POR FAIXA 116

1. INTRODUÇÃO 116

2. PROPOSIÇÃO DO MÉTODO 124

3. EXPERIMENTOS E RESULTADOS 130

4. CONCLUSÕES 148

REFERÊNCIAS BIBLIOGRÁFICAS 150

CONCLUSÕES GERAIS 154

APÊNDICES 158

Capítulo 1 158

a) Algoritmo da metodologia AEDO 158

Capítulo 2 168

a) Algoritmo da MAIB 168

Capítulo 3 186

a) Algoritmo do Método PP 186

b) Semivariogramas omnidirecionais experimentais e os modelos ajustados .. 190

c) Relatórios das análises geoestatísticas 192

d) Apresentação dos *outliers* detectados nas amostras de discrepâncias por meio do emprego do AEDO/Método δ 197

e) Análise gráfica exploratória das discrepâncias (dados sem *outliers*)..... 205

f) Análise de independência – Semivariogramas das discrepâncias (dados sem *outliers*). 210

LISTA DE FIGURAS

CAPÍTULO 1

Figura 1 – Perfil batimétrico eivado de *spikes*. 30
Figura 2 – Gráfico *boxplot* construído com base em um conjunto de dados sem *outliers*. 36
Figura 3 – Gráfico *boxplot* construído com base em um conjunto de dados com possíveis *outliers*. 36
Figura 4 – Valores de corte para detecção de *outliers* via gráfico *boxplot*. A área sombreada corresponde aos valores aceitáveis. As marcações em vermelho ilustram $Q1$ e $Q3$.. 37
Figura 5 – Fluxograma da metodologia proposta para detecção de *spikes* em dados de batimetria coletados a partir de sistemas de sondagem por varrimento. 39
Figura 6 – Exemplo de semivariogramas para dados espacialmente dependentes (a) e espacialmente independentes (b). 40
Figura 7 – Exemplo da técnica de Segmentação em Círculos empregada pelo AEDO. 44
Figura 8 – Superfície batimétrica tridimensional construída através de simulação computacional. 47
Figura 9 – *Spikes* detectados pelo AEDO a partir dos limiares do *Z-Score Modificado*, para um $P_{limiar} = 0{,}5$ (a) e um $P_{limiar} = 0{,}8$. 49
Figura 10 – *Spikes* implantados e *Spikes* detectados pelo AEDO a partir dos limiares das técnicas *Boxplot Ajustado* (a), *Z-Score Modificado* (b) e *Método* δ (c). 53
Figura 11 – Área de execução do Levantamento Hidrográfico (LH) e área de estudo. 55
Figura 12 – Análise gráfica exploratória. 56
Figura 13 – Nuvem de pontos da área em estudo. 57
Figura 14 – Semivariogramas experimentais com 75%, 50% e 25% da distância máxima. 58
Figura 15 – *Spikes* detectados pelo AEDO a partir dos limiares das técnicas *Boxplot Ajustado* (a), *Z-Score Modificado* (b) e *Método* δ (c). 59
Figura 16 – *Spikes* detectados através de processamento manual. 60

Figura 17 – *Spikes* excluídos através do processamento manual e *Spikes* detectados pelo AEDO a partir dos limiares das técnicas *Boxplot Ajustado* (a), *Z-Score Modificado* (b) e *Método* δ (c). 63

CAPÍTULO 2

Figura 1 – Fluxograma da metodologia proposta para avaliação intervalar da incerteza vertical em dados de sondagem batimétrica monofeixe. 77
Figura 2 – Exemplo de semivariogramas para dados espacialmente dependentes (a) e espacialmente independentes (b)..... 79
Figura 3 – Curva de dispersão padronizada para observações de profundidade..... 82
Figura 4 – Área de estudo 91
Figura 5 – Linhas Regulares de Sondagem (LRS) e Linhas de Verificação (LV)..... 93
Figura 6 – Interseções previstas e executas para o Reservatório 1 e Reservatório 2. 94
Figura 7 – Análise gráfica exploratória. 95
Figura 8 – Análise gráfica exploratória após a eliminação dos *outliers*. 96
Figura 9 – Semivariograma das discrepâncias para distância de 440m (18% da distância máxima). 98
Figura 10 – Histograma e gráficos *Q-Q Plot* da amostra *Bootstrap* dos estimadores: Φ, $RMSE$ e $\Phi_{Robusta}$ 99
Figura 11 – Análise gráfica exploratória. 102
Figura 12 – Análise gráfica exploratória após a eliminação dos *outliers*. 103
Figura 13 – Histograma dos dados brutos com 45 classes. 104
Figura 14 – Semivariogramas experimentais com 75%, 50% e 25% da distância máxima. 105
Figura 15 – Semivariograma das discrepâncias para distância de 1288m sobreposto ao envelope de Monte Carlo (50% da distância máxima). 105
Figura 16 – Agrupamentos e distribuição dos agrupamentos obtidos pelo algoritmo *k-medoids.* 106

CAPÍTULO 3

Figura 1 – Problemática da utilização apenas do modelo preditivo para determinação da incerteza..... 120
Figura 2 – Técnicas para obtenção das amostras de discrepâncias de levantamentos hidrográficos realizados a partir de sistemas de sondagem por faixa. 125

Figura 3 – Esquema do planejamento de um levantamento multifeixe com sobreposição entre linhas adjacentes de 100%..... 128
Figura 4 – Geometria do processo de formação de feixes de um sistema de sondagem multifeixe. 129
Figura 5 – Método proposto para avaliação intervalar da incerteza vertical em dados de sondagem por varrimento. 130
Figura 6 – Área de estudo. 132
Figura 7 – Histograma das bases de dados analisadas. 138
Figura 8 – Histograma das bases de dados geradas a partir do método PP. 139

LISTA DE TABELAS

INTRODUÇÃO GERAL

Tabela 1 – Padrões mínimos para Levantamentos Hidrográficos. 3

CAPÍTULO 1

Tabela 1 – Estatística descritiva da área de estudo simulada. 48
Tabela 2 – Análise da probabilidade da profundidade i ser um *spike* ($P_{limiar} = 0{,}5$). 49
Tabela 3 – *Spike*s inseridos aleatoriamente ao conjunto de dados. 50
Tabela 4 – Resultado do processamento dos dados da área de estudo simulada. 51
Tabela 5 – $P_{outlier(\%)}$ dos dados da área de estudo. 52
Tabela 6 – Estatística descritiva da área de estudo. 55
Tabela 7 – Profundidades válidas assinaladas como *spikes* pelo Método δ. 61

CAPÍTULO 2

Tabela 1 – Resumos das estatísticas empregadas pela MAIB. 90
Tabela 2 – Resumo dos padrões mínimos para Levantamentos Hidrográficos. 91
Tabela 3 – Estatística descritiva da área de estudo. 95
Tabela 4 – Estimativa pontual da incerteza vertical amostral. 97
Tabela 5 – Incerteza vertical amostral ao nível de confiança de 95% e viés das amostras *Bootstrap*. 100
Tabela 6 – Tolerâncias estipuladas para o Levantamento Hidrográfico da área de estudo e classificação via exame tradicional (profundidade média: 3,315 metros). 101
Tabela 7 – Estatística descritiva da área de estudo. 102
Tabela 8 – Estimativa pontual da incerteza vertical amostral. 103
Tabela 9 – Incerteza vertical amostral (TCL) ao nível de confiança de 95%. 107
Tabela 10 – Incerteza vertical amostral (robusta) ao nível de confiança de 95%. ... 107
Tabela 11 – Tolerâncias estipuladas para o Levantamento Hidrográfico da área de estudo e classificação via exame tradicional (profundidade média: 5,246 metros). 108

CAPÍTULO 3

Tabela 1 – Resumo dos padrões mínimos para Levantamentos Hidrográficos. 120

Tabela 2 – Informações gerais dos dados da área de estudo. 132
Tabela 3 – Resultados da análise geoestatística. 133
Tabela 4 – Síntese dos resultados obtidos por meio da aplicação dos métodos PP, SS e SP. 134
Tabela 5 – Resultados da detecção de *outliers* via metodologia AEDO..... 135
Tabela 6 – Estatística descritiva da área de estudo. 136
Tabela 7 - Análise tradicional do Levantamento Hidrográfico e estimativa da $\Phi_{Robusta}$ (profundidade média: 15,600 metros). 140
Tabela 8 – Intervalo da incerteza vertical amostral gerada com base na metodologia proposta. 144

LISTA DE ABREVIATURAS E SIGLAS

AEDO – Algoritmo Espacial para Detecção de *Outliers*

AJU – Águas Jurisdicionais Brasileiras

ANA – Agência Nacional de Águas

ANEEL – Agência Nacional de Energia Elétrica

CHM – Centro de Hidrografia da Marinha

CHS – *Canadian Hydrographic Service*

CUBE – *Combined Uncertainty and Bathymetry Estimator*

BCa – *Biased Corrected Accelerated*

DHN – Diretoria de Hidrografia e Navegação

ENC – *Eletronic Navigation Chart*

FGDC – *Federal Geographic Data Committee*

GNSS – *Global Navigation Satellite System*

IC – Intervalos de Confiança

IHO – *International Hydrographic Organization*

IHT – Incerteza Horizontal Total

INMETRO – Instituto Nacional de Metrologia Normalização, Qualidade e Tecnologia

I2NS – *Integrated Inertial Navigation System*

IODA – *Intelligent Outlier Detection Algorithm*

IPT – Incerteza Propagada Total

IVT – Incerteza Vertical Total

KS - *Kolmogorov-Smirnov*

LiDAR – Light Detection And Ranging

LINZ – *Land Information New Zealand*

LRS – Faixa Regular de Sondagem

LTS – *Least Trimmed Squares*

LV – Faixa de Verificação

MAIB – Metodologia para Avaliação da Incerteza de dados Batimétricos

MAD – Desvio Absoluto da Mediana

MBES – *Multibeam Echo Sounders*

NOAA – *National Oceanic and Atmospheric Administration*

NORMAM – Normas da Autoridade Marítima

NMAD – Desvio Absoluto da Mediana Normalizado

NR – Nível de Redução

PP – *Point to Point*

RMSE – *Root Mean Square Error*

RPs – Resíduos Padronizados

RTK – *Real Time Kinematic*

SBES – *Single Beam Echo Sounders*

SP – *Surface to Point*

SS – *Surface to Surface*

SSS – *Side Scan Sonar*

TCL – Teorema Central do Limite

THU – *Total Horizontal Uncertainty*

TIPLAM – Terminal Integrador Portuário Luiz Antônio Mesquita

TPU – *Total Propagated Uncertainty*

TVG – *Time Varying Gain*

TVU – *Total Vertical Uncertainty*

USACE – *U.S. Army Corps of Engineers*

USGS – *U.S. Geological Survey*

UTM – Universal Transversa de Mercator

WMO – *World Meteorological Organization*

RESUMO

FERREIRA, Italo Oliveira, D.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Controle de Qualidade em Levantamentos Hidrográficos**. Orientador: Afonso de Paula dos Santos. Coorientadores: Nilcilene das Graças Medeiros, Júlio César de Oliveira e Paulo César Emiliano.

O presente trabalho teve como objetivos o desenvolvimento e a proposição de metodologias para o controle de qualidade de levantamentos hidrográficos, especificamente o tratamento e avaliação estatística das profundidades coletadas por ecobatímetros monofeixe e sistemas de sondagem por faixa, tais como: ecobatímetros multifeixe e sonares interferométricos. Num primeiro momento, é apresentado o AEDO (Algoritmo Espacial para Detecção de *Outliers*), uma metodologia robusta destinada a detecção de *spikes* em conjuntos de dados batimétricos coletados através de sistemas de sondagem por varrimento. Em apoio ao AEDO, também foi proposto um novo limiar espacial para detecção de *outliers*. Ao aplicar o método proposto ao conjunto de dados simulados e reais, resultados bastante promissores foram alcançados. Em seguida, ao focar-se na avaliação estatística da qualidade vertical de levantamentos batimétricos monofeixe, foi proposta a MAIB (Metodologia para Avaliação da Incerteza de dados Batimétricos). Tal método, em todos os casos, mostrou-se mais eficiente ante as análises tradicionais, visto que leva em consideração pressupostos estatísticos quase sempre negligenciados. Conjuntamente a MAIB, foi proposto um novo estimador robusto, isto é, resistente a *outliers*, para estimativa da incerteza vertical amostral das profundidades coletadas numa sondagem batimétrica. Objetivando avaliar também as sondagens multifeixe, foi desenvolvido um novo método para extração de pontos homólogos através do levantamento das tradicionais varreduras de verificação. Esse método, chamado de PP *(Point to Point)*, apresentou-se bastante eficaz, principalmente quando comparado as técnicas usuais. Por fim, é apresentada uma alternativa à realização de varreduras de verificação para a estimativa da incerteza vertical amostral, apoiado pelo novo método desenvolvido (Método PP). De um modo geral, mediante a realização desta pesquisa pôde-se obter resultados significativos para o controle de qualidade dos levantamentos hidrográficos, visto que foram desenvolvidas diversas novas metodologias e métodos.

ABSTRACT

FERREIRA, Italo Oliveira, D.Sc., Universidade Federal de Viçosa, February, 2018. **Quality Control in Hydrographic Surveys**. Adviser: Afonso de Paula dos Santos. Co-advisers: Nilcilene das Graças Medeiros, Júlio César de Oliveira and Paulo César Emiliano.

The present work had as objective the development and the proposal of methodologies for the quality control of hydrographic surveys, specifically the treatment and statistical evaluation of depths collected by single beam echo sounders and swath systems, such as multibeam echo sounders and interferometric sonars. Firstly, the AEDO (Spatial Algorithm for Outliers Detection) is presented, a robust methodology for the detection of spikes in bathymetric data sets collected through swath systems. In support of AEDO, a new spatial threshold for detection of outliers was also proposed. By applying the proposed method to the set of simulated and real data, very promising results were obtained. Then, when focusing on the statistical evaluation of the vertical quality of single beam surveys, the MAIB (Methodology for the Evaluation of the Bathymetric Data Uncertainty) was proposed. This method, in all cases, proved to be more efficient than traditional analyzes, since it takes into account statistical assumptions almost always neglected. In conjunction with the MAIB, a new robust statistic was proposed, that is, resistant to outliers, to estimate the vertical uncertainty of the depths collected at the bathymetric survey. In order to evaluate also the multibeam echo sounders data, a new method was developed to extract homologous points by surveying the traditional check lines. This method, named PP (Point to Point), showed up very effective, especially when compared to the usual techniques. Finally, it's presented an alternative to estimate the vertical uncertainty, in replacement the traditional check lines. This technique is supported by the PP method. In general, the work obtained significant results for the quality control of the hydrographic surveys, since several new methodologies and methods have been developed.

INTRODUÇÃO GERAL

De acordo com IHO (2005), os levantamentos hidrográficos lidam com a configuração dos fundos de oceanos, rios, lagos, barragens, portos, entre outros. Em termos gerais, os levantamentos hidrográficos tratam da medição de profundidades, marés, correntes oceânicas, gravidade, magnetismo terrestre e a determinação das propriedades químicas e físicas da água.

Embora o objetivo principal de um levantamento hidrográfico seja compilar dados para a geração de cartas náuticas confiáveis, o conhecimento do relevo submerso é de essencial importância em diversas outras áreas, como por exemplo: obras civis (pontes, portos, píeres), a locação de cabos e dutos, a prospecção de recursos minerais e o monitoramento de assoreamento de reservatórios (abastecimento ou geração de energia) (FERREIRA et al., 2012).

A confecção, edição e publicações das cartas náuticas brasileiras, assim como a execução e o controle dos levantamentos hidrográficos realizados em Águas Jurisdicionais Brasileiras (AJU), são atribuições da Marinha do Brasil, através da DHN (Diretoria de Hidrografia e Navegação) e do CHM (Centro de Hidrografia da Marinha).

Além dos levantamentos executados pela Marinha, o CHM fiscaliza, por força de determinação legal, a execução dos levantamentos hidrográficos executados por entidades extra Marinha, visando, principalmente, manter as cartas náuticas brasileiras atualizadas. Nesse sentido, a Marinha do Brasil também é responsável por estabelecer normas e procedimentos específicos referentes à coleta, processamento e envio dos dados, bem como à confecção dos relatórios finais. Estas especificações técnicas são elaboradas segundo padrões internacionais de qualidade recomendados pela IHO *(International Hydrographic Organization)*, organização intergovernamental fundada em 1921 por 19 países, incluindo o Brasil.

As normas e os procedimentos para autorização e controle dos levantamentos hidrográficos estão, atualmente, definidos na NORMAM – 25 (DHN, 2014). A NORMAM-25 classifica os levantamentos hidrográficos em duas categorias (ALFA e BRAVO) de natureza administrativa:

- CATEGORIA A (ALFA): levantamentos hidrográficos que devem seguir especificações técnicas que permitam que os dados obtidos sejam aproveitados na atualização de cartas náuticas ou para as demais finalidades descritas no item 0206 da NORMAM-25.
- CATEGORIA B (BRAVO): levantamentos hidrográficos executados sem o propósito de produzir elementos que sirvam para atualização de cartas náuticas.

Os levantamentos hidrográficos Categorias ALFA devem cumprir integralmente as especificações previstas na Publicação Especial S-44, 5ª edição (DHN, 2014). Neste sentido, a S-44 especifica quatro ordens de levantamentos: Ordem Especial, Ordem 1a, Ordem 1b e Ordem 2, sumarizados na Tabela 1 (IHO, 2008).

Embora a NORMAM – 25 não estabeleça procedimentos técnicos específicos para os levantamentos hidrográficos categoria BRAVO, ela recomenda a adoção dos mesmos procedimentos dos levantamentos ALFA, visando uma eventual alteração de categoria, com vistas ao aproveitamento dos dados para confecção ou atualização da cartografia náutica.

A medição de profundidade é a principal tarefa de um levantamento hidrográfico (IHO, 2005). Esta tarefa requer conhecimento específico do meio físico, da acústica submarina, dos inúmeros equipamentos e sensores utilizados, além dos procedimentos apropriados para cumprir os requisitos e recomendações definidos pela Publicação Especial S-44, 5ª edição. As profundidades são obtidas através dos levantamentos batimétricos. O termo levantamento batimétrico é encontrado correntemente na literatura e trata-se da realização de medições de profundidades associadas a uma posição na superfície (JONG et al., 2010, FERREIRA et al., 2015; FERREIRA et al., 2016b).

O sensoriamento remoto acústico é, atualmente, o principal meio de investigação de fundos submersos, visto que os métodos tradicionais de sensoriamento remoto, óticos e radar, são pouco eficientes devido à alta atenuação das ondas eletromagnéticas pela água (AYRES NETO, 2000; FERREIRA et al., 2017a). Sendo assim, têm-se visto uma preferência pela utilização de sistemas acústicos para medição de profundidade, como ecobatímetros monofeixe (SBES – *Single Beam Echo Sounders*), multifeixe (MBES – *Multibeam Echo Sounders*) e sonares interferométricos. Sistemas de sondagem laser aerotransportados, conhecidos como *LiDAR (Light Detection And Ranging) batimétricos,* apesar do alto custo, também vem

sendo utilizados. É uma tecnologia em ascensão no mercado, destacando-se, principalmente, pelo grande ganho de produtividade (GUENTHER et al., 1996; IHO, 2005; ATHEARN et al., 2010; PASTOL, 2011; ELLMER et al., 2014).

Tabela 1 – Padrões mínimos para Levantamentos Hidrográficos.

Ordem	**Especial**	**1a**	**1b**	**2**
Descrição das áreas	Áreas onde a altura livre sob a quilha é de importância crítica.	Áreas com profundidades menores que 100 metros nas quais a lazeira de água debaixo da quilha não é de importância crítica, mas onde há possibilidade de existirem feições que ponham em risco a navegação.	Áreas com profundidades menores que 100 metros nas quais a lazeira de água debaixo da quilha não é um fator de risco em virtude do tipo de embarcações que deverão transitar nelas.	Áreas com profundidades maiores que 100 metros nas quais uma descrição geral do solo marítimo é considerada apropriada.
IHT (Incerteza Horizontal Total) máxima permitida. Nível de confiança de 95%	2 metros	5 metros + 5% da profundidade	5 metros + 5% da profundidade	20 metros + 10% da profundidade
IVT (Incerteza Vertical Total) máxima permitida. Nível de confiança de 95%	a = 0,25 metro b= 0,0075	a = 0,50 metro b= 0,013	a = 0,50 metro b= 0,013	a = 1,00 metro b= 0,023
Levantamento completo do solo marítimo	Feições cúbicas com aresta superior a 1 metro.	Feições cúbicas com aresta superior a 2 metros em fundos até aos 40 metros; em fundos superiores, aresta superior a 10% da profundidade.	Não aplicável.	Não aplicável.
Detecção de feições	Exigido	Exigido	Não aplicável	Não aplicável
Máximo espaçamentos entre linhas regulares de sondagem	Não aplicável. Requerida a busca total do fundo submerso.	Não aplicável. Requerida a busca total do fundo submerso.	3 vezes a profundidade média ou 25 metros, conforme o maior valor.	4 vezes a profundidade média.

Fonte: Adaptado de IHO (2008).

Sob outra perspectiva, a determinação da profundidade derivada de imagens orbitais ainda é uma área que carece de estudos mais detalhados. Pesquisas realizadas por Gao (2009), Cheng et al. (2015), Moura et al. (2016) e Ferreira et al. (2016b) mostraram que esta tecnologia possui um custo relativamente baixo e reduzido tempo de execução. Porém, sua utilização limita-se a águas pouco profundas (~10 metros) e as informações ainda são obtidas com acurácias incompatíveis com os requisitos atuais, restringindo seu uso para fins de planejamento, reconhecimento e modelagem ambiental. Assim, em levantamentos batimétricos, o uso de imagens orbitais permanece, principalmente como uma ferramenta de reconhecimento e de planejamento em áreas onde informações batimétricas são inexistentes ou insuficientes. Por outro lado, imagens oriundas de sensores orbitais e aerotransportados mostram-se uma ferramenta muito útil para a delineação de linhas de costa (FERREIRA et al., 2017a).

À primeira vista, o levantamento batimétrico pode parecer semelhante ao levantamento topográfico, porém essa semelhança se limita a representação por linhas de igual cota e ao tratamento computacional das superfícies. Os procedimentos seguidos no planejamento, coleta e análise dos dados são diferentes dos usados na topografia terrestre. Na topografia a superfície a ser mapeada é visível, sendo assim, pontos de mudança de declividade, acidentes geográficos, construções, dentre outros, podem ser facilmente localizados e levantados. Além disso, é possível materializar pontos estáveis de observação (marcos) e efetuar medições repetidas, para um posterior ajustamento de observações (FERREIRA et al., 2017b). No levantamento batimétrico, entretanto, a área a ser cartografada é dividida em uma malha de linhas equidistantes, doravante denominadas linhas regulares de sondagem, que são percorridas pela plataforma de sondagem permitindo a coleta de dados de profundidade e posição. O espaçamento entre linhas, bem como a orientação, é altamente dependente da tecnologia adotada na medição de profundidade (IHO, 2008; DHN, 2014).

Ainda num cenário atual, os produtos gerados com os sistemas de batimetria multifeixe apresentam um elevado ganho em resolução e acurácia, tanto em termos planimétricos quanto altimétricos (profundidade), e um grande adensamento de dados, descrevendo quase que por completo o fundo submerso (IHO, 2005; USACE, 2013; MALEIKA, 2015). Enquanto o SBES realiza um único registro de profundidade a cada pulso acústico transmitido *(ping)*, tendo como resultado uma linha de pontos

imediatamente abaixo da trajetória da embarcação, o multifeixe executa diversas medidas de profundidade com um mesmo *ping*, obtendo medições da coluna d'água em uma faixa *(swath)* perpendicular à trajetória da embarcação. Um número crescente de serviços hidrográficos adotou a tecnologia multifeixe como a metodologia principal para coleta de dados batimétricos visando a produção e atualização cartográfica (IHO, 2008; INSTITUTO HIDROGRÁFICO, 2009; LINZ, 2010; NOAA, 2011; USACE, 2013; DHN, 2014). Analisando o exposto na Tabela 1, verifica-se que levantamentos de Ordem Especial e 1a requerem a busca total do fundo submerso, ou seja, uso de MBES. Em alguns casos é válido o uso de SBES apoiado por SSS *(Side Scan Sonar)*, sonares interferométricos ou sistemas SBES multitransdutores (CLARKE, 2014; CRUZ et al., 2014).

Na batimetria multifeixe, para fins de construção ou atualização de cartas náuticas no âmbito dos levantamentos hidrográficos regulamentados pela NORMAM – 25, deve-se adotar como espaçamento entre as linhas regulares de sondagem a metade da largura de varredura (cobertura de fundo). Esse planejamento visa obter 100% de sobreposição entre linhas adjacentes, implicando na prática em uma ensonificação de 200%. Esse valor de sobreposição é recomendado para um correto processamento dos dados coletados. Deve-se atentar para o fato de que a cobertura do fundo é dada em função, entre outros, da profundidade, desta forma, o espaçamento entre linhas pode não ser constante para toda a área. As linhas são planejadas de forma paralela as isobátas, este direcionamento é exatamente o oposto ao adotado para levantamento com SBES. De acordo com DHN (2014), as linhas de verificação, fundamentais para o controle de qualidade dos dados, devem ser planejadas de modo, aproximadamente, perpendicular às linhas regulares de sondagem. O espaçamento máximo deve ser de até 15 vezes o adotado para as linhas regulares de sondagem.

Um sistema de sondagem multifeixe, assim como qualquer sistema de levantamento por varredura (sonares interferométricos, *LiDAR batimétricos*, sistemas multi-transdutores, *etc*.) é composto por diversos sensores, como por exemplo, ecobatímetros, receptores GNSS *(Global Navigation Satellite System)*, sensores inerciais e sensores de proa (IHO, 2005; JONG et al., 2010; CRUZ et al., 2014; FERREIRA et al., 2015). Para que as informações coletadas por esses sensores possam ser sincronizadas, é necessário o conhecimento da posição tridimensional de cada sensor em relação ao sistema de coordenadas da embarcação. É sabido que incertezas

no posicionamento destes sensores irão introduzir incertezas horizontais e verticais na medição da profundidade reduzida[1] (CLARKE, 2003).

Independente da configuração de equipamentos ou da disposição destes a bordo, nos levantamentos hidrográficos, assim como nos levantamentos fotogramétricos, geodésicos ou topográficos, as observações conterão incertezas. Tais incertezas são tradicionalmente divididas em erros grosseiros, efeitos sistemáticos e efeitos aleatórios.

Os erros grosseiros *(blunders)* são aqueles provocados por falhas ocasionais dos instrumentos e/ou do observador. Tais erros devem ser detectados, através de técnicas estatísticas ou geoestatísticas, e eliminados. Os efeitos sistemáticos são devido às deficiências na compensação dos erros fixos ou de desvios nas medições, e devem ser modelados, determinados e eliminados (ou ao menos minimizados) durante a calibração do sistema ou inseridos no modelo matemático. Por fim, restam ainda os efeitos aleatórios ou flutuações probabilísticas. Esses efeitos são uma das principais causas do valor verdadeiro de uma observação nunca ser conhecido.

É comum na literatura o uso do termo *outlier* como sinônimo de erro grosseiro, no entanto, é importante destacar que *outlier* é uma observação que, estatisticamente, se diferencia do conjunto de dados ao qual pertence, ou seja, é um valor atípico ou inconsistente. Nesse sentido, *outliers* podem ser causados por erros grosseiros, por efeitos sistemáticos ou, simplesmente, por efeitos aleatórios (SANTOS et al., 2016). Em levantamentos hidrográficos, profundidades que se configuram como *outliers* são designados de *spikes*[2]*,* enquanto que os erros de posicionamento são chamados de *tops.* Este trabalho foca, especificadamente, na componente vertical e, por esse motivo, o termo *spike* as vezes é tratado como sinônimo de *outlier.*

Atualmente, os sistemas multifeixe são capazes de coletar em águas rasas mais de 30 milhões de pontos por hora, assim, o processamento dos dados na sua forma tradicional tornou-se mais moroso do que o próprio levantamento hidrográfico (CALDER & MAYER, 2003; CALDER & SMITH, 2003; BJØRKE & NILSEN, 2009; VICENTE, 2011; LU et al., 2010). Em resposta ao aumento no volume de dados coletados pelas sondas multifeixe, a partir da década de 1990, pesquisadores começaram a desenvolver metodologias e algoritmos de processamento assistido por computador, com o objetivo de facilitar a tarefa do hidrógrafo (VICENTE, 2011).

[1] Profundidade referenciada a um Nível de Redução (NR), ou seja, corrigida dos efeitos de maré.
[2] Profundidades espúrias.

Segundo Debese (2007), esses algoritmos basicamente estimam a profundidade numa determinada localização, alguns ainda são capazes de avaliar qualitativamente o processo de estimação. Dentre os vários, merece destaque o algoritmo CUBE *(Combined Uncertainty and Bathymetry Estimator)* apresentado por Calder (2003). Este é considerado, até o momento, como um dos mais promissores algoritmos para processamento semiautomático de dados multifeixe, estando implementado em inúmeros pacotes comerciais de processamento.

Segundo Vicente (2011), o processamento tradicional de dados multifeixe, com recursos a pacotes comerciais de processamento, como: *Caris-Hips, Hysweep (Hypack), QPS, PDS2000 etc.*, pode ser realizado através das seguintes fases:

Fase 1, que consiste em um controle de qualidade:

- Conversão dos dados coletados pelos diversos sensores para o formato do *software* de processamento empregado;
- Análise dos dados dos sensores auxiliares (atitude, latência, velocidade do som, maré *etc*.) objetivando a identificação de possíveis falhas. Se necessário, interpolação ou rejeição de dados anômalos;
- Junção dos datagramas[3];
- Cálculo da *Total Propagated Uncertainty* (Horizontal e Vertical);
- Filtragem (Rejeição de dados com incertezas propagadas superiores ao admissível);
- Construção de um modelo batimétrico, utilizando, por exemplo, o CUBE (mais usual), e
- Análise do modelo batimétrico para detecção e inspeção de situações anômalas (falhas de cobertura, *spikes etc*.).

Fase 2, que consiste na validação dos dados:

- Análise completa dos dados através de ferramentas de visualização 2D e 3D;
- Aplicação de filtros automáticos;
- Possibilidade de, manualmente, requalificar as sondagens, e
- Efetuar uma nova interpolação dos dados (Re-CUBE).

[3] Entidade de dados completa e independente. Nesse caso, refere-se aos dados gerados pelos diversos sensores.

Fase 3, que consiste na geração dos produtos finais:

- O produto final de um levantamento hidrográfico é um conjunto de sondagens referidas a um sistema de coordenadas geodésicas ou a um sistema de coordenadas plano-retangular (por exemplo a projeção UTM – Universal Transversa de Mercator).

Esse fluxo de trabalho é, atualmente, utilizado no Instituto Hidrográfico Português. Ainda segundo Vicente (2011), o processamento tradicional é um processo lento, baseado num julgamento qualitativo, conservativo e subjetivo, no qual é dada primazia às sondagens mínimas visando a segurança na navegação. Todavia, deve-se atentar que se os requisitos inerentes a uma ordem de levantamento hidrográfico forem cumpridos, espera-se que a maioria dos dados adquiridos possuam qualidade para serem utilizados na cartografia náutica e nas demais finalidades previstas em norma. No entanto, é impossível ao hidrógrafo visualizar e validar todos os dados de uma forma coerente.

No caso do Brasil, a NORMAM-25, em sua versão mais recente (DHN, 2014), estipula de forma muito sucinta procedimentos a serem seguidos na coleta e processamento dos dados batimétricos. Porém, não especifica nada relacionado a interpolação dos dados ou geração de modelos batimétricos, apesar de esses auxiliarem na detecção de dados duvidosos. Contudo, cabe destacar que, embora as superfícies batimétricas sejam ferramentas de extrema utilidade para a visualização das características submersas, os dados batimétricos interpolados provenientes dessa superfície não podem ser usados na construção ou na atualização de cartas náuticas.

A 5ª edição da S-44 disponibiliza em seu anexo B, instruções para o processamento dos dados, estas devem ser seguidas tendo em vista a geração de produtos que cumpram os requisitos previstos em norma (IHO, 2008). Recomenda-se que as diretrizes do anexo B sejam analisadas conjuntamente com aquelas fornecidas no anexo A, que trata do controle de qualidade.

Dentre as diversas etapas do processamento multifeixe merece destaque a detecção, análise e eliminação de dados anômalos *(spikes)*, que ocorrem, conforme descrito, na fase 1. Geralmente esta tarefa é realizada manualmente pelo hidrógrafo que, visualizando os dados através de uma interface gráfica, decide subjetivamente qual sondagem pode, ou não, ser considerada um *outlier*. Devido ao grande volume de dados provenientes de uma sondagem multifeixe, essa tarefa tornou-se muito demorada e bastante subjetiva (WARE et al., 1992; ARTILHEIRO, 1998; CALDER

& MAYER, 2003; CALDER & SMITH, 2003; VICENTE, 2011). DHN (2014), por exemplo, destaca que, durante a edição manual de dados multifeixe por um analista, podem ocorrer a eliminação de profundidades mínimas e de alto fundo. Ambas oferecem perigos ao navegante.

Na medição da profundidade, dados discrepantes podem depender, entre outros fatores, da qualidade do ecobatímetro multifeixe e do seu algoritmo de detecção de fundo (detecção por fase, amplitude, transformada de *Fourier etc*.), da detecção por lóbulos secundários, de reflexões múltiplas, da presença de bolhas de ar na face do transdutor e de reflexões na coluna d'água causadas, principalmente, por: algas, cardumes de peixe, *deep scattering layer*[4], variações térmicas e sedimentos em suspensão (URICK, 1975; JONG et al., 2010).

Diversos autores desenvolveram pesquisas na área de detecção de valores anômalos em dados de batimetria advindos de sondadores acústicos. Ware et al. (1991), apresentaram um processo de detecção de *outliers* baseado na análise das propriedades estatísticas da amostra. Seguindo linhas similares, Eeg (1995) propôs um teste estatístico para validar o tamanho dos agrupamentos que são utilizados para detectar *spikes.* Debese & Bisquay (1999), Motao et al. (1999), Debese (2007) e Debese et al. (2012) aplicaram, basicamente, estimadores-M. Calder & Mayer (2003) utilizaram o filtro de *Kalman* para processar dados batimétricos automaticamente. De forma análoga, Bottelier et al. (2005) empregaram técnicas de krigagem e Bjørke & Nilsen (2009) apresentaram uma técnica para detecção de *spikes* fundamentada na construção de superfícies de tendência. Já Lu et al. (2010) desenvolveram um algoritmo baseado no estimador robusto LTS *(Least Trimmed Squares).*

Atualmente, existem algoritmos que permitem uma eficaz detecção de valores anômalos, como por exemplo, o IODA *(Intelligent Outlier Detection Algorithm)* aplicado na análise de séries temporais (WEEKLEY et al*.,* 2010). Todavia, estas metodologias são, em sua maioria, de difícil aplicação, semiautomatizadas ou encontram-se implementadas apenas em pacotes comerciais. Outra problemática dessas metodologias consiste no fato da maioria delas fundamentar-se em pressuposições teóricas dificilmente atendidas e/ou verificadas, como por exemplo, assumir que as variáveis estudadas são independentes e pertencem a conjuntos de variáveis normalmente distribuídas. Uma técnica baseada na análise de resíduos

[4] Consiste numa camada de plâncton que varia de profundidade ao longo do dia (IHO, 2005).

padronizados foi recentemente apresentada por Santos et al. (2017) para dados de altimetria terrestre. A metodologia, apesar de não ser automatizada, apresentou-se bastante eficiente para detecção de *outliers*.

Por outro lado, na estatística clássica umas das ferramentas mais utilizadas para detecção de *outliers* em amostras de dados contínuos univariados é o diagrama *boxplot* ou gráfico de caixa (TUKEY, 1977; CHAMBERS et al., 1983; HOAGLIN et al., 1983). O *boxplot* é uma ferramenta muito aplicada também para visualizar a distribuição do conjunto de dados. Outro método que pode ser empregado para pesquisar *outliers* é o *Z-Score Modificado*. Este método é baseado em estatísticas robustas, como a mediana e o desvio absoluto da mediana *(median absolute deviation),* que são capazes de garantir que os limiares definidos como valores de corte não sejam afetados, justamente, pela presença de *outliers* (IGLEWICZ & HOAGLIN, 1993). Diversos outros métodos podem ser aplicados para detecção de valores anômalos em conjuntos de dados univariados, compostos por variáveis quantitativas contínuas, tal como resumido em Seo (2006).

No entanto, essas metodologias não levam em consideração a posição geográfica dos dados, o que, *a priori*, as torna ineficientes no controle de qualidade de dados batimétricos. Soma-se a isto, o fato destas técnicas assumirem que as observações são variáveis aleatórias independentes e identicamente distribuídas (MORETTIN & BUSSAB, 2004; SEO, 2006), pressupostos indispensáveis para um tratamento estatístico clássico e coerente. Porém, quando as variáveis são georreferenciadas, a autocorrelação espacial torna-se uma característica inerente.

Outra problemática da maioria dessas metodologias reside no fato dos valores de corte para detecção de *outliers* serem derivados da distribuição normal, o que reduz a eficiência da metodologia quando a amostra não é simétrica (HUBERT & VANDERVIEREN, 2008). Todavia, são mecanismos de simples aplicação e análise. Assim, pode-se vislumbrar a possibilidade do desenvolvimento e aplicação de algoritmos para detecção automatizada de *spikes* através da utilização desses mecanismos em amostras de dados espaciais, desde que as metodologias desenvolvidas levem em consideração a estrutura de dependência espacial, intrínseco aos dados batimétricos, e os pressupostos estatísticos básicos.

Conforme afirmam Santos et al. (2017), a Geoestatística destaca-se como uma potencial ferramenta para o desenvolvimento de metodologias para detecção de *outliers* em dados com geolocalização e consequente dependência espacial. Além do

mais, essa ferramenta tem como principal característica a modelagem espacial sem tendência e com variância mínima, atributos que podem robustecer quaisquer técnicas de detecção de *outliers*. Sendo assim, pode-se utilizá-la como ferramenta de suporte às metodologias desenvolvidas neste estudo, tendo em vista suas características ideais. Tais características foram confirmadas por Ferreira et al. (2013, 2015 e 2017b) durante estudos para modelagem de superfícies batimétricas.

Após a detecção, análise e eliminação de dados discrepantes, resta avaliar a qualidade dos dados e/ou dos produtos gerados a partir do levantamento hidrográfico. Segundo Monico et al. (2009), nas ciências geodésicas e cartográficas corriqueiramente encontra-se os termos acurácia e precisão interpretados de forma equivocada. Mikhail & Ackermann (1976) apresentam acurácia como sendo o grau de proximidade de uma estimativa com seu valor de referência, enquanto precisão expressa o grau de consistência da grandeza medida em relação a sua média, estando esta ligada diretamente com a dispersão da distribuição das observações. Ainda conforme os mesmos autores, a acurácia incorpora efeitos aleatórios e sistemáticos e a precisão está associada apenas com efeitos aleatórios. Em suma, pode-se concluir que, matematicamente, o termo acurácia por si só envolve a medida de precisão. Kirkup & Frenkel (2006) destacam que a alta acurácia implica em alta precisão, porém o contrário não se verifica quando as observações estão eivadas de significativos efeitos sistemáticos.

É também comum o emprego do termo exatidão, tomado, na maioria das vezes, como sinônimo de acurácia (ANDRADE, 2003; KIRKUP & FRENKEL, 2006; INMETRO, 2012b). Entretanto, Rodrigues (2008) faz uma distinção entre os termos acurácia e exatidão, definindo este último como sendo o grau de aderência de uma estimativa em relação ao seu valor verdadeiro, enquanto acurácia é quantificada em relação a um valor de referência, tal como definido.

Em levantamentos hidrográficos frequentemente termos como: erro, exatidão, precisão (quantificada pelo desvio padrão), desvio padrão, repetibilidade, acurácia, *etc*., são utilizados indevidamente ou confundidos com a incerteza estimada ou resultante de uma profundidade observada.

O termo "erro" é tradicionalmente definido como sendo a diferença entre o valor real (valor exato) e o valor observado. Nesse sentido, o termo "erro" está intimamente ligado à expressão exatidão, ou seja, uma observação é dita exata quando

está isenta de erros, ou ainda, uma observação será mais exata quanto menor for a magnitude dos possíveis erros cometidos (IHO, 2008; INMETRO, 2012a, b).

No entanto, o termo "erro" é cientificamente vago, pois é sabido que as observações podem estar contaminadas por erros grosseiros, efeitos sistemáticos e aleatórios, sendo assim, mesmo se o valor verdadeiro de uma observação for conhecido, o que nas ciências hidrográficas é improvável, devido as flutuações probabilísticas, a magnitude da diferença entre a observação e o valor verdadeiro seria uma junção de possíveis erros grosseiros, efeitos sistemáticos e aleatórios. Além do mais, o "erro" é um termo pejorativo, ligado a falhas, enganos e negligências, sendo então, um indutor de conclusões equivocadas. Logo, o seu uso deve ser evitado. Do mesmo modo, é um equívoco usar o termo exatidão em levantamentos hidrográficos quando se tratar de controle de qualidade.

A coleta de dados redundantes nos levantamentos hidrográficos não é tão simples como no mapeamento terrestre, sendo assim, termos como precisão, desvio padrão e repetibilidade também devem ser evitados. Por fim, o uso do termo acurácia fica limitado à complexidade de definir valores de referência em ambientes submersos.

Devido aos conceitos expostos anteriormente e seguindo as recomendações de IHO (2008), INMETRO (2012a, b) e Ferreira et al. (2016a) neste trabalho, será dada preferência ao termo incerteza. De acordo com Vicente (2011), a definição de incerteza assume um papel fundamental no meio hidrográfico e pode ser entendida como sendo o intervalo em torno de um valor de profundidade, que contém o valor medido num nível de confiança específico.

As profundidades observadas pelos sistemas acústicos derivam do intervalo de tempo entre a saída e a chegada de um mesmo pulso acústico ao transdutor. A metade desse tempo multiplicado pela velocidade de propagação do som na água produzirá uma estimativa da profundidade local, denominada sondagem. Às sondagens devem ser acrescidas diversas correções para que seja possível a obtenção da profundidade corrigida (IHO, 2008; USACE, 2013; FERREIRA et al., 2015). O termo profundidade reduzida é comumente utilizado na comunidade hidrográfica e pode ser entendida como a profundidade corrigida referenciada a um Nível de Redução (NR), ou seja, corrigida também dos efeitos de maré.

De acordo com USACE (2013), a incerteza de medição da profundidade possui muitas fontes em potencial. Estas incluem: o método de medição, a velocidade de propagação do som na água, a largura do feixe *(beamwidth)*, o tipo e formato de fundo,

os movimentos da plataforma de sondagem *(roll-pitch-heave-heading)* e a profundidade de imersão do transdutor *(draft* ou *draught).* Todos esses fatores compõem o modelo de incertezas de medição da profundidade reduzida.

Hare et al. (2011) evidenciam que em uma sondagem batimétrica existem fontes de incerteza que contribuem apenas com a incerteza vertical, fontes de incerteza que contribuem apenas com a incerteza horizontal e aquelas que contribuem com ambas.

Conforme sintetizado por IHO (2008), incertezas individuais associadas com a **posição horizontal** de um feixe incluem:

a) Incertezas de posicionamento do sistema;

b) Incertezas de alcance e de feixe;

c) Incertezas associadas com o modelo de trajetória do raio acústico (incluindo o perfil da velocidade do som) e o ângulo de direcionamento do feixe;

d) Incertezas na determinação do rumo/proa *(heading)* da embarcação;

e) Incertezas resultantes do desalinhamento do transdutor;

f) Incertezas devido à localização dos sensores, como, por exemplo, o *heave* induzido;

g) Incertezas nas medições realizadas pelo sensor de movimentos da embarcação como, por exemplo, *roll, pitch* e *heave*;

h) Incertezas na medição dos afastamentos *(offsets)* dos diversos sensores a bordo; e

i) Incertezas associadas a sincronização do tempo/latência.

Fatores que podem contribuir com a incerteza **vertical** incluem:

a) Incertezas associados à redução ao *datum* vertical (quando aplicável);

b) Incertezas do sistema de posicionamento vertical;

c) Incertezas associadas a medição de marés, incluindo erros cotidais (quando aplicável);

d) Incertezas instrumentais;

e) Incertezas associadas a determinação do perfil de velocidade do som;

f) Incertezas elipsoidais/incertezas de modelo de separação do *datum* vertical (quando aplicável);

g) Incertezas associadas aos movimentos da embarcação como, por exemplo*, roll, pitch* e, principalmente, *heave;*

h) Incertezas devido a medição do *draft*;

i) Incertezas associadas aos movimentos de *settlement* e *squat* da embarcação;

j) Incertezas associadas a inclinação e variação de relevo submerso; e

k) Incertezas associadas a sincronização do tempo/latência.

Todos estes elementos podem ser combinados através da aplicação da lei de propagação de incertezas, desde que todos os pressupostos sejam atendidos, para fornecer uma estimativa da IPT (Incerteza Propagada Total, do Inglês *TPU – Total Propagated Uncertainty*) do sistema de sondagem.

Ainda de acordo com IHO (2008), a IPT é definida como a incerteza propagada total no processo de determinação da profundidade reduzida. É composta pelas componentes horizontal ou IHT (Incerteza Horizontal Total, do Inglês *THU – Total Horizontal Uncertainty*) e vertical ou IVT (Incerteza Vertical Total, do Inglês *TVU – Total Vertical Uncertainty*). O fato dos termos IHT e IVT não terem a palavra "propagada" em sua expressão, pode conduzir a erros de interpretação, desse modo, evidencia-se que ambas são incertezas propagadas.

Na prática não se calcula a IPT como um todo. Estimam-se a IVT e a IHT a partir de propagação de covariâncias, considerando apenas as fontes de incerteza que as afetam individualmente. Nessas previsões estatísticas, considera-se que todas as componentes individuais são variáveis aleatórias, não correlacionadas, cujas incertezas seguem uma distribuição normal (FERREIRA et al., 2016a).

Apesar da IPT ser mencionada como um único numeral, a IHT é uma quantidade bidimensional (IHO, 2008). Uma metodologia para estimação da IPT para sistemas multifeixe foi documentada por Hare (1995). Segundo Hare et al. (2011), a mesma metodologia pode ser aplicada para quantificação da IPT de sistemas monofeixe, considerando o caso especial de uma sonda multifeixe operando somente com feixe nadiral (central), tal como realizado por Ferreira et al. (2016a).

Após calculadas a IHT e a IVT, desde que seguidos todos os procedimentos previstos em norma, pode-se classificar a ordem do levantamento de acordo com as especificações da S-44 (Tabela 1). Neste caso, todas as profundidades do levantamento em questão devem ter IHT e IVT, expressas ao nível de confiança de 95%, iguais ou inferiores aos valores máximos permitidos. De modo contrário, o levantamento deve ser reclassificado, refeito ou as profundidades com incertezas superiores as tolerâncias estabelecidas em norma devem ser desconsideradas.

Todavia, uma propagação de covariâncias, apesar de considerar incertezas obtidas em todas as etapas de um levantamento hidrográfico, sejam elas sistemáticas

ou aleatórias, estabelece apenas uma estimativa da qualidade do levantamento baseada nos possíveis desvios não correlacionados do sistema de sondagem (IHO, 2005; LINZ, 2010; Ferreira et al., 2015). Além disso, as incertezas utilizadas na computação da IPT são, em sua maioria, resultantes de testes de laboratório que não consideram as reais condições de operação (FERREIRA et al., 2016b).

Pode-se concluir que a metodologia citada apenas deveria ser utilizada para demonstrar a capacidade do sistema de levantamento, pois não leva em consideração a aleatoriedade de medidas obtidas naturalmente, como é o caso das sondagens batimétricas. Em outras palavras, tal critério, por si só, não atesta a qualidade dos dados coletados. Além disso, conforme afirma IHO (2008), simplesmente fazer uso de um equipamento que teoricamente cumpre a incerteza requerida, não é necessariamente o bastante. Fatores tais como: a maneira como o equipamento é montado, utilizado e o modo como ele interage com os demais componentes do sistema de levantamento interferem na incerteza do produto final. O profissional neste caso é uma peça fundamental, uma vez que precisa conhecer as técnicas de medição e os impactos das incertezas inerentes ao processo. Sendo assim, é preferível que a estimativa da incerteza dos dados coletados seja baseada em observações redundantes.

Porém, conforme supracitado, a coleta de dados redundantes em ambientes submersos não é tão simples como no mapeamento terrestre, onde é possível efetuar inúmeras observações repetidas para um posterior ajustamento de observações além de fixar pontos de controle. Em um levantamento hidrográfico até mesmo a detecção de *spikes* torna-se uma tarefa árdua, principalmente, devido à superfície do fundo submerso não ser visível (USACE, 2002; FERREIRA et al., 2015). Portanto, estimar a acurácia de um levantamento hidrográfico torna-se algo, *a priori,* impraticável e a precisão somente pode ser obtida através de suposições estatísticas (USACE, 2002, 2013).

Diante da dificuldade de obtenção de medidas repetidas em fundos submersos, em levantamentos hidrográficos realizam-se linhas de verificação, que cruzam as linhas regulares de sondagem ortogonalmente. As linhas de verificação devem ser coletadas, preferencialmente, em momentos distintos e em condições atmosféricas favoráveis. Supondo-se que os *spikes* e *tops* tenham sido eliminados ou minimizados, durante a fase de processamento dos dados, procede-se a avaliação da qualidade vertical do levantamento a partir da comparação entre as profundidades próximas às interseções entre as linhas regulares de sondagem e as linhas de verificação. A partir

das discrepâncias entre as profundidades são efetuadas análises estatísticas objetivando obter uma estimativa da qualidade vertical do levantamento (IHO, 2008; DHN, 2014; FERREIRA et al., 2015).

A técnica supracitada é exaustivamente utilizada para avaliar as profundidades coletadas por ecobatímetros monofeixe. Entretanto, a abundância de dados gerados numa sondagem multifeixe, implica, em um primeiro momento, na necessidade de uma adaptação desse procedimento. É comum a criação de modelos batimétricos das faixas sondadas através das linhas regulares e das linhas de verificação e, assim, efetuar a comparação entre as superfícies visando obter um arquivo de discrepâncias (SUSAN & WELLS, 2000; EEG, 2010). Porém, modelos batimétricos são resultantes de interpolações matemáticas que, como sabido, possuem incertezas em suas estimativas (FERREIRA et al., 2013, 2015). Dessa forma, notavelmente, a análise estatística da incerteza proveniente do levantamento hidrográfico ficaria comprometida.

Alternativamente, com a finalidade de reduzir as incertezas introduzidas pelas interpolações, pode-se gerar superfícies batimétricas apenas para as linhas de sondagem regulares e, então, efetuar a comparação entre os valores de profundidade armazenados no modelo, com as profundidades obtidas da varredura de verificação, embora a análise ainda fique, mesmo que em partes, comprometida. Assim, um algoritmo mais eficiente seria aquele que não necessitasse de predições matemáticas ainda que, inicialmente, o esforço computacional necessário possa limitar o emprego dessa metodologia.

Conforme exposto, as sondagens batimétricas realizadas para fins de construção ou atualização de cartas náuticas, devem adotar como espaçamento entre as linhas regulares de sondagem a metade da largura de varredura. Tal procedimento sugere que varreduras adjacentes irão se sobrepor. Diante disso, pode-se vislumbrar a utilização dessas informações para avaliar a qualidade da sondagem batimétrica por meio da geração de amostras de discrepâncias.

De acordo com IHO (2014), linhas de verificação ou áreas de varredura sobrepostas indicam o nível de conformidade ou repetição das medidas; porém não indicam acurácia absoluta uma vez que os dados são coletados a partir da mesma plataforma de sondagem e, neste caso, há um grande número de fontes de incertezas comuns em potencial entre os dados das linhas regulares e das linhas de verificação. USACE (2002) afirma que esses métodos de avaliação somente fornecem uma

estimativa da acurácia das medidas de profundidade, pois não são um teste independente. Todavia, as linhas de verificação fornecem um bom indicador de qualidade vertical do levantamento, e nesse caso, seu uso é recomendável e exigido por normas como IHO (2008), Instituto Hidrográfico (2009), LINZ (2010), NOAA (2011), USACE (2013) e DHN (2014).

Embora as normativas exijam a execução de linhas de verificação, elas não apresentam os procedimentos para avaliação estatística das discrepâncias. No entanto, diversos pacotes comerciais possuem ferramentas para a comparação estatística das varreduras de verificação com a sondagem regular. Basicamente, os algoritmos realizam uma confrontação entre as profundidades, calculando estatísticas como média, desvio-padrão, máximo e mínimo, dentre outras. Na prática, a principal estatística utilizada para classificação da ordem do levantamento são as diferenças entre as profundidades ou discrepâncias. Na hipótese de 95% das discrepâncias serem iguais ou estarem abaixo da tolerância prevista na S-44 para a ordem requerida, é comum, entre a comunidade hidrográfica, considerar que determinado levantamento cumpre os requisitos de incerteza para ser classificado naquela ordem, bem como enquadrado na categoria ALFA da NORMAM-25. Contudo, a avaliação baseada apenas na percentagem é insuficiente para avaliar e classificar o levantamento hidrográfico. Susan & Wells (2000) e Eeg (2010) utilizam o estimador *RMSE (Root Mean Square Error)* para estimar a incerteza do levantamento e então classificá-lo de acordo com as ordens previstas na S-44. Todavia, visto que as tolerâncias estipuladas em norma são intervalares, medidas pontuais e/ou intervalos de confiança inconsistentes, tornam a avaliação estatística pouco eficaz e equivocada.

Em todos os casos, nota-se que as análises são realizadas negligenciando-se pressupostos básicos da estatística clássica, como a avaliação da presença, ou não, de *outliers*, aplicação de testes de normalidade, uma vez que a distribuição normal é assumida, e a verificação da independência espacial das discrepâncias. Conforme afirmam Li et al. (2005), Maune (2007) e Santos (2015), esses pressupostos teóricos nem sempre são verdadeiros e podem, nesses casos, mascarar a análise da qualidade do levantamento.

Portanto, é nítida a necessidade do desenvolvimento de metodologias para uma avaliação estatística adequada, consistente e, especialmente, acurada de levantamentos hidrográficos realizados com auxílio à ecobatímetros monofeixe e sistemas de sondagem por varrimento.

1. HIPÓTESES

Diante do exposto, definem-se como hipóteses deste trabalho:

i. É possível desenvolver uma nova metodologia para localizar *spikes* em dados de batimetria coletados por sistemas de sondagem por faixa, assistida por técnicas estatísticas que a princípio não levem em consideração a espacialização geográfica dos dados.

ii. É viável desenvolver e aplicar, nas metodologias apresentadas neste trabalho, um novo limiar baseado em estatísticas robustas para localização de *spikes.*

iii. É possível estimar a qualidade vertical de um levantamento hidrográfico, realizado com um ecobatímetro monofeixe, a partir da sondagem de linhas de verificação, tendo discrepâncias que inicialmente não apresentem normalidade e/ou independência espacial.

iv. É viável o desenvolvimento de um novo estimador robusto, resistente a *outliers*, para estimar pontualmente a incerteza vertical amostral de levantamentos batimétricos.

v. É viável desenvolver e aplicar algoritmos computacionais para obtenção de amostras de discrepâncias que objetivem a avaliação estatística dos dados coletados por sistemas de sondagem por faixa, sem recorrer a acomodações teóricas, ou seja, interpolações matemáticas e/ou geoestatísticas.

vi. É possível estimar a qualidade vertical de um levantamento hidrográfico, realizada através de sistemas de sondagem por varrimento, tendo discrepâncias espacialmente dependentes e que não possuam uma distribuição normal.

vii. É viável a avaliação estatística das profundidades coletadas por sistemas de varrimento a partir de discrepâncias geradas via sobreposição de sucessivas varreduras regulares de sondagem, eliminando, assim, a necessidade de execução de linhas de verificação.

2. OBJETIVOS

Este trabalho objetiva, de forma geral, desenvolver estudos na área de controle de qualidade em levantamentos hidrográficos, especificadamente, o tratamento e avaliação estatística de dados batimétricos oriundos de sondagem monofeixe e multifeixe.

Como objetivos específicos têm-se:

i. Propor uma metodologia robusta e inovadora, teoricamente fundamentada na estatística clássica e Geoestatística, para detecção de *spikes* em dados batimétricos coletados a partir de sistemas de sondagem por faixa.
ii. Propor um novo limiar robusto para localização de *outliers*, inicialmente, aplicado em dados batimétricos.
iii. Propor uma metodologia para a avaliação estatística das profundidades coletadas através de sistemas de sondagem batimétrica monofeixe, abordando normalidade e independência espacial, bem como a presença, ou não, de *outliers* na base de dados.
iv. Propor um estimador robusto para cálculo da incerteza vertical amostral de levantamentos batimétricos, verificando a sua eficiência frente aos estimadores tradicionalmente utilizados no meio hidrográfico.
v. Propor um novo método para obtenção de amostras de discrepâncias de levantamentos hidrográficos realizados a partir de sistemas de sondagem por faixa, comparando-o com as técnicas usuais de extração de pontos homólogos.
vi. Adaptar e aplicar o método desenvolvido para avaliação estatística dos levantamentos batimétricos monofeixe à dados oriundos de sistemas de sondagem por varrimento.
vii. Realizar a estimativa da incerteza vertical amostral por meio de discrepâncias advindas da sobreposição de sucessivas varreduras regulares de sondagem, comparando os resultados com aqueles alcançados pelo método usual, isto é, por meio de varreduras de verificação.

3. JUSTIFICATIVAS E IMPORTÂNCIA

O principal objetivo dos levantamentos hidrográficos é produzir informações que suportam a segurança na navegação marítima e fluvial e a preservação do ambiente subaquático, assim como a sua defesa e exploração. Nesse sentido, o controle de qualidade do dado coletado torna-se cada vez mais importante. Avaliar a qualidade, bem como detectar, analisar e eliminar dados discrepantes, tornaram-se tarefas indispensáveis.

Com o advento dos sistemas de sondagem por faixa, o número de profundidades coletadas aumentou exponencialmente, melhorando a qualidade vertical das profundidades coletadas, porém, tornando o processamento tradicional bastante lento e subjetivo, quase sempre baseado num julgamento qualitativo e conservativo pelo hidrógrafo. Diante disso, o desenvolvimento de metodologias para detecção automatizada de *spikes* são altamente válidos pois diminuem, consideravelmente, o tempo e a subjetividade do processamento manual.

Devido à natureza dos dados batimétricos, a estimativa da qualidade vertical das sondagens não é tarefa simples, havendo na literatura inúmeras divergências entre metodologias de avaliação. A maioria das normativas de levantamentos hidrográficos, por exemplo, assumem, mesmo que implicitamente, que a base de dados é livre de *outliers*, possui distribuição normal e é espacialmente independente. Esses pressupostos, apesar de dificilmente atendidos e/ou verificados, são quase sempre assumidos com o objetivo de justificar o uso da estatística clássica (LI et al., 2005). Em vista disso, para um controle de qualidade estatisticamente coerente, torna-se necessário o desenvolvimento de metodologias para avaliação da incerteza do dado batimétrico, que considerem a independência e normalidade dos dados coletados, bem como a presença, ou não, de dados anômalos.

Tais exames e validações são sempre realizadas por meio de um conjunto de discrepâncias geradas a partir de pontos homólogos. Entretanto, como sabido, em levantamentos hidrográficos, redundâncias estão, geralmente, indisponíveis. Diante disso, artifícios teóricos e práticos são utilizados pela comunidade hidrográfica, principalmente, nos levantamentos batimétricos monofeixe. Já nas sondagens por varrimento, observa-se uma série de problemáticas inerentes as características da técnica. Assim, é nítida a necessidade do desenvolvimento de métodos para extração de pontos homólogos em dados batimétricos coletados por sistema de sondagem por

faixa, a fim de se proceder à avaliação estatística da qualidade vertical das profundidades.

4. ESTRUTURAÇÃO DO TRABALHO

Esta pesquisa visa, de modo geral, o estudo sobre o controle de qualidade em levantamentos hidrográficos, especificadamente, aqueles realizados através de sistemas de sondagem acústica. No Brasil, são poucos os estudos nessa área, embora, seja de extrema importância, principalmente, para a produção de cartas náuticas confiáveis.

O presente documento está dividido em cinco tópicos, conforme segue:

INTRODUÇÃO GERAL – apresenta uma introdução sobre o controle de qualidade em levantamentos hidrográficos, os objetivos gerais e específicos do projeto, justificativas e importância do presente estudo.

CAPÍTULO 1: *Metodologia robusta para detecção de spikes em dados batimétricos* – tem como objetivo propor metodologias e técnicas inovadoras para detecção de *spikes* em dados batimétricos coletados a partir de sistemas de sondagem por faixa, tais como: sistemas multifeixe e sonares interferométricos.

CAPÍTULO 2: *Proposta metodológica para avaliação da qualidade vertical de sondagens batimétricas monofeixe, com ênfase em testes de normalidade e independência* – tem como objetivo propor uma metodologia que permita a avaliação estatística das profundidades coletadas por sistemas de sondagem batimétrica monofeixe através de amostras de discrepâncias, abordando normalidade e independência, bem como a presença, ou não, de dados discrepantes *(outliers)*. Neste capítulo também é apresentado um estimador, resistente a *outliers*, para o cálculo da incerteza vertical amostral de levantamentos batimétricos.

CAPÍTULO 3: *Proposta metodológica para avaliação da qualidade vertical de dados batimétricos coletados a partir de sistemas de sondagem por faixa* – tem como objetivo propor um novo método para obtenção de amostras de discrepâncias de profundidades coletadas a partir de sistemas de sondagem por faixa. A partir dessas, avalia-se a qualidade vertical das sondagens aplicando uma metodologia baseada naquela proposta no Capítulo 2. Na fase de localização de discrepâncias anômalas, utiliza-se uma sutil adaptação do método proposto do Capítulo 1. Também é

apresentada, neste capítulo, uma alternativa à realização de varreduras de verificação para estimativa da incerteza vertical amostral.

CONCLUSÕES GERAIS – expõe as considerações finais e recomendações para trabalhos futuros.

REFERÊNCIAS BIBLIOGRÁFICAS

ANDRADE, J. B. **Fotogrametria**. 2ª ed. SBEE, 274p., 2003.

ARTILHEIRO, F. M. F. **Analysis and Procedures of Multibeam Data Cleaning for Bathymetric Charting.** M. Eng. report, Department of Geodesy and Geomatics Engineering, Technical Report n. 191, University of New Brunswick, Fredericton, New Brunswick, Canada, 140p., 1998.

ATHEARN, N.; TAKEKAWA, J.; JAFFE, B.; HATTENBACH, B. Mapping elevations of tidal wetlands restoration sites in San Francisco Bay: comparing accuracy of aerial Lidar with a singlebeam echosounder. **Journal of Coastal Research**, v. 26, n. 2, p. 312–319, 2010.

AYRES NETO, A. Uso da sísmica de reflexão de alta resolução e da sonografia na exploração mineral submarina. **Brazilian Journal of Geophysics**, v. 18, n. 3, p. 241-256, 2000.

BJØRKE, J. T. & NILSEN, S. Fast trend extraction and identification of spikes in bathymetric data. **Computers & Geosciences**, v. 35, n. 6, p. 1061-1071, 2009.

BOTTELIER, P.; BRIESE, C.; HENNIS, N.; LINDENBERGH, R.; PFEIFER, N. Distinguishing features from outliers in automatic Kriging-based filtering of MBES data: a comparative study. **Geostatistics for Environmental Applications**, **Springer**, p. 403-414, 2005.

CALDER, B. R. Automatic statistical processing of multibeam echosounder data. **The International Hydrographic Review**, v. 4, n. 1, p. 53-68, 2003.

CALDER, B. R. & MAYER, L. A. Automatic processing of high-rate, high-density multibeam echosounder data. **Geochemistry, Geophysics, Geosystems**, v. 4, n. 6, 2003.

CALDER, B. R. & SMITH, S. A time/effort comparison of automatic and manual bathymetric processing in real-time mode. In: Proceedings of the US Hydro 2003 Conference, **The Hydrographic Society of America**, Biloxi, MS. 2003.

CHAMBERS, J. M.; CLEVELAND, W. S.; KLEINER, B.;TUKEY, P. A. **Graphical Methods for Data Analysis**. Pacific Grove, CA: Wadsworth & Brooks/Cole, 1983.

CHENG, L.; MA, L.; CAI, W.; TONG, L.; LI, M.; DU, P. Integration of Hyperspectral Imagery and Sparse Sonar Data for Shallow Water Bathymetry Mapping**. Geoscience and Remote Sensing**. IEEE Transactions on, v. 53, n. 6, p. 3235-3249, 2015.

CLARKE, J. E. H. **A reassessment of vessel coordinate systems: what is it that we are really aligning?** In: US Hydrographic Conference. 2003.

CLARKE, J. E. H. **Imaging and Mapping II: Submarine Acoustic Imaging Methods**. Notes of classes. Ocean Mapping Group. University of New Brunswick. 2014.

CRUZ, J.; VICENTE, J.; MIRANDA, M.; MARQUES, C.; MONTEIRO, C.; ALVES, A. Benefícios da utilização de sondadores interferométricos. **3as Jornadas de Engenharia Hidrográfica**. Instituto Hidrográfico Português, Lisboa, Portugal, 2014.

DEBESE, N. & BISQUAY, H. Automatic detection of punctual errors in multibeam data using a robust estimator. **The International Hydrographic Review**, v. 76 n. 1, p. 49-63, 1999.

DEBESE, N. Multibeam Echosounder Data Cleaning Through an Adaptive Surface-based Approach. **In: US Hydro 07 Norfolk**, 18p., 2007.

DEBESE, N.; MOITIÉ, R.; SEUBE, N. Multibeam echosounder data cleaning through a hierarchic adaptive and robust local surfacing. **Computers & Geosciences**, v. 46, p. 330-339, 2012.

DHN – Diretoria de Hidrografia e Navegação. **NORMAM 25: Normas da Autoridade Marítima para Levantamentos Hidrográficos**. Marinha do Brasil, Brasil, 52p., 2014.

EEG, J. On the identification of spikes in soundings. **The International Hydrographic Review**, v. 72, n. 1, p. 33-41, 1995.

EEG, J. Multibeam Crosscheck Analysis: A Case Study. **The International Hydrographic Review**, n. 4, p. 25-33, 2010.

ELLMER, W.; ANDERSEN, R. C.; FLATMAN, A.; MONONEN, J.; OLSSON, U.; ÖIÅS, H. Feasibility of Laser Bathymetry for Hydrographic Surveys on the Baltic Sea. **The International Hydrographic Review**, n. 12, p. 33-50, 2014.

FERREIRA, Í. O. ; RODRIGUES, D. D. ; SANTOS, A. P. Levantamento batimétrico automatizado aplicado à gestão de recursos hídricos. Estudo de caso: represamento do ribeirão São Bartolomeu, Viçosa-MG. **In: IV Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação**, 2012, Recife. Geotecnologias para o Planejamento e a Gestão Eficiente do território, 2012.

FERREIRA, Í. O.; RODRIGUES, D. D.; NETO, A. A.; MONTEIRO, C. S. Modelo de incerteza para sondadores de feixe simples**. Revista Brasileira de Cartografia**, v. 68, n. 5, p. 863-881, 2016a.

FERREIRA, Í. O.; NETO, A. A.; MONTEIRO, C. S. O uso de embarcações não tripuladas em levantamentos batimétricos**. Revista Brasileira de Cartografia**, v. 68, n. 10, p. 1885-1903, 2017a.

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.; **Coleta, processamento e análise de dados batimétricos**. 1ª ed. Saarbrucken: Novas Edições Acadêmicas, v. 1, 100p., 2015.

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.; ROSA, L. M. F. In bathymetric surfaces: IDW or Kriging? **Boletim de Ciências Geodésicas**, v. 23, n. 3, p. 493-508, 2017b.

FERREIRA, Í. O.; SANTOS, G. R.; RODRIGUES, D. D. **Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas**. Revista Brasileira de Cartografia, Rio de Janeiro, v. 65, n. 5, p. 831-842, 2013.

FERREIRA, Í. O.; ZANETTI, J.; GRIPP, J. S.; MEDEIROS, N. G. **Viabilidade do uso de imagens do sistema Rapideye na determinação da batimetria de águas rasas**. Revista Brasileira de Cartografia, v. 68, n. 7, p. 1331-1340, 2016b.

GAO, J. Bathymetric mapping by means of remote sensing: methods, accuracy and limitations. **Physical Geography**, v. 33, n. 1, p. 103-116, 2009.

GUENTHER, G. C.; THOMAS, R. W. L. ; LAROCQUE, P. E. Design considerations for achieving high accuracy with the Shoals bathymetric Lidar system. In: CIS Selected Papers: Laser Remote Sensing of Natural Waters-From Theory to Practice. **International Society for Optics and Photonics**, p. 54-71, 1996.

HARE, R. Depth and position error budgets for multibeam echosounding. **The International Hydrographic Review**, v. 72, n. 2, p. 37-69, 1995.

HARE, R.; EAKINS, B.; AMANTE, C. Modelling bathymetric uncertainty. **The International Hydrographic Review,** n. 6, p. 31-42, 2011.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. **Understanding robust and exploratory data analysis**. New York: Wiley, 433p., 1983.

HUBERT, M. & VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Journal of Computational statistics & data analysis**, v. 52. n. 12, p. 5186-5201, 2008.

IGLEWICZ, B. & HOAGLIN, D. **How to detect and handle outliers**. Milwaukee, Wis.: ASQC Quality Press, 87p., 1993.

IHO – International Hydrographic Organization. **C-13: IHO Manual on Hydrography**. Mônaco: International Hydrographic Bureau, 540p., 2005.

IHO – International Hydrographic Organization. **S-44: IHO Standards for Hydrographic Surveys**. Special Publication n. 44 – 5th. Mônaco: International Hydrographic Bureau, 36p., 2008.

INMETRO – Instituto Nacional de Metrologia Normalização, Qualidade e Tecnologia. **Avaliação de dados de medição: guia para a expressão de incerteza de medição (GUM 2008)**. Duque de Caxias, RJ: INMETRO/CICMA/SEPIN, 141 p., 2012a.

INMETRO – Instituto Nacional de Metrologia Normalização, Qualidade e Tecnologia. **Vocabulário Internacional de Metrologia: conceitos fundamentais e gerais de termos associados (VIM 2012)**. Duque de Caxias, RJ : INMETRO, 94 p., 2012b.

INSTITUTO HIDROGRÁFICO. **Especificação Técnica para Produção de cartografia hidrográfica**. Marinha Portuguesa, Lisboa, Portugal, v. 0, 24p., 2009.

JONG, C. D.; LACHAPELLE, G.; SKONE, S.; ELEMA, I. A. **Hydrography**. 2ª ed. Delft University Press: VSSD, 354p., 2010.

KIRKUP, L. & FRENKEL, R. B. **An introduction to uncertainty in measurement: using the GUM (guide to the expression of uncertainty in measurement)**. Cambridge University Press, 2006.

LI, Z.; ZHU, Q.; GOLD, C. M. **Digital terrain modelling. Principles and methodology**. New York: CRC Press, 319p., 2005.

LINZ – Land Information New Zealand. **Contract Specifications for Hydrographic Surveys**. New Zealand Hydrographic Authority, V. 1.2, 111p., 2010.

LU, D.; LI, H.; WEI, Y.; ZHOU, T. Automatic outlier detection in multibeam bathymetric data using robust LTS estimation. In: 3rd International Congress on Image and Signal Processing (CISP), **IEEE**, v. 9, p. 4032-4036, 2010.

MALEIKA, W. The influence of the grid resolution on the accuracy of the digital terrain model used in seabed modeling. **Marine Geophysical Research**, v. 36, n. 1, p. 35-44, 2015.

MAUNE, D. F. Digital Elevation Model Technologies and Applications: The DEM Users Manual. **American Society for Photogrammetry and Remote Sensing**, 2007.

MIKHAIL, E. & ACKERMAN, F. **Observations and Least Squares**. University Press of America, 497p., 1976.

MONICO, J. F. M.; DAL POZ, A. P.; GALO, M.; SANTOS, M. C.; OLIVEIRA, L. C. Acurácia e Precisão: Revendo os conceitos de forma acurada. **Boletim de Ciências Geodésicas**, v. 15, n. 3, p. 469-483, 2009.

MORETTIN, P. A. & BUSSAB, W. O. **Estatística básica**. 5ª ed. São Paulo: Editora Saraiva, 526p., 2004.

MOTAO, H.; GUOJUN, Z.; RUI, W.; YONGZHONG, O.; ZHENG, G. Robust method for the detection of abnormal data in hydrography. **The International Hydrographic Review**, v. 76, n. 2, p. 93-102, 1999.

MOURA, A.; GUERREIRO, R.; MONTEIRO, C. As potencialidades da derivação de batimetria a partir de imagens de satélite multiespetrais na produção de cartografia náutica. **4as Jornadas de Engenharia Hidrográfica**. Instituto Hidrográfico Português, Lisboa, Portugal, 2016.

NOAA – National Oceanic and Atmospheric Administration. **Field Procedures Manual**. Office of Coast Survey, 2011.

PASTOL, Y. Use of Airborne lidar Bathymetry for Coastal Hydrographic Surveying: The French Experience. **Journal of Coastal Research**, n. 62, p. 6-18, 2011.

RODRIGUES, D. D. **Topografia: Planimetria para Engenheiros Agrimensores e Cartógrafos**. Apostila. Universidade Federal de Viçosa. 2008.

SANTOS, A. M. R. T.; SANTOS, G. R.; EMILIANO, P. C.; MEDEIROS, N. G.; KALEITA, A. L.; PRUSKI, L. O. S. Detection of inconsistencies in geospatial data with geostatistics. **Boletim de Ciências Geodésicas,** v. 23, n. 2, p. 296-308, 2017.

SANTOS, A. P. **Controle de qualidade cartográfica: metodologias para avaliação da acurácia posicional em dados espaciais**. Tese (Doutorado). Programa de Pós-Graduação em Engenharia Civil, Departamento de Engenharia Civil, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 172p., 2015.

SANTOS, A. P.; RODRIGUES, D. D.; SANTOS, N. T.; GRIPP JUNIOR, J. Avaliação da acurácia posicional em dados espaciais utilizando técnicas de estatística espacial: proposta de método e exemplo utilizando a norma brasileira. **Boletim de Ciências Geodésicas**, v. 22, n. 4, p. 630-650, 2016.

SEO, S. **A review and comparison of methods for detecting outliers in univariate data sets**. Master Of Science, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, USA, 59p., 2006.

SUSAN, S. & WELLS, D. Analysis of Multibeam Crosschecks Using Automated Methods. **In: US Hydro 2000 Conference paper**, Biloxi, Mississippi. 2000.

TUKEY, J.W. **Exploratory Data Analysis**. Princeton, Ed. Pearson (1977).

URICK, R. I. **Principles of Underwater Acoustics.** Toronto: McGraw-Hill, 1975.

USACE – U.S. Army Corps of Engineers. **HYDROGRAPHIC SURVEYING**. Engineer Manual n. 1110-2-1003. Department of the Army. Washington, D. C., USA, 2002.

USACE – U.S. Army Corps of Engineers. **Hydrographic Surveying**. Engineer Manual n. 1110-2-1003. Department of the Army. Washington, D. C. USA, 2013.

VICENTE, J. P. D. **Modelação de dados batimétricos com estimação de incerteza**. Dissertação (Mestrado). Programa de Pós-Graduação em Sistemas de Informação Geográfica Tecnologias e Aplicações, Departamento de Engenharia Geográfica, Geofísica e Energia, Universidade de Lisboa, Portugal, 158p., 2011.

WARE, C.; SLIPP, L.; WONG, K. W.; NICKERSON, B.; WELLS, D. E.; LEE, Y. C.; DODD, D.; COSTELLO, G. A System for Cleaning High Volume Bathymetry. **The International Hydrographic Review**, v. 69. n. 2, p. 77-94, 1992.

WARE, C.; KNIGHT, W.; WELLS, D. Memory intensive statistical algorithms for multibeam bathymetric data. **Computers & Geosciences**, v. 17, n. 7, p. 985-993, 1991.

WEEKLEY, R. A.; GOODRICH, R. K.; CORNMAN, L. B. An algorithm for classification and outlier detection of time-series data**. Journal of Atmospheric and Oceanic Technology**, v. 27, n. 1, p. 94-107, 2010.

CAPÍTULO 1. METODOLOGIA ROBUSTA PARA DETECÇÃO DE *SPIKES* EM DADOS BATIMÉTRICOS

Resumo:
Atualmente, os sistemas de sondagem por varrimento são capazes de coletar milhares de pontos em um pequeno intervalo de tempo, promovendo uma maior cobertura do fundo submerso, com consequente aumento na capacidade de detecção de objetos. Embora tenha ocorrido uma melhora na acurácia das profundidades coletadas, o processamento na sua forma tradicional ainda é requerido. Contudo, devido principalmente ao aumento da massa de dados coletada, o processamento manual tornou-se extremamente moroso e subjetivo, especialmente, na fase de detecção e eliminação de *spikes.* Diversos são os algoritmos que se propõe a executar tal tarefa, todavia, a maioria deles são de difícil aplicação ou encontram-se implementados apenas em pacotes comerciais. Nesse sentido, o objetivo deste estudo é apresentar a metodologia AEDO (Algoritmo Espacial para Detecção de *Outliers*), um novo método de detecção de *spikes* concebido para tratar dados batimétricos coletados através de sistemas de sondagem por faixa.

1. INTRODUÇÃO

A coleta de profundidades é tarefa essencial em diversas áreas, com destaque para aquelas relacionadas a produção e atualização da cartografia náutica. Diferentemente das ondas eletromagnéticas, as ondas acústicas apresentam uma boa propagação nos meios aquáticos e, por este motivo, a maioria dos sensores utilizados na determinação da profundidade utilizam ondas sonoras, tais como: ecobatímetros monofeixe, multifeixe e sonares interferométricos (IHO, 2005, FERREIRA, et al., 2017a).

Apesar da atenuação sofrida pelas ondas eletromagnéticas, os sistemas de sondagem laser também têm sido utilizados no mapeamento batimétrico, destacando-se, principalmente, pelo grande ganho de produtividade (GUENTHER et al., 1996; IHO, 2005; PASTOL, 2011; ELLMER et al., 2014). O uso de imagens orbitais para estimar a batimetria em águas rasas também vem sendo objeto de pesquisa (GAO, 2009; CHENG et al., 2015; MOURA et al., 2016; FERREIRA et al., 2016b).

Porém, num cenário atual, a realização de levantamentos hidrográficos, principalmente aqueles destinados a atualização cartográfica, restringe-se ao uso de ecobatímetros multifeixe e sonares interferométricos. Em comparação com os sondadores de feixe simples, esses sistemas apresentam um elevado ganho em resolução e acurácia, tanto em termos planimétricos quanto altimétricos

(profundidade), e um grande adensamento de dados, descrevendo quase que por completo o fundo submerso, melhorando, inclusive, a capacidade de detecção de objetos (CRUZ et al., 2014; MALEIKA, 2015). Sistemas menos eficientes já são capazes de coletar em águas rasas mais de 30 milhões de pontos por hora (BJØRKE & NILSEN, 2009).

Enquanto os sistemas de batimetria monofeixe realizam um único registro de profundidade a cada pulso acústico transmitido *(ping)*, tendo como resultado uma linha de pontos imediatamente abaixo da trajetória da embarcação, o sistema de sondagem por varrimento executa diversas medidas de profundidade com um mesmo *ping*, obtendo medições da coluna d'água em uma faixa *(swath)* perpendicular à trajetória da embarcação. Um número crescente de serviços hidrográficos adotou a tecnologia multifeixe como a metodologia principal para coleta de dados batimétricos visando a produção cartográfica (IHO, 2008; INSTITUTO HIDROGRÁFICO, 2009; LINZ, 2010; NOAA, 2011; USACE, 2013; DHN, 2014). Os sonares interferométricos são uma tecnologia relativamente nova, porém passível de alcançar resultados semelhantes ou superiores aos da batimetria multifeixe, com vantagens, principalmente, na cobertura de fundo em águas rasas (CRUZ et al., 2014).

Embora os sistemas de sondagem por faixa tragam uma melhoria na resolução e acurácia da batimetria, o processamento dos dados na sua forma tradicional tornou-se mais moroso do que o próprio levantamento. Dentre as diversas fases merece destaque a detecção, análise e eliminação de dados discrepantes *(spikes)* (WARE et al., 1992; ARTILHEIRO, 1998; CALDER & MAYER, 2003; CALDER & SMITH, 2003; BJØRKE & NILSEN, 2009; VICENTE, 2011). O termo *outlier* pode ser definido como uma observação que, estatisticamente, se diferencia do conjunto de dados ao qual pertence, ou seja, é um valor atípico ou inconsistente (SANTOS et al., 2017). Nesse sentido, *outliers* podem ser causados por erros grosseiros, por efeitos sistemáticos ou, simplesmente, por efeitos aleatórios (SANTOS et al., 2016). Em levantamentos hidrográficos, profundidades que se configuram como *outliers* são conhecidos como *spikes*, enquanto que os erros de posicionamento são chamados de *tops*. Este trabalho foca, especificadamente na componente vertical e, por esse motivo o termo *spike* por vezes é tratado como sinônimo de *outlier.* A Figura 1 ilustra um perfil batimétrico na presença de *spikes.*

Na sondagem batimétrica os valores anômalos são causados, principalmente, pelo desempenho deficiente dos algoritmos utilizados pelo ecobatímetro para detecção

de fundo (detecção por fase, amplitude, transformada de *Fourier*, *etc*.), pela detecção por lóbulos secundários, reflexões múltiplas, presença de bolhas de ar na face do conjunto de transdutores, por reflexões na coluna d'água e, até mesmo, por equipamentos operando simultaneamente na mesma frequência (URICK, 1975; JONG et al., 2010).

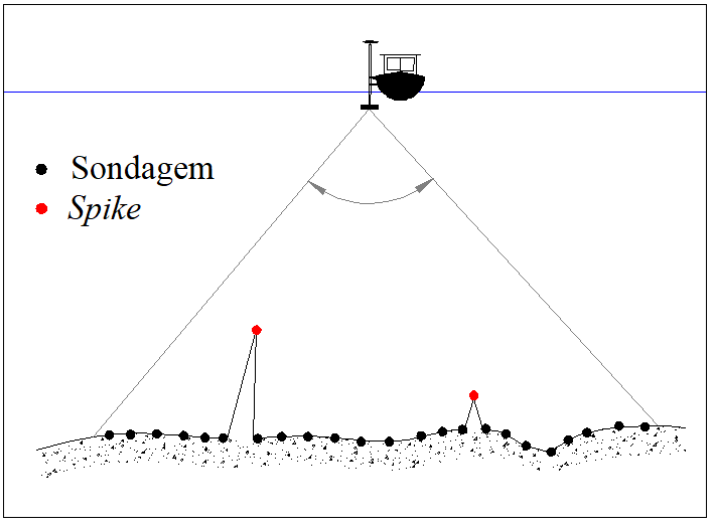

Figura 1 – Perfil batimétrico eivado de *spikes*.

Geralmente, a detecção, análise e eliminação de *spikes* são realizadas manualmente pelo hidrógrafo que, visualizando os dados através de uma interface gráfica, decide subjetivamente qual sondagem pode, ou não, ser considerada um valor anormal. A *priori* esta tarefa pode parecer simples, visto que *spikes* são pontos aleatórios que não representam a superfície de fundo, destoando, visivelmente, dela e devendo, nesses casos, serem eliminados. Todavia, devido ao grande volume de dados provenientes de uma sondagem multifeixe, essa tarefa tornou-se muito demorada e ainda mais subjetiva (WARE et al., 1992; CALDER & SMITH, 2003). É importante destacar que a análise dos pontos anômalos não deve ser ambígua, isto é, ora são interpretados como dados espúrios, ora como pertencente a superfície de fundo.

Desse modo, com o objetivo de facilitar a tarefa do hidrógrafo, diversos autores desenvolveram pesquisas na área de detecção de *spikes* em dados de sondagem batimétrica, como por exemplo: Ware et al. (1991; 1992), Eeg (1995), Debese & Bisquay (1999), Motao et al. (1999), Calder & Mayer (2003), Bottelier et al. (2005), Debese (2007), Bjørke & Nilsen (2009), Lu et al. (2010) e Debese et al. (2012).

Os primeiros algoritmos baseavam-se na geração de superfícies batimétricas, principalmente obtidas a partir de funções polinomiais ou de médias ponderadas, seguidos da utilização de filtros para a detecção e eliminação de *outliers* (WARE et

al., 1991; EEG, 1995). Com o incremento da tecnologia computacional, algoritmos mais robustos foram desenvolvidos, baseados em estimadores-M (DEBESE & BISQUAY, 1999; MOTAO et al., 1999; DEBESE, 2007; DEBESE et al. 2012), filtros de *Kalman* (CALDER & MAYER, 2003), técnicas de Krigagem (BOTTELIER et al., 2005), superfícies de tendência (BJØRKE & NILSEN, 2009) e estimador LTS *(Least Trimmed Squares)* (LU et al., 2010).

Dentre os vários procedimentos sobressai-se o algoritmo CUBE *(Combined Uncertainty and Bathymetry Estimator),* apresentado por Calder (2003). Esse algoritmo encontra-se implementado nos principais pacotes hidrográficos e é, talvez por isso, a ferramenta semiautomatizada mais utilizada no processamento de dados multifeixe (VICENTE, 2011), incluindo a pesquisa por *spikes*. Todavia, essas metodologias são, em sua maioria, de difícil aplicação, semiautomatizadas ou encontram-se implementadas apenas em pacotes comerciais. Além do mais, a maioria desses métodos são fundamentados em pressuposições teóricas dificilmente atendidas e quase nunca verificadas. Conforme afirma Vicente (2011), o problema dos algoritmos, exceto no caso do CUBE (Calder, 2003; Calder & Mayer, 2003), permanece na incapacidade destes em estimarem a incerteza associada a profundidade reduzida.

Uma técnica baseada na análise de resíduos padronizados foi recentemente apresentada por Santos et al. (2017) para dados de altimetria terrestre. A metodologia, apesar de não ser automatizada, uma vez que necessita de uma análise geoestatística, apresentou-se bastante eficiente para detecção de *outliers*. Todavia, a natureza da técnica impõe um certo padrão de subjetividade ao processo, sobretudo, na etapa de modelagem do semivariograma, que conforme exposto por Ferreira et al. (2013), é uma fase crucial do processo de modelagem geoestatística e não deve ser automatizada ou negligenciada.

Sob outra perspectiva, na estatística clássica umas das ferramentas mais utilizadas para detecção de *outliers* em conjunto de dados contínuos univariados é o diagrama *boxplot* ou gráfico de caixa (TUKEY, 1977; CHAMBERS et al., 1983; HOAGLIN et al., 1983). Outro método comumente empregado é o *Z-Score Modificado*, que ao contrário do tradicional *Z-Score*, utiliza estatísticas robustas, como a mediana e o desvio absoluto da mediana *(median absolute deviation),* que podem garantir que os valores de corte não foram afetados, justamente, pela presença de *outliers* (IGLEWICZ & HOAGLIN, 1993). Diversos outros métodos podem ser

aplicados para detecção de valores anômalos em conjuntos de dados univariados, compostos por variáveis quantitativas contínuas, tal como resumido em Seo (2006).

O problema da aplicação dessas metodologias reside no fato destas, além de desconsiderarem a localização espacial do dado analisado, assumirem como pressupostos básicos que as observações são variáveis aleatórias independentes e identicamente distribuídas (MOOD et al., 1974; MORETTIN & BUSSAB, 2004; SEO, 2006), pressuposições indispensáveis para um tratamento estatístico clássico e coerente, porém, dificilmente atendidas ou teoricamente comprovadas. Além do mais, na maioria dessas técnicas, os valores de corte para detecção de *outliers* derivam da distribuição normal, o que reduz a eficiência dos métodos quando a distribuição amostral é assimétrica (HUBERT & VANDERVIEREN, 2008). Todavia, são mecanismos de simples aplicação e análise.

Sendo assim, pode-se vislumbrar a possibilidade do desenvolvimento e aplicação de métodos para detecção automatizada de *spikes* através da utilização destes mecanismos em dados de batimetria, desde que as metodologias desenvolvidas levem em consideração os pressupostos estatísticos básicos e a estrutura de dependência espacial, inerente aos dados espacialmente contínuos. Tendo isso em vista, a Geoestatística apresenta-se como uma potencial ferramenta de suporte, dado suas características ideais, isto é, modelagem espacial sem tendência e com variância mínima, atributos que podem robustecer quaisquer técnicas de detecção de *outliers* (SANTOS et al., 2017). Tais características foram, ainda, confirmadas por Ferreira et al. (2013, 2015 e 2017b) durante estudos para modelagem de superfícies batimétricas. Assim, a Geoestatística pode ser empregada como ferramenta de suporte as técnicas e algoritmos desenvolvidos neste estudo.

Diante do exposto, o objetivo principal deste trabalho é propor uma nova metodologia de detecção de *spikes* para dados batimétricos coletados por sistemas de sondagem por faixa, denominada AEDO (Algoritmo Espacial para Detecção de *Outliers*). O método proposto emprega três técnicas ou limiares de detecção de *outliers*, a saber: o *Boxplot Ajustado*, *Z-Score Modificado* e o *Método* δ. Este último também desenvolvido e apresentado neste trabalho. Visando robustecer a metodologia, toda a fundamentação teórica baseia-se em teoremas da estatística clássica e Geoestatística.

2. PRINCIPAIS FONTES DE *SPIKES* EM SONDAGENS BATIMÉTRICAS

O processo de determinação da profundidade baseia-se na integração de diversas medições individuais além daquelas efetivamente realizadas pelo ecobatímetro, tais como: a profundidade de imersão do transdutor, a altura de maré, a atitude da embarcação de sondagem, o perfil de velocidade do som na água *etc*. O ecobatímetro mede, de um modo geral, apenas o tempo decorrido desde o instante em que um pulso acústico é transmitido na água até o momento em que ele retorna ao transdutor após refletir-se no fundo, a direção de onde provém esse retorno e a intensidade do sinal. Sendo assim, diversas são as fontes de incertezas envolvidas neste processo, que podem, em quase sua totalidade, serem causadores direta ou indiretamente de medições anômalas de profundidade (FERREIRA et al., 2016a).

Ao focar-se na propagação da onda sonora, em especial nos sistemas de sondagem por faixa, depara-se com um processo bastante complexo e que envolve inúmeros princípios físicos e geométricos, que podem, dependendo de uma série de condições, causar dados anormais na medição de profundidade.

É evidente que dependendo das profundidades, características de fundo e potências envolvidas, as ondas sonoras poderão sofrer várias reflexões entre a superfície e o fundo, retornando, em alguns casos, ao transdutor, o que poderá gerar ruídos ou medições inconsistentes, sendo caracterizados como um *outlier*. A reflexão da onda sonora também pode causar o efeito da reverberação que, ao contrário do eco, ocorre quando o intervalo de tempo não é suficiente para se distinguir o som refletido do som transmitido (LURTON, 2002; IHO, 2005; JONG et al., 2010).

Durante a propagação, o nível de intensidade do eco diminui rapidamente com o tempo em consequência das perdas por transmissão sofrida pela onda sonora na interação com o meio. As perdas por transmissão, um dos muitos fenômenos associados com a propagação do som na água, também causa a deformação do sinal para uma frequência ligeiramente menor (JONG et al., 2010). Devido a isso, durante a recepção, o nível do eco deve ser amplificado através do ganho e do ganho variado no tempo *(TVG – Time Varying Gain)* (CHU & HUFNAGLE JR., 2006). Deve-se atentar que a amplificação do sinal, também amplifica os ruídos, gerando dados incosistentes. O ajuste do ganho depende, basicamente, do tipo de fundo e do nível de intensidade do pulso transmitido (potência). O ajuste da potência do pulso transmitido

define a quantidade de energia da onda sonora e é dependente de fatores como profundidade, frequência e tipo de fundo. Em suma, a potência deve ser mantida em níveis mínimos, mas que garantam a correta detecção de fundo. Potências elevadas do sinal conduzirão ao efeito da reverberação, que por sua vez podem gerar dados ruidosos.

A superfície da água e o fundo causam reflexão, como visto anteriormente, mas podem também dispersar a onda sonora. A dispersão do som na superfície é influenciada pela presença de bolhas de ar e pelas condições da superfície da água (rugosidade). No caso do fundo submerso, a dispersão pode ocorrer devido ao tipo de fundo (composição e rugosidade), ao ângulo de incidência do feixe e a frequência de operação do sondador (JONG et al., 2010). Durante a propagação, a dispersão da onda sonora pode ser causada pela presença de partículas ou corpos presentes na coluna d'água (URICK, 1975).

Dependendo das dimensões e da impedância desses alvos, durante o processo de dispersão poderá ocorrer a reflexão da onda sonora de volta ao transdutor, essa parcela de energia que retorna a fonte é chamada de *backscatter*. Técnicas de processamento de sinal utilizadas pelo ecobatímetro, poderão, em alguns casos, detectar esses retornos que, provavelmente, apresentaram-se como profundidades discrepantes, ou seja, *spikes*.

O padrão de transmissão dos feixes sonoros de um sistema de sondagem por faixa, associado a variação do perfil de velocidade do som, causa problemas de refração muito aparente nos feixes mais externos, implicando em dados ruidosos. Soma-se a isso o fato da determinação da velocidade do som ser talvez o fator mais crítico numa sondagem batimétrica, devido a sua variação temporal, local (à superfície) e ao longo da coluna d'água (USACE, 2013).

Para determinar o caminho percorrido pelo feixe, o sistema de sondagem por faixa utiliza modelagens matemáticas baseadas, especificadamente, na *lei de Snell-Descartes,* constatando novamente, que a medição da profundidade é altamente dependente do perfil de propagação da velocidade do som. Seguindo a teoria apresentada por *Snell-Descartes,* feixes centrais sofrem pouco com o efeito da refração, por esse motivo, o sistema monofeixe é capaz de estimar a profundidade apenas utilizando uma velocidade do som média, geralmente uma média harmônica, uma vez que as variáveis envolvidas são inversamente proporcionais.

Os levantamentos hidrográficos são conduzidos em condições dinâmicas a partir de embarcações que possuem, durante o seu deslocamento, seis possíveis movimentos: três translações e três rotações (IHO, 2005; CLARKE, 2014). Sendo assim, para a medição de profundidade através de sistemas de varrimento, tal como o ecobatímetro multifeixe, é imprescindível a determinação, preferencialmente em tempo real, da atitude da plataforma de sondagem, dada pelas rotações: *roll* (balanço), *pitch* (caturro/cabeceio/arfagem) e *heading* (proa/giro/guinada), e da translação ao longo do eixo vertical (*heave*/afundamento) (JONG et al., 2010; USACE, 2013). Na hipótese desses efeitos não serem compensados ou serem maiores que a resolução ou precisão nominal de medida dos sensores utilizados na mensuração, estes poderão, mesmo que de forma indireta, serem agentes causadores de *spikes*.

Sendo assim, nota-se que de um modo geral, os *spikes* são causados, dentre outros, pela ineficiência do ecobatímetro multifeixe e do seu algoritmo de detecção de fundo (detecção por fase, amplitude, transformada de *Fourier etc*.), pela detecção por lóbulos secundários, pela ineficácia dos instrumentos de medição da velocidade de propagação do som ao longo da coluna d'água e à face do conjuntos de transdutores, por falhas na compensação da atitude da plataforma de sondagem, por reflexões múltiplas, pela presença de bolhas de ar na face do transdutor e por reflexões na coluna d'água causadas, principalmente, por: algas, cardumes de peixe, *deep scattering layer*, variações térmicas e sedimentos em suspensão (URICK, 1975; JONG et al., 2010).

3. MÉTODOS PARA DETECÇÃO DE *OUTLIERS* EM CONJUNTOS DE DADOS UNIVARIADOS

O *boxplot*, método introduzido por Tukey (1977), é uma ferramenta gráfica simples e muito utilizada durante a análise exploratória de dados para a detecção de possíveis *outliers* (TUKEY, 1977; CHAMBERS et al., 1983; HOAGLIN et al., 1983). De acordo com Santos et al. (2017), devido a simplicidade de construção, utilização e interpretação, diversos trabalhos apresentam essa metodologia como principal ferramenta para detecção de dados discrepantes.

A sua construção consiste em montar gráficos em formas de caixas onde são representadas a mediana ou segundo quartil ($Q2$), o primeiro quartil ($Q1$) e o terceiro quartil ($Q3$), além dos dados (CHAMBERS et al., 1983). De forma conjunta, exibe-se também os limites superior (LS) e inferior (LI) através de linhas retas verticais que se

originam em $Q1$ e $Q3$, respectivamente. Esses seguimentos são conhecidos como *"Whisker"* ou, simplesmente, "fio de bigode" (CHAMBERS et al., 1983; WILLIAMSON, 1989). As informações que se encontram fora dos limites são consideradas possíveis *outliers.* A largura da caixa pode ser utilizada para se avaliar a dispersão dos dados, assim como a medida de desvio padrão (HOAGLIN et al., 1983).

A Figura 2 ilustra um exemplo do formato de um *boxplot* onde não houve a detecção de *outliers*, enquanto a Figura 3 apresenta dados eivados de possíveis *outliers,* representados por círculos acima e abaixo, respectivamente, dos limites superior e inferior.

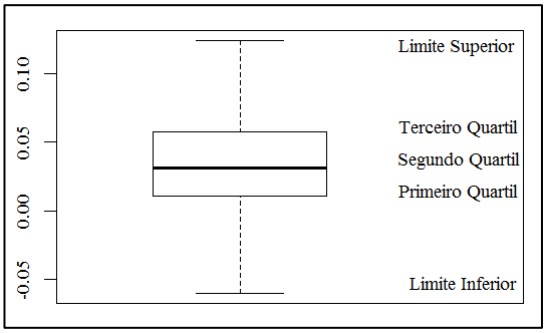

Figura 2 – Gráfico *boxplot* construído com base em um conjunto de dados sem *outliers*.

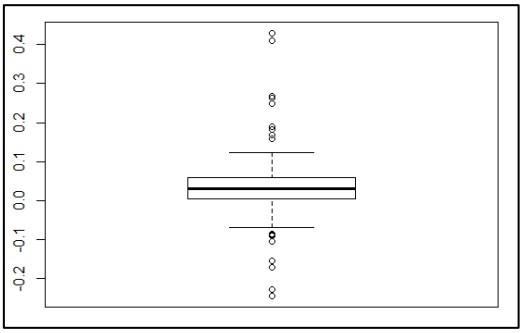

Figura 3 – Gráfico *boxplot* construído com base em um conjunto de dados com possíveis *outliers.*

Os limites superior e inferior, ou valores de corte, são dados, respectivamente, pelas Equações (1) e (2) (TUKEY, 1977; HUBERT & VANDERVIEREN, 2008).

$$LS = Q3 + 1{,}5 \cdot AIQ \quad (1)$$

$$LI = Q1 - 1{,}5 \cdot AIQ \quad (2)$$

em que AIQ corresponde ao intervalo interquartílico, ou seja, $Q3 - Q1$. Os limites de corte estabelecidos por Tukey (1977) são baseados na distribuição normal padrão, conforme a justificativa a seguir. Considerando uma curva normal com média zero e variância unitária, têm-se: $Q1 = -0{,}6745$, $Q2 = 0$, $Q3 = 0{,}6745$ e, portanto, $AIQ = 1{,}349$. Sendo assim, pelas Equações (1) e (2), os valores de corte seriam $LI = -2{,}698$ e $LS = 2{,}698$, que correspondem a uma área abaixo da curva normal igual a 0,993, ou 99,3% (Figura 4). Isso significa que, para uma $N(0{,}1)$, 0,7% das observações são de ocorrência rara, podendo ser consideradas possíveis *outliers* (MORETTIN & BUSSAB, 2004).

Deve-se atentar que, embora a metodologia proposta por Tukey (1977) seja fundamentada em estatísticas robustas, os limites definidos para detecção de valores anômalos supõem que os dados sigam uma distribuição normal. Neste sentido, a sua

aplicação a dados assimétricos ou que possuam algum tipo de afastamento da normalidade seja realizada com cuidado. Para contornar este problema, Vandervieren & Hubert (2004) introduziram o *boxplot ajustado*, que leva em conta a assimetria da variável em que os dados são amostrados.

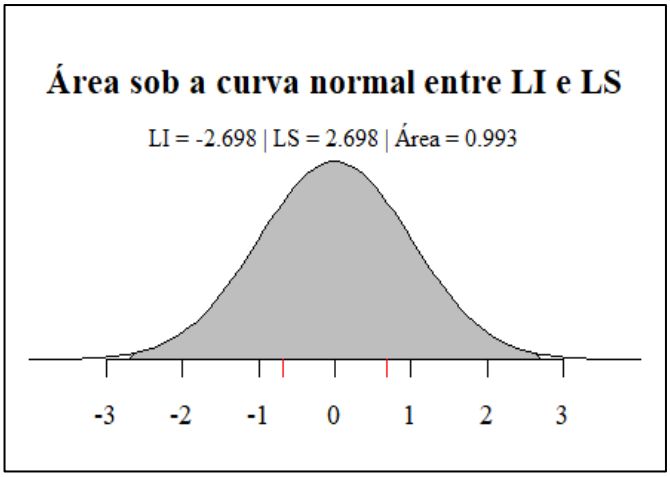

Figura 4 – Valores de corte para detecção de *outliers* via gráfico *boxplot*. A área sombreada corresponde aos valores aceitáveis. As marcações em vermelho ilustram $Q1$ e $Q3$.
Fonte: Adaptado de Morettin & Bussab (2004).

O *boxplot ajustado* é baseado na estatística *medcouple (MC)*, uma medida robusta de assimetria, introduzida por Brys et al. (2004). O MC é semelhante ao *Quartile Skewness* (BOWLEY, 1920) e ao *Octile Skewness* (HINKLEY, 1975).

De acordo com Seo (2006), quando $X_n = \{x_1, x_2, \dots, x_n\}$ é um conjunto de dados contínuos univariados, amostrados independentemente e devidamente ordenados, isto é, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, a medida MC dos dados é definida como:

$$MC\left(x_{(1)}, x_{(2)}, \dots, x_{(n)}\right) = mediana\left(\frac{(x_{(j)} - med_k) - (med_k - x_{(i)})}{x_{(j)} - x_{(i)}}\right) \quad (3)$$

em que med_k é a mediana de X_n, i e j são subgrupos de X_n formados, respectivamente, pelos relativos menores que a mediana e os relativos maiores que a mediana, ou seja, $x_{(i)} \leq med_k \leq x_{(j)}$, e $x_{(i)} \neq x_{(j)}$.

Computada a medida de assimetria *medcouple*, os limites superior e inferior para detecção de *outliers* são obtidos através das seguintes equações (VANDERVIEREN & HUBERT, 2004; HUBERT & VANDERVIEREN, 2008):

$$MC \geq O \rightarrow \begin{cases} LS = Q3 + 1{,}5 \cdot e^{3 \cdot MC} \cdot AIQ \\ LI = Q1 - 1{,}5 \cdot e^{-4 \cdot MC} \cdot AIQ \end{cases} \quad (4)$$

$$MC < O \rightarrow \begin{cases} LS = Q3 + 1{,}5 \cdot e^{4 \cdot MC} \cdot AIQ \\ LI = Q1 - 1{,}5 \cdot e^{-3 \cdot MC} \cdot AIQ \end{cases} \quad (5)$$

O valor de MC varia entre -1 e 1. Quando $MC = 0$ a distribuição dos dados é simétrica e o *boxplot ajustado* torna-se idêntico ao método proposto por Tukey (1977).

Outro método de vasta utilização na busca por dados anômalos é o *Z-Score*, em que a estatística do método é computada através das medidas de média (μ) e desvio padrão (σ), conforme Equação 6:

$$Z_i = \frac{X_i - \mu}{\sigma}, \text{ com } X_i \sim N(\mu, \sigma^2) \quad (6)$$

O método é baseado nas propriedades da distribuição normal, isto é, se $X_i \sim N(\mu, \sigma^2)$, então $Z_i \sim N(0,1)$ e, assim, toda observação com $|Z_i| > 3$ é considerada um *outlier.* Quando os dados seguem uma distribuição normal, este método apresenta um critério razoável (ALTMAN, 1968; SEO, 2006).

Uma problemática deste método, além de não ser eficiente para tratar dados não-normais, reside no fato da média e o desvio padrão serem medidas de localização pouco resistentes a valores extremos, ou seja, são medidas afetadas, de forma exagerada, por *outliers* (MORETTIN & BUSSAB, 2004).

Uma alternativa é utilizar o método *Z-Score modificado,* que utiliza estimadores mais robustos. Nesse método, a média é substituída pela mediana ($Q2$) e o desvio padrão pelo Desvio Absoluto da Mediana (MAD) (IGLEWICZ & HOAGLIN, 1993). A estatística do método é dada por:

$$M_i = \frac{0{,}6745 \cdot (x_i - Q2)}{MAD} \quad (7)$$

em que $Q2$ é a mediana da amostra, $MAD = mediana\{|x_i - Q2|\}$ e a constante 0,6745 é devido a $E\ (MAD) = 0{,}6745 \cdot \sigma$ para um grande conjunto de dados normais. Iglewicz & Hoaglin (1993) sugerem que, se $|M_i| > 3{,}5$, a observação i é um *outlier.*

4. PROPOSIÇÃO DO MÉTODO

O método proposto neste estudo fundamenta-se, prioritariamente, em teoremas da estatística clássica e Geoestatística. Toda a metodologia, incluindo a parte inovadora, foi implementada no *software* livre R (R Core Team, 2017) e o *script* pode ser consultado no Apêndice. Para a análise geoestatística, quando necessário, recorre-se ao pacote geoR, desenvolvido por Ribeiro Júnior & Diggle (2001).

A Figura 5 ilustra a metodologia proposta, chamada, neste trabalho, de ***AEDO (*Algoritmo Espacial para Detecção de *Outliers).*** A primeira etapa consiste na importação da nuvem de pontos (conjunto de dados espaciais). Nessa fase o algoritmo

desenvolvido é capaz de importar as coordenadas tridimensionais no formato XYZ (arquivo *Shapefile* ou arquivo de texto), em que X e Y representam, respectivamente, as coordenadas posicionais, sejam elas locais, projetadas ou geodésicas, e Z denota a profundidade reduzida. Nos casos em que o usuário optar por importar um arquivo de texto, é necessário informar o sistema de projeção adotado.

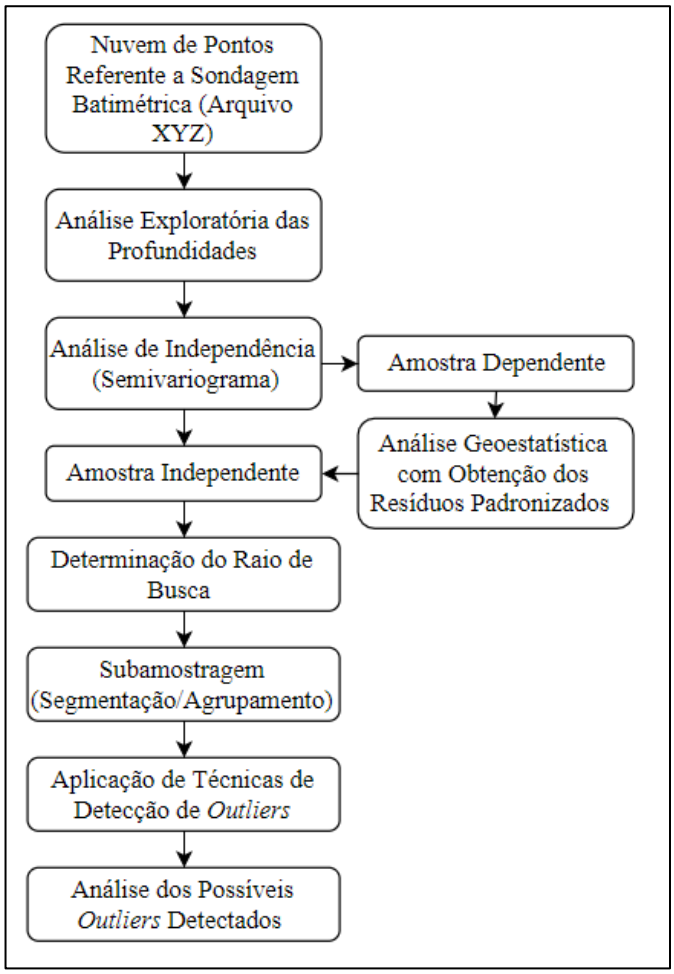

Figura 5 – Fluxograma da metodologia proposta para detecção de *spikes* em dados de batimetria coletados a partir de sistemas de sondagem por varrimento.

Seguidamente, realiza-se a análise exploratória dos dados de profundidade, fase indispensável em qualquer análise estatística e/ou geoestatística (FERREIRA et al., 2013). Basicamente, nessa etapa, o método propõe a construção e interpretação de gráficos (histogramas, *Q-Q Plot etc.*) e de estatísticas, como: média, desvio padrão, mínimo, máximo, coeficientes de assimetria e curtose, dentre outros.

A próxima etapa consiste na verificação da presença de independência espacial entre os dados de profundidade, condição assumida pelas técnicas de detecção de *outliers* utilizadas neste estudo (MORETTIN & BUSSAB, 2004; SEO, 2006). Para isso, devido principalmente à sua eficiência, sugere-se o uso do semivariograma,

ferramenta utilizada pela Geoestatística para avaliar a autocorrelação espacial dos dados (MATHERON, 1965; FERREIRA et al., 2015).

O semivariograma dos dados (Figura 6), doravante denominado de semivariograma experimental, é um gráfico construído através da função de semivariância *versus* cada valor h, em que h é a distância euclidiana entre as profundidades amostradas. Esse gráfico é também conhecido como variograma (BACHMAIER & BACKES, 2011).

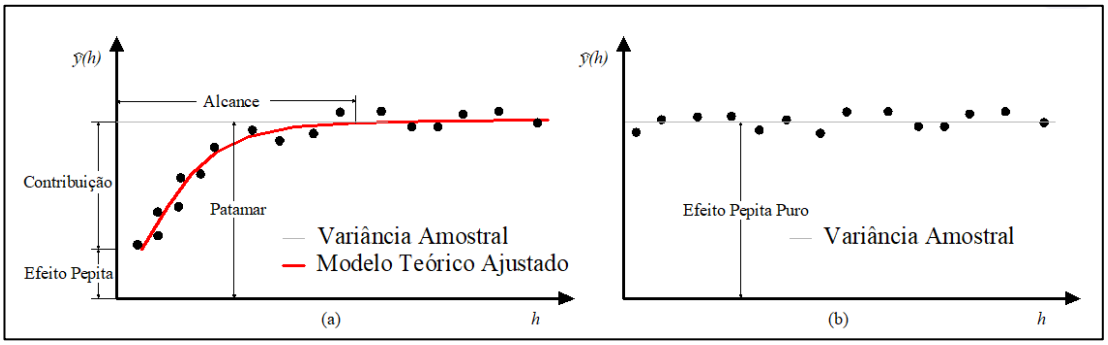

Figura 6 – Exemplo de semivariogramas para dados espacialmente dependentes (a) e espacialmente independentes (b).

Segundo Matheron (1965), a função de semivariância é definida como sendo a metade da esperança matemática do quadrado da diferença entre as realizações de duas variáveis localizadas no espaço, separada pela distância h.

Dentre os estimadores de semivariâncias, destaca-se o baseado no método dos momentos, dado pela Equação 8:

$$\hat{y}(h) = \frac{1}{2 \cdot N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \quad (8)$$

em que $\hat{y}(h)$ é o valor estimado da semivariância para a distância h e $N(h)$ é o número de pares dos valores $Z(x_i)$ e $Z(x_i + h)$, separados por uma distância h. É esperado que $\hat{y}(h)$ aumente com a distância h, até um valor máximo, teoricamente a variância amostral, no qual se estabiliza em um ***patamar*** correspondente à distância dentro da qual as amostras apresentam-se correlacionadas espacialmente, a essa distância dar-se o nome de ***alcance***. Conforme Equação 8, pode-se facilmente concluir que $\hat{y}(0) = 0$. Todavia, é comum para a maioria das variáveis estudadas, o semivariograma experimental apresentar uma descontinuidade para distâncias menores do que a menor distância amostral, assim, $\hat{y}(0) \neq 0$. Esse fenômeno é conhecido como ***efeito pepita***. A diferença entre o patamar e o efeito pepita é chamado de ***contribuição*** (Figura 6a)**.** Por fim, nos casos em que a profundidade não apresentar autocorrelação espacial, o semivariograma apresentará apenas o efeito pepita puro (Figura 6b).

Para construir o semivariograma experimental, deve-se definir uma distância de passo para que sejam selecionados os pares de profundidade, e uma distância limite para o crescimento dos passos. Segundo Santos (2015), deve-se escolher uma distância limite que melhor represente a dependência espacial da variável analisada. Geralmente, o valor máximo utilizável corresponde à metade da distância máxima entre os pontos (RIBEIRO JÚNIOR & DIGGLE, 2001; DIGGLE & RIBEIRO JÚNIOR, 2007).

Nesse sentido, o AEDO calcula a distância máxima entre as profundidades e, baseado nesta informação, constrói três semivariogramas, o primeiro com alcance igual a 75% da distância máxima; o segundo com 50% da distância máxima e o terceiro com 25%. Diante desses gráficos, o analista pode decidir sobre a existência, ou não, de dependência espacial entre os dados, isto é, se pelo menos um dos semivariogramas não apresentar efeito pepita puro, a autocorrelação espacial é confirmada. Nessa etapa, o algoritmo também é capaz de gerar o envelope de Monte Carlo (simulação de Monte Carlo) para confirmar, de forma explanatória, a existência de autocorrelação espacial (ISAAKS, 1990).

Se constatada a dependência espacial, sugere-se recorrer a Geoestatística para o correto tratamento estatístico dos dados. A escolha da geoestatística como metodologia de apoio baseia-se em suas características ideais. Conforme afirma Vieira (2000), a Geoestatística, além de considerar a estrutura de dependência espacial dos dados, é capaz de modelar e predizer sem tendência e com variância mínima, sendo, portanto, uma ferramenta de apoio muito eficiente para tratar dados geoespaciais.

O semivariograma é a ferramenta básica de suporte as técnicas geoestatísticas, sendo, por esse motivo, a etapa mais importante da análise. A inferência geoestatística baseia-se na pressuposição de três hipóteses de estacionariedade, a estacionariedade de primeira e segunda ordem e do semivariograma (MATHERON, 1965; FERREIRA et al., 2013). Porém, conforme afirma Vieira (2000), comumente assume-se apenas a hipótese intrínseca ou de estacionariedade do semivariograma, isto é, assume-se que o variograma existe e é estacionário para a variável na área de estudo.

Quando o semivariograma apresenta um comportamento idêntico para todas as direções, ele é dito isotrópico, caso contrário, ele é dito anisotrópico. Quando for detectada a anisotropia, esta deve ser corrigida, geralmente através de transformações lineares, pois a mesma impossibilita a existência da estacionariedade, condição

necessária para a exatidão na análise e estimativas para a área em estudo (ISAAKS, 1990; VIERA, 2000; FERREIRA et al., 2013; 2015).

Uma vez que o semivariograma experimental é obtido, pode-se então ajustá-lo através de modelos teóricos. Esse ajuste consiste na modelagem da dependência espacial, propriamente dita, sendo assim, deve ser feita com cautela. Incertezas nesse ajuste, conduzirão a incertezas de predição (FERREIRA et al., 2013). Com o modelo teórico ajustado, podem-se predizer valores em locais não-amostrados, considerando a variabilidade espacial dos dados (VIEIRA, 2000, SANTOS, 2015).

Diversos são os modelos isotrópicos existentes na literatura, esses contemplam semivariogramas com e sem patamar. Dentre os modelos sem patamar, cita-se o modelo de potência e dentre os com patamar (mais comuns), destacam-se o modelo exponencial, o modelo esférico e o modelo gaussiano (VIEIRA, 2000). Após a modelagem do semivariograma, pode-se predizer valores não amostrados, sem viés e com variância mínima, através do método de interpolação geoestatística denominada krigagem. Maiores detalhes sobre a modelagem geoestatística pode ser consultada, por exemplo, em Vieira (2000) e Ferreira et al. (2013).

Após a modelagem geoestatística, realiza-se o processo de validação cruzada (autovalidação *leave-one-out)* que, de acordo com Ferreira et al. (2013), é o procedimento que quantifica as incertezas inerentes ao processo de modelagem e predição, devidas às hipóteses assumidas ou, mais comumente, ao ajuste do modelo. Essa técnica consiste em, temporariamente, retirar um valor amostrado e predizer o valor do mesmo com o uso do modelo teórico ajustado aos demais valores amostrados. Ao final obtém-se os resíduos da modelagem, isto é, diferença entre os valores observados e seus correspondentes preditos (VIEIRA, 2000). A partir desses resíduos pode-se avaliar a qualidade da estimativa.

Segundo Santos et al. (2017), esses resíduos são conhecidos como ruído branco, ruído aleatório ou passeio aleatório e na sua forma padronizada, doravante denominados de RP (Resíduo Padronizado), possuem importantes características estatísticas a saber: seguem distribuição normal com média nula e variância unitária, são independentes, não-tendenciosos e homogêneos.

Verificada a independência espacial, seja das profundidades ou dos RPs, prossegue-se com a aplicação da metodologia proposta (Figura 5). Assim, o próximo passo consiste na segmentação da amostra que visa, prioritariamente, preservar a

característica espacial da análise (análise local). Essa subamostragem também permite uma redução considerável do tempo de processamento de máquina.

Como já exposto, as metodologias de detecção de *outliers* baseadas na estatística clássica pressupõe que as observações são variáveis aleatórias independentes e identicamente distribuídas (MORETTIN & BUSSAB, 2004; SEO, 2006). Assim, a etapa de subamostragem proposta neste estudo fundamenta-se no seguinte teorema: Se $X_1, \cdots, X_k$ são variáveis aleatórias independentes e $g_1(\cdot), \cdots, g_s(\cdot)$ são s funções tal que $Y_j = g_j(X_j), j = 1, \cdots, k$ são variáveis aleatórias, então, $Y_1, \cdots, Y_s$ são independentes. A demonstração desse teorema, bem como exemplos teóricos, pode ser consultada em Mood (1913) e Mood et al. (1974).

O AEDO aplica uma segmentação intitulada, neste estudo, de ***Segmentação em Círculos*** (Figura 7). Diante disso, o algoritmo gera um círculo centrado em cada profundidade ou RP, identificando e armazenando todos os dados presentes dentro do círculo em subamostras. Toda a análise, a partir desse momento, é então realizada apenas nessas subamostras.

O raio do círculo ou ***Raio de Busca*** poderá ser definido pelo usuário ou baseado na análise espacial. Destaca-se que essa grandeza é intimamente ligada à morfologia de fundo. Como o relevo submerso não é visível, a determinação deste raio pelo analista torna-se bastante subjetiva. Por hora, é sabido apenas que naqueles locais onde for nítida a presença de um relevo plano, pode-se adotar círculos com maiores dimensões.

Alternativamente, sugere-se que o raio seja equivalente ao triplo da distância mínima. Nesse caso, o algoritmo é capaz de computar a menor distância entre os pontos e atribuir três vezes esse valor ao raio do círculo. Tal sugestão, *a priori,* não possui fundamentação teórica, porém elimina a intervenção do analista, automatizando o processo. É baseada no pressuposto de que a nuvem de pontos, adquirida a partir de um sistema de sondagem por faixa, é densa e sem feriados[5]. Assim, este raio é capaz de garantir uma investigação local, com subamostras contendo pontos suficientes para a análise.

A Figura 7 a seguir ilustra o procedimento.

[5] Termo utilizado entre a comunidade hidrográfica para descrever falhas de cobertura nas áreas sondadas.

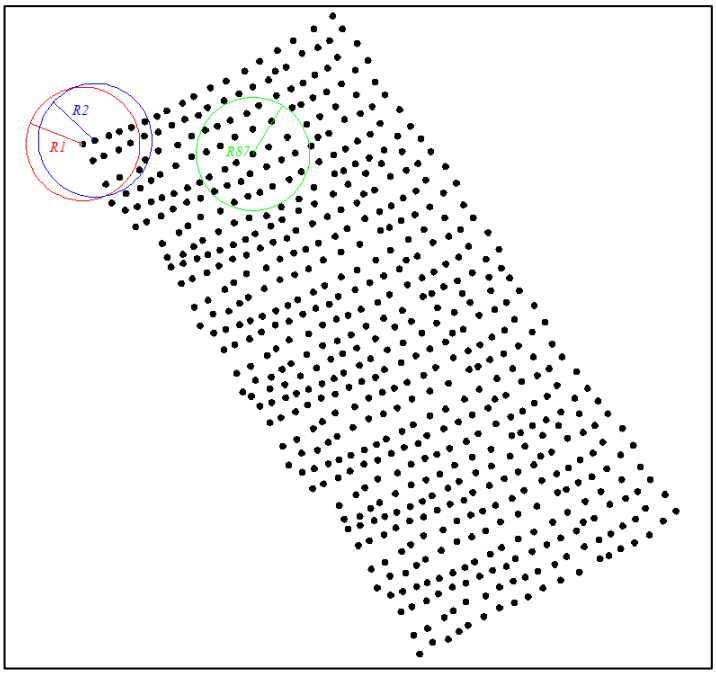

Figura 7 – Exemplo da técnica de Segmentação em Círculos empregada pelo AEDO.

A Figura 7 apresenta uma nuvem de pontos contendo dados de batimetria coletados por um sistema de sondagem por faixa. O raio de busca foi definido pelo usuário. O círculo gerado para o ponto 1, em vermelho, captou 11 pontos, enquanto o círculo gerado para o ponto 2, em azul, captou exatos 15 pontos. Note que, desses 15 pontos, 11 coincidem com aqueles captados pelo círculo anterior. Sugerindo que, no AEDO, as profundidades ou RPs serão analisadas mais de uma vez, dependendo, logicamente, do raio de busca e da densidade da nuvem de pontos. Esse fato traz ganhos à metodologia proposta, e por isso, o algoritmo armazena essas informações, objetivando utilizá-las posteriormente.

Em verde, tem-se o círculo gerado para o ponto 87, contendo em seu interior 29 pontos. É importante garantir que as subamostras possuam pontos suficientes para uma análise estatística coerente. Sendo assim, o algoritmo verifica a quantidade de pontos presentes em cada subamostra. Caso o número seja inferior a 7 (embasamento empírico), a subamostra é desconsiderada nas análises posteriores.

De posse das subamostras, o AEDO aplica três técnicas de detecção de *outliers*: o *Boxplot Ajustado* (VANDERVIEREN & HUBERT, 2004); o *Z-Score Modificado* (IGLEWICZ & HOAGLIN, 1993) e o ***Método δ***, proposto neste trabalho. O *Método* δ foi inspirado, em partes, na técnica proposta por Lu et al. (2010), que consiste em aplicar limiares de detecção de *spikes* baseados na variância amostral global e local, onde a variância local, refere-se a variância de subamostras. Nesse método, se a

variância global for maior que a variância local, o valor de corte é configurado para ser igual a variância global, caso contrário, o limiar é definido como $0{,}5 \cdot (\sigma^2_{Global} + \sigma^2_{Local})$, em que σ^2 é a variância. E assim, qualquer observação que possua resíduo maior, em valor absoluto, que o valor de corte, é considerado um *spike.* Porém, como já discutido, o desvio padrão e, portanto, a variância amostral do conjunto de dados, são medidas de dispersão pouco resistentes a *outliers*.

Por outro lado, a teoria dos erros afirma que, quando uma distribuição normal pode ser assumida, 68,3% dos dados avaliados estão no intervalo $\mu \pm \sigma$; 95% estão no intervalo $\mu \pm 1{,}96\sigma$ e 99,7% dos dados avaliados estão no intervalo $\mu \pm 3\sigma$ (MOOD, 1913; MOOD et al., 1974). Fundamentando-se nessas conjecturas, é muito comum, principalmente nas ciências geodésicas, efetuar a eliminação de *outliers* através da aplicação do limiar $3 \cdot \sigma$ ou $3 \cdot RMSE$, em que $RMSE$ é a raiz do erro quadrático médio ou, do Inglês, *Root Mean Square Error* (MIKHAIL & ACKERMAN, 1976; COOPER, 1987; HÖHLE & HÖHLE, 2009).

Diante disso, o Método δ é uma proposição que consiste num novo limiar de detecção de *outliers,* baseado em estimadores robustos, dado pela Equação 9:

$$Limiar = Q2_{Local} \pm c \cdot \delta \tag{9}$$

em que $Q2_{Local}$ é a mediana dos dados subamostrados e c e δ são constantes que dependem da variabilidade dos dados. A constante c assume o valor 1 para relevos irregulares ou canais artificiais (variabilidade alta); 2 para relevos ondulados (variabilidade média) e 3 para relevos planos (variabilidade baixa). Este valor pode ser entendido como um ponderador da constante δ e deve ser inserido pelo usuário. A constante δ é determinada automaticamente pelo algoritmo através da avaliação do Desvio Absoluto da Mediana Normalizado Global ($NMAD_{Global}$) ou Local ($NMAD_{local}$), isto é, $\delta = 0{,}5 \cdot (NMAD_{Global} + NMAD_{local})$, se $NMAD_{Global} > NMAD_{local}$, ou, em caso contrário, $\delta = NMAD_{Global}$.

O $NMAD$ corresponde a $1{,}4826 \cdot mediana\{|x_i - Q2|\}$, ou seja, $1{,}4826 \cdot MAD$. É considerado uma estimativa para a dispersão dos dados mais resistente a *outliers* que o tradicional desvio padrão. Nos casos em que a distribuição normal for verificada, o $NMAD$ é equivalente ao desvio padrão (HOAGLIN et al., 1983; HÖHLE & HÖHLE, 2009). Contudo deve-se destacar que a constante $1{,}4826$ remete a dados normalmente distribuídos, isto é, o desvio absoluto da mediana de uma distribuição normal padrão é igual a $1/0{,}6745 \approx 1{,}4826$, ou ainda, $(\Phi^{-1}(3/4)) \approx 1{,}4826$, em

que Φ^{-1} é a função inversa de distribuição acumulada da distribuição normal padrão (ROUSSEEUW & CROUX, 1993).

Diante do exposto, na hipótese do AEDO empregar os limiares definidos pelo *Método* δ, bem como pelo *Z-Score Modificado*, indiretamente se pressupõe que as subamostras possuam uma distribuição normal. Isto é devido a dois fatores principais. O primeiro consiste na inviabilidade de se efetuar testes de hipóteses para determinar a distribuição de probabilidade de cada subamostra e o segundo fato, inclusive utilizado para justificar o primeiro, reside no fato da distribuição normal ser a mais importante distribuição de probabilidade contínua e, por este motivo, utilizada na maioria das técnicas de estatística aplicada (MOOD, 1913; MOOD et al., 1974).

Aplicado os limiares de detecção de *outliers,* no próximo passo, o método proposto determina a probabilidade do dado ser um *outlier* ($P_{outlier}$) em cada uma das três técnicas, baseando-se no número de vezes que o dado foi analisado ($N^{\underline{o}}{}_{analisado}$) e o número de vezes que ele foi considerado um *outliers* ($N^{\underline{o}}{}_{outlier}$) (Equação 10):

$$P_{outlier}(\%) = \frac{N^{o}{}_{outlier}}{N^{o}{}_{analisado}} \cdot 100 \qquad (10)$$

Por exemplo, considere que, dado um raio de busca qualquer, a observação i foi subamostrada 20 vezes (Figura 7). Sendo assim, ela foi analisada pelas três técnicas de detecção de *outliers* nessas 20 vezes. Considere, ainda, que das 20 vezes, em 10 vezes a observação i foi considerada um *outlier* pelo *Método* δ, logo, $P_{outlier} = (10/20) = 0{,}5$, ou seja, a observação i tem 50% de probabilidade de ser um *outlier* se o limite de corte considerado for aquele dado pelo *Método* δ.

Por fim, definido um P_{limiar} padrão pelo usuário, o AEDO plota espacialmente todas as observações, destacando os *spikes* detectados pelos limiares utilizados, isto é, todos os *outliers* com $P_{outlier} \geq P_{limiar}$. O usuário, então, efetua uma inspeção visual com objetivo de confirmar os *spikes* e, posteriormente, elimina-los. Em todos os casos, novos arquivos XYZ, para cada técnica, são criados, isto é, o AEDO associado, respectivamente, ao *Boxplot Ajustado, Z-Score Modificado* e *Método* δ.

Essa última etapa requer um cuidado adicional no sentido de que, se houver qualquer dúvida sobre o possível *spike*, deve-se refinar as análises e, dependendo da finalidade do levantamento, retornar à área de sondagem para execução de pesquisa de perigo. É muito comum, dependendo da densidade de sondagem, o analista confundir feições marinhas ou mesmo objetos afundados com *spikes* e assim, erroneamente tratá-los como tal.

5. EXPERIMENTOS E RESULTADOS

5.1. Aplicação do método proposto em dados simulados

Objetivando avaliar a robustez do AEDO, bem como efetuar ajustes, em um primeiro momento, recorreu-se a simulação computacional. Uma área de estudo semelhante a um canal de navegação foi construída através de dados simulados, conforme ilustrado na Figura 8.

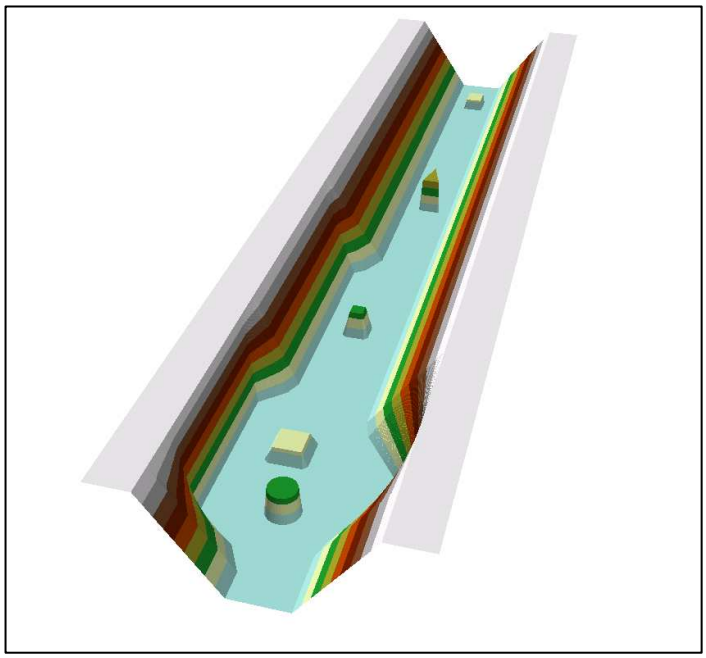

Figura 8 – Superfície batimétrica tridimensional construída através de simulação computacional.

A superfície batimétrica simulada possui uma área de 1.600m² (80 x 20m) com relevo submerso variando de 8 a 15 metros, dois taludes laterais com inclinações médias de 135% e cinco estruturas submarinas, posicionadas no leito do canal, que reproduzem perigos à navegação, tais como: bancos de areia, rochas e cascos soçobrados. Essas feições são representadas por sólidos geométricos conhecidos (paralelepípedo, tronco de cone, tronco de pirâmide, *etc.*) e possuem alturas que variam entre 1 e 3 metros. A partir desta superfície extraiu-se o conjunto de dados, composto por 40.000 pontos batimétricos, inicialmente sem *outliers*, espaçados de 20 em 20 centímetros, ou seja, 25 pontos/m².

A Tabela 1 resume as informações.

Tabela 1 – Estatística descritiva da área de estudo simulada.

Número de Observações	40.000
Profundidade Média (m)	11,607
Profundidade Mínima (m)	8,000
Profundidade Máxima (m)	15,000
Variância (m²)	8,6130
Coeficiente de Curtose	1,300
Coeficiente de Assimetria	-0,050

Como tratam-se de dados simulados, as fases de análise exploratória e de independência amostral não são detalhadas, todavia, evidencia-se que o conjunto de dados é espacialmente independente. Assim sendo, o método proposto foi aplicado diretamente sobre as profundidades. O raio de busca, conforme exposto na seção 4, foi definido como o triplo da distância mínima entre os pontos, isto é, 60 centímetros. A partir deste raio, a amostra original foi segmentada em 40.000 subamostras e os limiares de detecção de *outliers* foram aplicados. Na primeira etapa, o *Boxplot Ajustado*, *Z-Score Modificado* e *Método* δ, localizaram, respectivamente, 3.476 (8,69%), 7.742 (19,35%) e 453 (1,13%) possíveis *spikes*. Como trata-se de um canal de navegação, a constante c do *Método* δ foi configurada com o valor unitário.

Com o objetivo de avaliar a concordância entre os métodos, foi realizada uma análise comparativa dos possíveis *spikes* localizados por cada limiar, concluindo-se que, dos 453 pontos detectados pelo *Método* δ, 391 foram também detectados pelo *Z-Score Modificado* e 182 pelo *Boxplot Ajustado*. Isto é, tomando o menor conjunto de dados como referência, têm-se uma concordância de, respectivamente, 86,31% e 40,18%. A concordância entre o *Boxplot Ajustado* e o *Z-Score Modificado* foi de 98,36%, ou seja, dos 3.476 possíveis *spikes* localizados pelo limiar *Boxplot Ajustado*, cerca de 3.419 foram também detectados pelo limiar *Z-Score Modificado.*

Na próxima etapa, a análise foi refinada através do cálculo da probabilidade de o dado ser um *spike* para cada uma das três técnicas, baseando-se no número de vezes que a profundidade foi analisada e o número de vezes que ela foi considerada um *outlier*. Para a execução desta etapa, *a priori,* sugere-se um $P_{limiar} = 50\%$, isto é, se $P_{outlier} \geq 0{,}5$, a profundidade analisada é considerada um *spike.* A Tabela 2 ilustra essa etapa, realizada para o *Método* δ.

Tabela 2 – Análise da probabilidade da profundidade *i* ser um *spike* ($P_{limiar} = 0{,}5$).

Ponto	$\boldsymbol{Nº_{analisado}}$	$\boldsymbol{Nº_{outlier}}$	$\boldsymbol{P_{outlier}\ (\%)}$	***Método δ*** $\boldsymbol{(P_{outlier}(\%) \geq 50)}$
1	6	6	100,00	Sim
2	15	5	33,33	Não
⋮	⋮	⋮	⋮	⋮
n	27	20	74,07	Sim

Após essa etapa, os limiares do *Boxplot Ajustado* e *Método* δ, como esperado, não localizaram nenhum *spike*, ou seja, uma concordância de 100%. Em contrapartida, o método *Z-Score Modificado*, de modo equivocado, sinalizou 287 (0,72%) pontos como possíveis *outliers*, dentre estes, merece destaque as profundidades das estruturas submersas, conforme é ilustrado na Figura 9a, onde a malha de pontos batimétricos é plotada na cor azul e os *spikes* destacados em vermelho. No caso de um processamento de dados reais, a eliminação dessas sondagens representativas de perigos à navegação, poderiam causar problemas graves, como encalhes de navios e embarcações, avarias no casco e até mesmo um naufrágio.

Analisando os metadados dos *outliers* detectados percebeu-se a necessidade de efetuar um ajuste no P_{limiar} desse valor de corte em específico. Após alguns testes e simulações, chegou-se a um valor ótimo de 80%, isto é, $P_{limiar} = 0,8$ (Figura 9b). Assim, baseando-se nos dados simulados, recomenda-se um $P_{limiar} = 0{,}5$ para o *Boxplot Ajustado* e *Método* δ e $P_{limiar} = 0,8$ para o *Z-Score Modificado*.

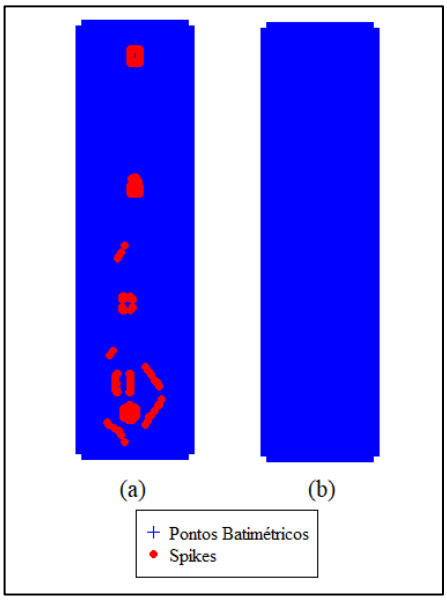

Figura 9 – *Spikes* detectados pelo AEDO a partir dos limiares do *Z-Score Modificado*, para um $P_{limiar} = 0{,}5$ (a) e um $P_{limiar} = 0{,}8$.

Continuando as simulações, foram introduzidos dez *spikes* ao conjunto de dados, com posições horizontais escolhidas aleatoriamente. A magnitude destes *spikes* foram determinadas com base em conhecimento prático adquirido a partir do processamento de dados reais (Tabela 3). O número reduzido de *spikes* permitirá, posteriormente, a realização de uma análise mais criteriosa.

Tabela 3 – *Spikes* inseridos aleatoriamente ao conjunto de dados.

ID do Ponto	$\boldsymbol{X}$ $\boldsymbol{(m)}$	$\boldsymbol{Y}$ $\boldsymbol{(m)}$	$\boldsymbol{Z}$ $\boldsymbol{(m)}$	$\boldsymbol{Z_{spike}}$ $\boldsymbol{(m)}$	**Magnitude do *Spike* (m)**
2751	599.574,540	7.700.293,157	9,00	10,50	1,50
2659	599.576,140	7.700.293,357	8,00	10,00	2,00
15051	599.574,540	7.700.268,557	11,00	15,00	4,00
15260	599.576,340	7.700.268,157	8,00	5,00	3,00
17573	599.578,940	7.700.263,557	11,50	12,10	0,60
19611	599.566,540	7.700.259,357	15,00	17,00	2,00
23557	599.575,740	7.700.251,557	8,00	3,20	4,80
25943	599.572,940	7.700.246,757	10,00	5,00	5,00
35141	599.572,540	7.700.228,357	8,00	11,00	3,00
38738	599.580,940	7.700.221,157	14,30	10,30	4,00

Aplicando a metodologia proposta de modo análogo ao anterior, num primeiro instante os limiares *Boxplot Ajustado*, *Z-Score Modificado* e *Método* δ, localizaram, respectivamente, 3.458 (8,65%), 7.749 (19,37%) e 458 (1,15%) possíveis *spikes*. Das 458 profundidades espúrias determinadas, *a priori,* pelo Método δ, 187 foram também localizadas pelo *Boxplot Ajustado* (40,83%) e 396 pelo *Z-Score Modificado* (86,46%). Enquanto que a concordância entre o *Boxplot Ajustado* e o *Z-Score Modificado* foi de, aproximadamente, 98%.

Posteriormente foram definidos o $P_{limiar} = 0{,}5$ para o *Boxplot Ajustado* e *Método* δ e $P_{limiar} = 0,8$ para o *Z-Score Modificado.* Os resultados são sumarizados na Tabela 4.

Tabela 4 – Resultado do processamento dos dados da área de estudo simulada.

ID do Ponto	**Magnitude do *Spike* (m)**	***Boxplot Ajustado*** $(P_{outlier}(\%) \geq 50)$	***Z-Score Modificado*** $(P_{outlier}(\%) \geq 80)$	***Método δ*** $(P_{outlier}(\%) \geq 50)$
2751	1,50	Detectado	Detectado	Não detectado
2659	2,00	Detectado	Detectado	Não detectado
15051	4,00	Detectado	Detectado	Detectado
15260	3,00	Detectado	Detectado	Detectado
17573	0,60	Não detectado	Não detectado	Não detectado
19611	2,00	Detectado	Detectado	Não detectado
23557	4,80	Detectado	Detectado	Detectado
25943	5,00	Não detectado	Detectado	Detectado
35141	3,00	Detectado	Detectado	Detectado
38738	4,00	Detectado	Detectado	Detectado

O limiar *Boxplot Ajustado* detectou todos os *spikes* implantados, exceto os pontos de ID 17573 e 25943, cuja magnitude do erro é de, respectivamente, 0,60 e 5 metros. Sendo assim, a porcentagem de acerto foi de 80%, isto é, dos 10 *spikes* implantados, o limiar *Boxplot Ajustado* localizou 8.

O ponto de ID 17573 não foi detectado devido, especificadamente, a sua baixa magnitude para o relevo analisado. Todavia, das 29 vezes que este ponto foi analisado, ele apresentou-se como um *spike* em 11 ocasiões, ou seja, uma probabilidade de 38%. Por outro lado, é nítido que a falha na detecção do ponto de ID 25943 não está relacionada com a magnitude do erro ou com o limiar aplicado, uma vez que *spikes* de magnitudes inferiores foram localizados. Assim, tal fato pode estar intimamente ligado com a vizinhança do *outlier* analisado. O ponto 25943 encontra-se posicionado, horizontalmente, sobre uma estrutura submersa, próximo a borda, isto é, na crista do talude. Todavia, como pode ser visto na Tabela 5, este ponto teve um $P_{outlier}(\%) = 48\%$, muito próximo do P_{limiar} adotado. Todos os demais pontos, considerados *outliers* no primeiro passo do método AEDO, obtiverem $P_{outlier}$ menor que 28%, na média, menor que 8%.

O *Z-Score Modificado* obteve uma porcentagem de acerto de 90%, isto é, ele foi capaz de detectar todos os *spikes*, exceto o ponto de ID 17573, que possui, conforme supracitado, um erro com magnitude muito inferior daqueles experimentados na prática hidrográfica. Esse ponto apresentou um $P_{outlier}(\%) =$

21%. Dos demais *spikes* implantados, 8 deles atingiram $P_{outlier} = 100\%$, o que mostra a eficiência desse limiar. Em contrapartida, aproximadamente 130 pontos, obtiveram um $P_{outlier}$ variando entre 60% e 79%, muitos deles, representativos de estruturas submersas, sugerindo maiores cuidados durantes as análises posteriores.

Por fim, o *Método* δ detectou 60% dos *spikes* implantados, todos apresentando um $P_{outlier} = 100\%$. Os pontos de ID 2751, 2659, 17573 e 19611 obtiveram um $P_{outlier} = 0\%$ e, assim, não foram detectados (Tabela 4). Analisando a Tabela 3, facilmente nota-se que a falha na localização desses pontos está relacionada com a magnitude dos erros, ou seja, o *Método* δ mostrou-se capaz de detectar, para o relevo em questão, apenas os *spikes* com magnitude maior que 2 metros. Nessa análise deve-se ter ciência que o limiar em questão se baseia em um estimador robusto de variabilidade dos dados, o que pode ser pouco eficaz para dados demasiadamente regulares, como o conjunto analisado, que possuem diversas faixas de dados exatamente planas, isto é, com mesma profundidade e consequente $NMAD = 0$.

Todos os demais possíveis *spikes* tiveram uma probabilidade menor que 28%, exceto 6 pontos, que tiveram um $P_{outlier}$ variando entre 35% e 48%. Esses pontos representam a crista da estrutura submersa de formato plano triangular, com altura de 1 metro (Figura 8).

A Tabela 5 e a Figura 10 resumem e ilustram, respectivamente, as informações discutidas.

Tabela 5 – $P_{outlier}(\%)$ dos dados da área de estudo.

ID do Ponto	***Boxplot Ajustado***	***Z-Score Modificado***	***Método δ***
	$\boldsymbol{P_{outlier}}$	$\boldsymbol{P_{outlier}}$	$\boldsymbol{P_{outlier}}$
2751	59%	100%	0%
2659	59%	100%	0%
15051	86%	97%	100%
15260	83%	100%	100%
17573	38%	21%	0%
19611	76%	100%	0%
23557	79%	100%	100%
25943	48%	100%	100%
35141	97%	100%	100%
38738	100%	100%	100%

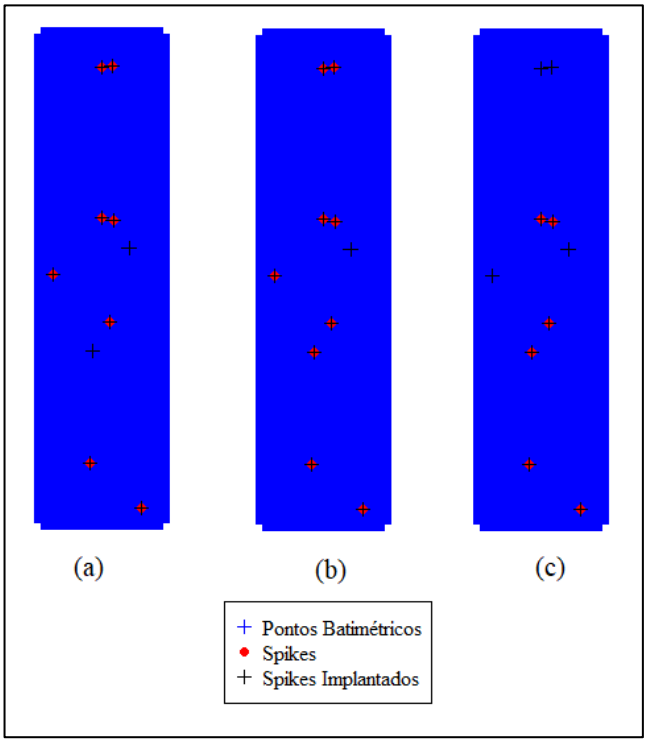

Figura 10 – *Spikes* implantados e *Spikes* detectados pelo AEDO a partir dos limiares das técnicas *Boxplot Ajustado* (a), *Z-Score Modificado* (b) e Método δ (c).

No geral, a metodologia proposta, isto é, o AEDO apresentou-se bastante eficiente e versátil. Embora o Método δ tenha apresentado apenas 60% de acerto, ele foi capaz de detectar todos os *spikes* com magnitude superior a 2 metros e, talvez o ponto mais importante, nenhuma estrutura submersa pertencente ao relevo do canal foi sinalizada como *spike*, preservando, nesses casos, a segurança na navegação. De modo semelhante, os demais limiares utilizados pela metodologia proposta também obtiveram resultados ótimos.

Destaca-se que o algoritmo implementado realizou todo o processamento desse conjunto de dados em, aproximadamente, 2 horas e 45 minutos. Foi utilizada uma máquina com sistema operacional Windows 10, memória RAM de 8GB (parcialmente dedicada ao *software R*) e processador Intel® Core™ i7-4500U CPU @ 1,80GHz 2,40 GHZ.

5.2. Aplicação do método proposto em dados reais

Os dados reais que serviram de base para a aplicação da metodologia proposta foram obtidos a partir de uma parceria com a empresa *A2 Marine Solution*. O levantamento hidrográfico foi realizado em abril de 2017, nas proximidades da bacia

de evolução do Terminal Integrador Portuário Luiz Antônio Mesquita (TIPLAM). Bacia de evolução consiste na área fronteiriça às instalações de acostagem, reservada para as evoluções necessárias às operações de atracação e desatracação dos navios.

O TIPLAM localiza-se na cidade de Santos, estado de São Paulo. Atualmente este terminal movimenta cerca de 2,5 milhões de toneladas por ano, sendo responsável pela descarga, principalmente, de enxofre, rocha fosfática, fertilizantes e amônia.

Para a coleta dos dados batimétricos foi utilizado um sistema de sondagem por faixa, composto pelo ecobatímetro multifeixe modelo *Sonic 2022* da marca *R2* Sonic, integrado com o sistema inercial, modelo *I2NS (Integrated Inertial Navigation System)* da marca *Applanix*.

Visto que os dados ainda se encontram em avaliação junto à DHN, maiores informações foram omitidas e apenas parte dos dados coletados foram utilizados. Sabe-se que o planejamento, execução, bem como a análise dos dados seguiram as recomendações da NORMAM-25 e S-44 para a categoria A e Ordem Especial, respectivamente.

As sondagens foram primeiramente submetidas a um pré-processamento no *software Hysweep* (Hypack, 2012), que consistiu nas seguintes etapas:

- Conversão dos dados coletados pelos diversos sensores para o formato do *Hysweep*;
- Análise dos dados dos sensores auxiliares (atitude, latência, velocidade do som, maré, *etc*.), objetivando a identificação de possíveis falhas. Se necessário, interpolação ou rejeição de dados anômalos;
- Junção dos datagramas[6];
- Cálculo da *Total Propagated Uncertainty* (Horizontal e Vertical);
- Cálculo das coordenadas tridimensionais no formato XYZ (profundidades reduzidas georreferenciadas), e
- Filtragem de pontos duplicados.

Nesta última etapa, foram eliminados todos os pontos duplicados, advindos, principalmente, das faixas sobrepostas. Essa diminuição da amostra original, permitiu uma redução considerável do tempo de processamento. Posteriormente, a área de

[6] Entidade de dados completa e independente. Neste caso, refere-se aos dados gerados pelos diversos sensores.

estudo foi selecionada e as coordenadas tridimensionais no formato XYZ (profundidades reduzidas georreferenciadas) foram exportadas.

A Figura 11 ilustra a área total de sondagem e a área de estudo, que corresponde a cerca de 11% da área total levantada.

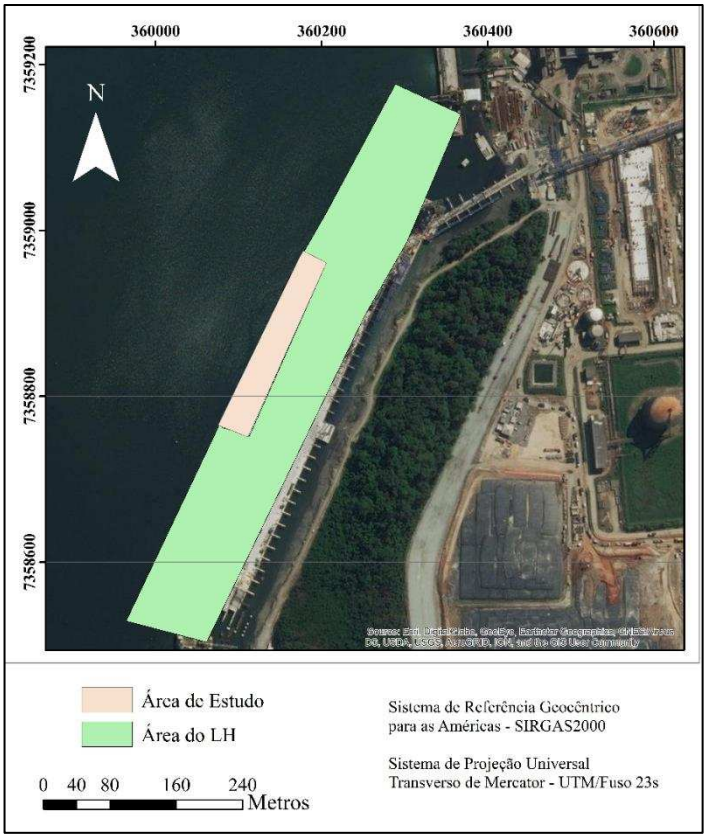

Figura 11 – Área de execução do Levantamento Hidrográfico (LH) e área de estudo.

De posse do arquivo XYZ, pôde-se dar continuidade a aplicação do método AEDO. Assim, a nuvem de pontos foi importada pelo algoritmo desenvolvido, onde os dados foram, então, submetidos a uma análise exploratória (Tabela 6).

Tabela 6 – Estatística descritiva da área de estudo.

Número de Observações	8090
Profundidade Média (m)	14,854
Profundidade Mínima (m)	9,220
Profundidade Máxima (m)	15,690
Variância (m²)	0,1240
Coeficiente de Curtose	101,660
Coeficiente de Assimetria	-6,900

Percebe-se que os dados apresentam uma variabilidade baixa, considerando o valor da variância (WARRICK & NIELSEN, 1980). Os coeficientes de assimetria e curtose que quantificam, respectivamente, o desvio da distribuição das profundidades em relação a uma distribuição simétrica e o grau de achatamento da distribuição, indicam uma distribuição assimétrica à esquerda (negativa), leptocúrtica e, *a priori,* com uma grande concentração de valores em torno da média. Em suma, conclui-se que a distribuição é não normal e/ou está eivada de valores anormais.

A seguir são apresentados alguns gráficos que auxiliam na análise exploratória e, assim sendo, são construídos e gerados pelo AEDO.

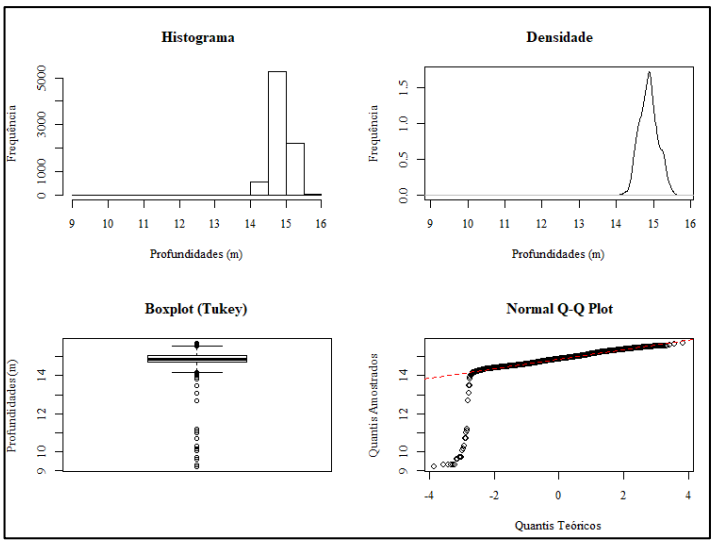

Figura 12 – Análise gráfica exploratória.

O histograma das profundidades conjuntamente com o gráfico da curva de densidade são ferramentas capazes de fornecer indicativos acerca da normalidade dos dados, que pode, ou não, ser comprovada através da análise do *Q-Q Plot*. Este consiste em um gráfico que permite checar a adequação da distribuição de frequência dos dados (empírica/real) à uma distribuição de probabilidades qualquer, resumidamente, os quantis da função de distribuição empírica são plotados contra os quantis teóricos da distribuição de probabilidades, nesse caso, a distribuição normal. Se a distribuição empírica é normal, o gráfico será apresentado como uma linha reta (HÖHLE & HÖHLE, 2009). Após a análise gráfica pôde-se constatar a não normalidade dos dados que poderá posteriormente, caso seja constatada a independência espacial, ser confirmada por testes de normalidade univariada.

Por fim, apresenta-se o *boxplot* de *Tukey,* no qual se nota, no geral, a presença de possíveis *outliers*. Contudo, conforme já discutido, tal afirmação não pode ser confirmada uma vez que o método de *Tukey* não considera a estrutura de dependência

espacial das profundidades, além do mais, os limiares de corte são derivados da distribuição normal. Todavia, observado o histograma, pode-se notar algumas profundidades ligeiramente afastadas da média.

A Figura 13 ilustra a nuvem de pontos, em que se percebe, na cor azul escuro, a presença de algumas profundidades destoando de seus valores vizinhos. Salienta-se que alguns dos possíveis *spikes* coincidem com aquelas profundidades anormais detectadas pelo método de *Tukey.*

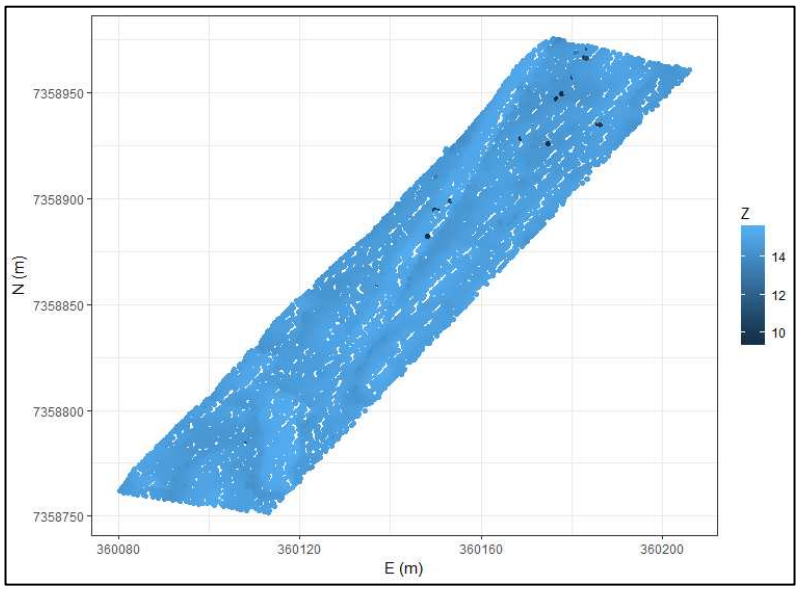

Figura 13 – Nuvem de pontos da área em estudo.

Após a análise exploratória, efetua-se a análise de independência espacial. Caso confirmada, dar-se continuidade a metodologia proposta determinando o raio de busca e, a seguir, segmentando a amostra. Caso contrário, uma análise geoestatística, com a obtenção dos resíduos padronizados (RPs), deve ser realizada. Então, toda a análise posterior é executada sobre os RPs. Conforme exposto na seção 4, o AEDO constrói três semivariogramas que permitem, através de uma análise visual, confirmar, ou não, a independência espacial.

A seguir são apresentados os semivariogramas, em que a independência espacial pode ser confirmada pela presença de efeito pepita puro (Figura 14).

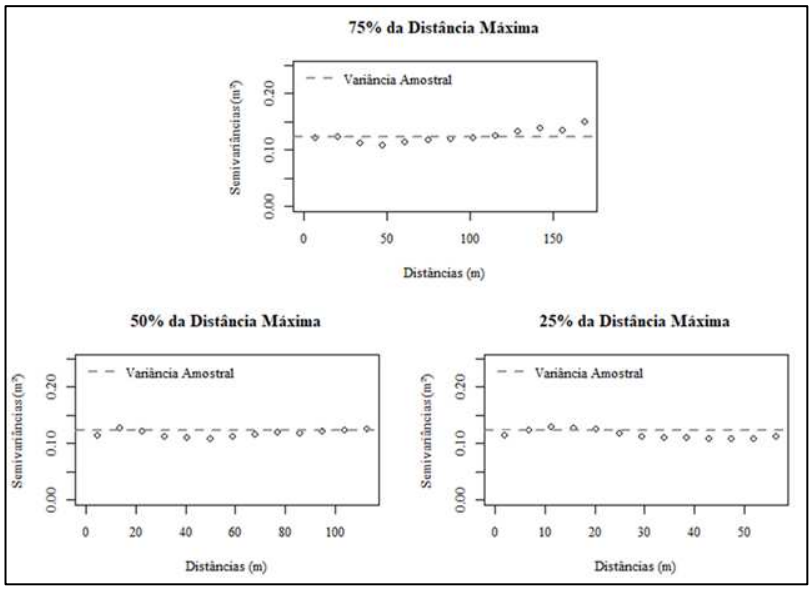

Figura 14 – Semivariogramas experimentais com 75%, 50% e 25% da distância máxima.

A determinação do raio de busca, conforme seção 4, é, preferencialmente, realizada a partir de uma análise espacial. Nesse caso, seleciona-se um raio igual a três vezes a distância mínima. No presente estudo obteve-se um raio igual a 1,799 metros. A partir desse raio o algoritmo segmenta e aplica, para cada subamostra, os três limiares de detecção de *outliers* propostos. Num primeiro momento, os limiares do *Boxplot Ajustado*, *Z-Score Modificado* e *Método* δ, detectaram, respectivamente, 2.028 (25,07%), 1.204 (14,88%) e 161 (1,99%) possíveis *spikes*. Como trata-se de uma área com relevo plano, a constante c do *Método* δ foi configurada com o valor 3.

Efetuando uma comparação entre os limiares empregados pelo AEDO na área analisada, nota-se uma concordância de 95,65% entre o *Método* δ e o limiar *Z-Score Modificado*, isto é, dos 161 possíveis *spikes* detectados pelo *Método* δ, 154 foram detectados também pelo *Z-Score Modificado*. De modo análogo, a concordância entre o *Método* δ e o *Boxplot Ajustado* foi de 57,76% e entre o *Z-Score Modificado* e o *Boxplot Ajustado* de 43,35%.

Na próxima etapa, o AEDO efetua um refinamento da análise calculando a probabilidade de o dado ser um *spike* para cada uma das três técnicas. Para a execução dessa etapa, foi utilizado um $P_{limiar} = 0{,}5$ para o *Boxplot Ajustado* e *Método* δ e $P_{limiar} = 0,8$ para o *Z-Score Modificado*, conforme sugerido na seção 5.1. Feito isso, a metodologia proposta associada as técnicas *Boxplot Ajustado*, *Z-Score Modificado* e *Método* δ, detectaram, respectivamente, 66 (0,82%), 24 (0,30%) e 34 (0,42%) *spikes*. Dos 24 *spikes* detectados pelo *Z-Score Modificado*, 16 foram detectados pelo *Boxplot*

Ajustado e pelo *Método δ*, uma concordância de cerca de 66%. Já a concordância entre o *Método δ* e o *Boxplot Ajustado* foi de, aproximadamente, 38,23%, sugerindo que os 16 *outliers* localizados em concordância com o *Z-Score Modificado* não são os mesmos.

A Figura 15 ilustra a área de estudo, com destaque em vermelho, para os *spikes* detectados.

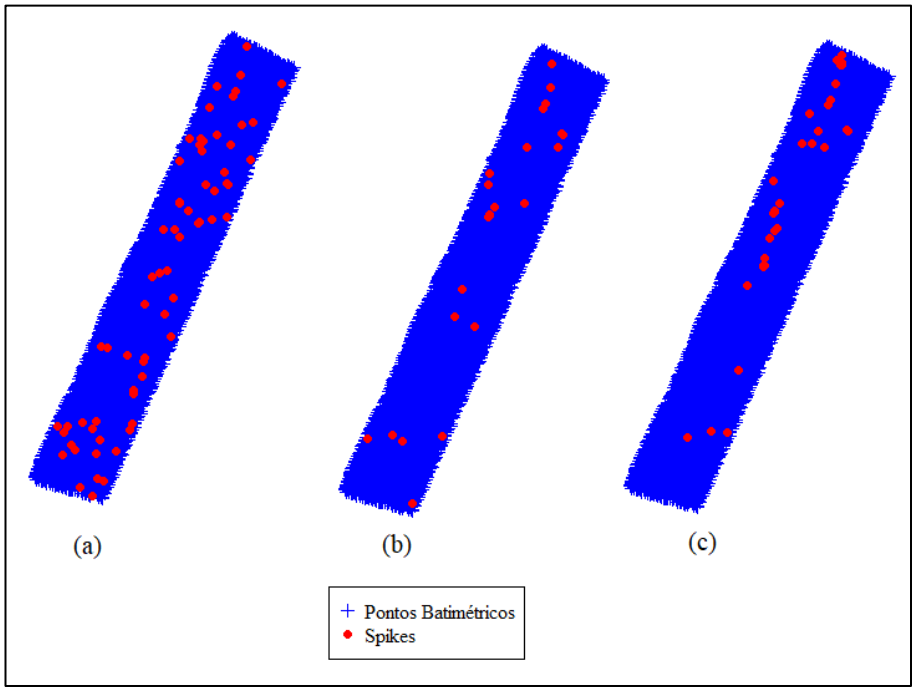

Figura 15 – *Spikes* detectados pelo AEDO a partir dos limiares das técnicas *Boxplot Ajustado* (a), *Z-Score Modificado* (b) e Método *δ* (c).

Destaca-se que para o processamento desse conjunto de dados foi utilizada uma máquina com sistema operacional Windows 10, memória RAM de 8GB (parcialmente dedicada ao *software R*) e processador Intel® Core™ i7-4500U CPU @ 1,80GHz 2,40 GHZ. O tempo de processamento foi de aproximadamente 9 minutos.

Objetivando avaliar minuciosamente a robustez da metodologia proposta e, particularmente, do *Método δ*, a área de estudo foi submetida ao processamento manual. Nessa fase, o profissional visualiza os dados através de uma interface gráfica, onde podem ser apresentados, dentre outras informações, a nuvem de pontos geral e parcial, perfis batimétricos e imagens de retroespalhamento (*backscattering),* e decide, baseado num julgamento qualitativo e analítico, qual profundidade pode ser considerada válida. De um modo geral, a área de estudo é relativamente pequena e possui um relevo plano, o que facilita a pesquisa manual por *spikes*. Soma-se a isso o fato das sondagens terem sido coletadas com altíssima qualidade, o que proporcionou

um processamento manual rápido e fácil. Todavia, a coleta de profundidades utilizando ecobatímetros multifeixe de baixa qualidade, sonares interferométricos ou sistemas *LiDAR (Light Detection And Ranging) Batimétricos*, geram, quase sempre, dados muito ruidosos, tornando o processamento manual com qualidade muito lento.

Através do processamento tradicional, foram detectadas 38 profundidades espúrias, aproximadamente, 0,47% de *spikes*, conforme é apresentado na Figura 16.

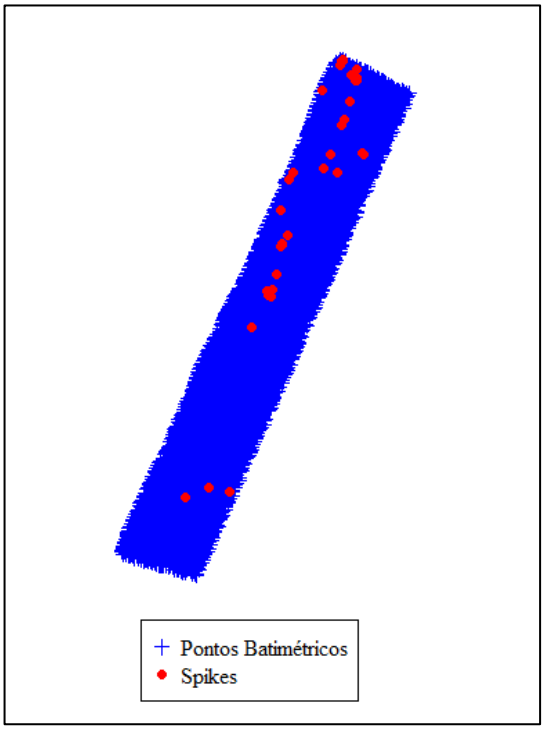

Figura 16 – *Spikes* detectados através de processamento manual.

Tomando o processamento manual como referência, percebe-se que a metodologia proposta aliada, principalmente, aos limiares do *Método δ*, apresentou-se bastante robusta, obtendo uma concordância de 82,35%, isto é, dos 34 *spikes* localizados pelo *Método δ,* 28 foram deletados durante as análises pelo processamento tradicional. Em outras palavras, a porcentagem de acerto foi de 74%, uma vez que *Método δ* detectou 28 dos 38 *spikes* presentes no conjunto de dados.

Os demais pontos não detectados, em quase sua totalidade, são *outliers* com magnitude muito baixa, em torno de 0,30 metros. Sendo assim, a falha na detecção destes pontos, possivelmente, não traz prejuízos ou ganhos aos produtos finais.

Essa problemática deve-se a dois motivos principais. O primeiro está relacionado com os pontos de borda. Conforme pode ser visto na Figura 17, o *Método δ* apresentou-se falho na detecção de alguns poucos *spikes* presentes nas extremidades da área de estudo. Isso é devido à natureza da metodologia desenvolvida, uma vez que

as buscas por *spikes* são realizadas nas subamostras formadas a partir de círculos e, assim, as análises nas bordas da nuvem de pontos são prejudicadas pelo número reduzido de verificações (redundância). Soma-se a isso a necessidade de pelo menos 7 profundidades para que as técnicas de detecção de *outliers* sejam aplicadas (seção 4).

O segundo motivo está associado com a aglomeração de *spikes*, o que pode mascarar a análise local, e/ou a presença destes em áreas de feriados. Nesses casos, a redundância de análises também é reduzida e, consequentemente, há uma diminuição da confiabilidade do processo. Em casos mais extremos, *spikes* presentes em área de feriados ou em áreas com baixa densidade de pontos podem não serem analisados, visto que as subamostras tendem a apresentar menos que 7 profundidades. Sendo assim, em concordância com o exposto na seção 4, a aplicação do método proposto presume que a nuvem de pontos é densa e não possui falhas de cobertura, isto é, feriados.

O *Método* δ*,* aparentemente, também apresentou problemas ao sinalizar 6 profundidades válidas como *spikes*. Ao investigar estas profundidades, percebeu-se que algumas delas estão presentes em áreas com falhas de cobertura e, conforme discutido, nestas ocasiões a metodologia pode apresentar falhas. As demais (ID 6026 e ID 2184), quando analisadas localmente, mostraram-se como profundidades destoantes dos vizinhos, com magnitude em torno de 50 centímetros, indicando, assim, falhas no processamento manual.

Outrem, conforme exposto da Tabela 7, estas profundidades obtiverem um $P_{outlier}$ em torno de 50%. Isso significa que, se o $P_{outlier}$ fosse configurado pelo analista com o valor de 51%, isto é, $P_{outlier} = 51\%$, o método localizaria apenas 2 falsos *spikes* (ID 6026 e ID 2184).

Tabela 7 – Profundidades válidas assinaladas como *spikes* pelo *Método* δ.

ID do Ponto	$N^{o}_{analisado}$	$N^{o}_{outlier}$	$P_{outlier}$ (%)
6026	14	8	57%
2184	11	6	55%
1497	12	6	50%
3312	8	4	50%
3313	8	4	50%
3760	10	5	50%

O *Boxplot Ajustado* superestimou quantitativamente a detecção de *spikes*, localizando profundidades que, em comparação com o processamento manual, não se caracterizam como *spikes*. Foram detectados 66 supostos *spikes*, dentre os quais, apenas 17 foram eliminados através do processamento manual. Sendo assim, a concordância e a porcentagem de acerto deste limiar foram, respectivamente, 25,76% e 44,74%.

Os pontos detectados de forma equivocada quando analisados localmente, através de perfis batimétricos brutos parciais, são profundidades que de certo modo destoam dos vizinhos, apresentando-se como ruídos locais. Entretanto, em um processamento para fins de navegação, a maioria dos pontos detectados por esse limiar são, de fato, profundidades válidas. Salienta-se que nenhum suposto *spike* detectado pelo limiar representa perigo direto à navegação e, assim, a exclusão desses pontos não traz perdas aos produtos finais gerados. Na verdade, tal fato pode representar ganhos, uma vez que o método proposto se mostrou como uma excelente ferramenta para suavização de superfícies batimétricas. Nesse sentido, nos casos em que a suavização do relevo submerso é requerida, por exemplo, quando se deseja construir curvas de nível ou modelos digitais de profundidade, a aplicação desse limiar pode manifestar-se bastante útil.

Por outro lado, nota-se que o limiar *Boxplot Ajustado*, mesmo localizando 66 supostos *outliers* de 38 possíveis, obteve apenas 44,74% de acerto. Analisando os *spikes* não detectados, nota-se a mesma problemática discutida anteriormente, isto é, a metodologia possui baixa eficiência para localizar *spikes* presentes nas extremidades da área de estudo, em feriados e em áreas com baixa densidade de pontos. Destaca-se também, a falha na detecção de profundidades espúrias nos casos em que existem aglomerações de *spikes*, como aquele presente na parte central da área de estudo (Figura 17).

Por fim, o limiar *Z-Score Modificado*, ao contrário do ocorrido na área de estudo simulada (seção 5.1), subestimou a detecção de *spikes*. Foram localizados 24 *spikes* dos quais 16 são realmente profundidades anômalas, isto é, uma concordância de 66,67% e uma porcentagem de acerto de 42,15%. Em termos gerais, esse limiar apresentou-se mais eficiente que o *Boxplot Ajustado.* Analisando os *spikes* sabidamente não detectados, em acordo com o ocorrido anteriormente, o *Z-Score Modificado* apresentou a mesma problemática, ou seja, *spikes* aglomerados, presentes

nas extremidades da área de estudo e em áreas de feriados não foram detectados. Por outro lado, 8 profundidades válidas foram, de forma equivocada, assinaladas como *spikes*, dentre as quais, 6 foram detectadas também pelo *Boxplot Ajustado* e as outras duas (ID 6377 e ID 7801) possuem uma diferença em relação a vizinhança de cerca de 20 centímetros.

Na Figura 17 é apresentado os *spikes* detectados pelo processamento manual e aqueles localizados pela metodologia proposta associada as técnicas de detecção de *outliers*.

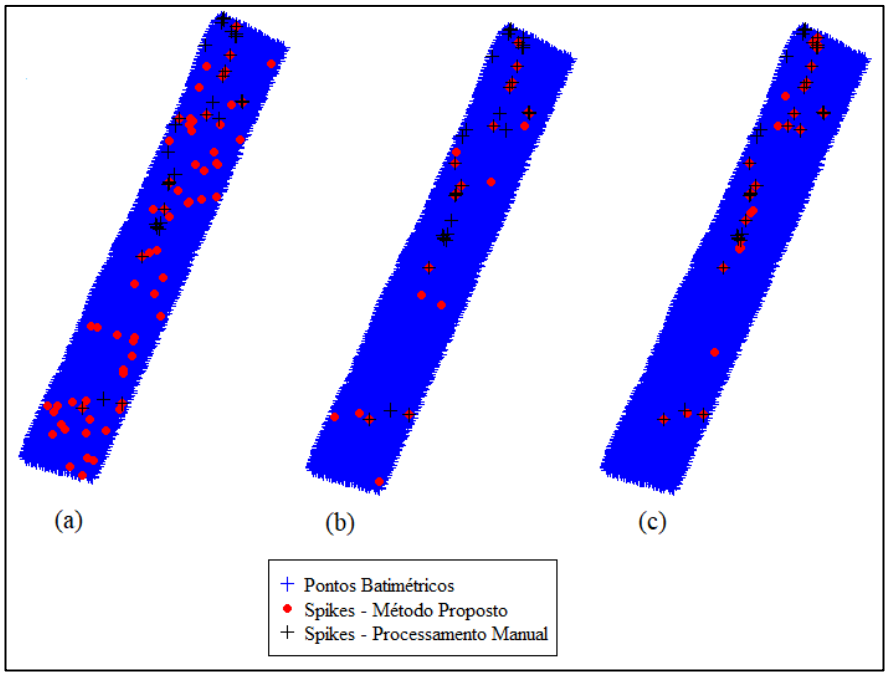

Figura 17 – *Spikes* excluídos através do processamento manual e *Spikes* detectados pelo AEDO a partir dos limiares das técnicas *Boxplot Ajustado* (a), *Z-Score Modificado* (b) e *Método* δ (c).

De acordo com o discutido, a metodologia proposta associada ao *Método* δ mostrou-se bastante robusta, obtendo cerca de 74% de acerto no processamento dos dados reais, enquanto os limiares *Boxplot Ajustado* e *Z-Score Modificado* atingiram, respectivamente, 44,73% e 42,10%. Em termos gerais, ambos os limiares, quando comparados com o processamento manual, mostraram-se como poderosas ferramentas de localização e eliminação de *spikes*. Alguns ajustes, como a escolha do raio de busca e a definição da constante c para o *Método* δ ainda são requeridas.

6. CONCLUSÕES

A metodologia proposta neste trabalho associada ao *Método* δ mostrou-se bastante eficiente, obtendo cerca de 74% de acerto no processamento dos dados reais, enquanto os limiares *Boxplot Ajustado* e *Z-Score Modificado* atingiram, respectivamente, 44,73% e 42,10%. Numa comparação entre os limiares adotados pelo método proposto, destaca-se que o δ e *Z-Score Modificado,* apresentaram uma concordância equivalente em todas as situações. A medida que as concordâncias entre o *Boxplot Ajustado* e *Z-Score Modificado* variaram de alta, para os dados simulados, a mediana, para os dados reais. Já a concordância entre *Boxplot Ajustado* e *Método* δ mostrou-se sempre mediana.

Em termos de robustez, nos testes realizados na área de estudos simulada, o AEDO foi capaz de detectar os *spikes* presentes no conjunto de dados sem eliminar estruturas submersas sabidamente pertencentes ao relevo submarino. Foram obtidas porcentagens de acerto variando de 60% a 90%. A partir da aplicação aos dados reais, pôde-se confirmar a eficiência do método proposto, bem como a sua versatilidade. Nesse caso, como supracitado, o limiar estabelecido pelo *Método* δ destacou-se em relação aos demais limiares.

Sendo assim, pode-se constatar que a aplicação dos métodos desenvolvidos poderá contribuir de sobremaneira com o tratamento de dados coletados através de sistemas de sondagem por faixa, uma vez que se apresenta como uma alternativa versátil e eficiente quando comparada com as técnicas atualmente empregadas, em especial, o processamento manual.

Embora o foco tenha sido dado às sondagens multifeixe, a metodologia proposta pode ser aplicada a sistemas de batimetria laser aerotransportados, sonares interferométricos e até mesmo em áreas afins, tais como: topografia, geodésia, fotogrametria, mineração, estatística, *etc*.

Para trabalhos futuros, recomenda-se a realização de testes e simulações, em diferentes tipos de relevo, visando melhorar a performance dos algoritmos em termos de tempo de processamento, definição do raio de busca para áreas com falhas de cobertura e definição da constante c do Método δ. No que tange a determinação do raio de busca, sugere-se ainda que diferentes raios de busca sejam aplicados simultaneamente durante o processamento. Isso permitirá uma maior redundância de análises, bem como uma melhora na qualidade dos resultados do processamento em

nuvem de pontos com falhas de cobertura. Outra problemática a ser solucionada é a questão das análises nas extremidades da nuvem de pontos. Todavia, acredita-se que computacionalmente a solução para tal fato é, ainda, inviável. Uma solução prática seria sempre realizar uma extrapolação da área de sondagem durante os levantamentos de campo. Por fim, a aplicação desta metodologia em áreas correlatas é desejável.

REFERÊNCIAS BIBLIOGRÁFICAS

ALTMAN, E. I. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. **Journal of Finance,** v. 23, n.4, p. 589-609, 1968.

ARTILHEIRO, F. M. F. **Analysis and Procedures of Multibeam Data Cleaning for Bathymetric Charting.** M. Eng. report, Department of Geodesy and Geomatics Engineering, Technical Report n. 191, University of New Brunswick, Fredericton, New Brunswick, Canada, 140p., 1998.

BACHMAIER, M & BACKES, M. Variogram or Semivariogram? Variance or Semivariance? Allan Variance or Introducing a New Term? **Mathematical Geosciences**, v. 43, n. 6, p. 735-740, 2011.

BJØRKE, J. T. & NILSEN, S. Fast trend extraction and identification of spikes in bathymetric data. **Computers & Geosciences**, v. 35, n. 6, p. 1061-1071, 2009.

BOTTELIER, P.; BRIESE, C.; HENNIS, N.; LINDENBERGH, R.; PFEIFER, N. Distinguishing features from outliers in automatic Kriging-based filtering of MBES data: a comparative study. **Springer Berlin Heidelberg**, 2005.

BOWLEY, A. L. **Elements of Statistics**. New York: Charles Scribner's Sons, 1920.

BRYS, G.; HUBERT, M.; STRUYF, A. A robust measure of skewness. **Journal of Computational and Graphical Statistics**, v. 13, n. 4, p. 996–1017, 2004.

CALDER, B. R. & MAYER, L. A. Automatic processing of high-rate, high-density multibeam echosounder data. **Geochemistry, Geophysics, Geosystems**, v. 4, n. 6, 2003.

CALDER, B. R. & SMITH, S. A time/effort comparison of automatic and manual bathymetric processing in real-time mode. In: Proceedings of the US Hydro 2003 Conference, **The Hydrographic Society of America**, Biloxi, MS. 2003.

CALDER, B. R. Automatic statistical processing of multibeam echosounder data. **The International Hydrographic Review**, v. 4, n. 1, p. 53-68, 2003.

CHAMBERS, J. M.; CLEVELAND, W. S.; KLEINER, B.; TUKEY, P. A. **Graphical Methods for Data Analysis**. Pacific Grove, CA: Wadsworth & Brooks/Cole, 1983.

CHENG, L.; MA, L.; CAI, W.; TONG, L.; LI, M.; DU, P. Integration of Hyperspectral Imagery and Sparse Sonar Data for Shallow Water Bathymetry Mapping**. Geoscience and Remote Sensing**. IEEE Transactions on, v. 53, n. 6, p. 3235-3249, 2015.

CHU, D. & HUFNAGLE JR, L. C. Time varying gain (TVG) measurements of a multibeam echo sounder for applications to quantitative acoustics. **IEEE**, 2006.

CLARKE, J. E. H. **Imaging and Mapping II: Submarine Acoustic Imaging Methods**. Notes of classes. Ocean Mapping Group. University of New Brunswick. 2014.

COOPER, M. A. R. **Control surveys in civil engineering**. Nichols Pub Co, 381p., 1987.

CRUZ, J.; VICENTE, J.; MIRANDA, M.; MARQUES, C.; MONTEIRO, C.; ALVES, A. Benefícios da utilização de sondadores interferométricos. **3as Jornadas de Engenharia Hidrográfica**. Instituto Hidrográfico Português, Lisboa, Portugal, 2014.

DEBESE, N. & BISQUAY, H. Automatic detection of punctual errors in multibeam data using a robust estimator. **The International Hydrographic Review**, v. 76 n. 1, p. 49-63, 1999.

DEBESE, N. Multibeam Echosounder Data Cleaning Through an Adaptive Surface-based Approach. **In: US Hydro 07 Norfolk**, 18p., 2007.

DEBESE, N.; MOITIÉ, R.; SEUBE, N. Multibeam echosounder data cleaning through a hierarchic adaptive and robust local surfacing. **Computers & Geosciences**, v. 46, p. 330-339, 2012.

DHN – Diretoria de Hidrografia e Navegação. **NORMAM 25 – Normas da Autoridade Marítima para Levantamentos Hidrográficos**. Marinha do Brasil, 2014.

DIGGLE, P. J. & RIBEIRO JÚNIOR, P. J. **Model-based Geostatistics**. New York: Springer, 229p., 2007.

EEG, J. On the identification of spikes in soundings. **The International Hydrographic Review**, v. 72, n. 1, p. 33-41, 1995.

ELLMER, W.; ANDERSEN, R. C.; FLATMAN, A.; MONONEN, J.; OLSSON, U.; ÖIÅS, H. Feasibility of Laser Bathymetry for Hydrographic Surveys on the Baltic Sea. **The International Hydrographic Review**, n. 12, p. 33-50, 2014.

FERREIRA, Í. O.; RODRIGUES, D. D.; NETO, A. A.; MONTEIRO, C. S. Modelo de incerteza para sondadores de feixe simples. **Revista Brasileira de Cartografia**, v. 68, n. 5, p. 863-881, 2016a.

FERREIRA, Í. O.; NETO, A. A.; MONTEIRO, C. S. O uso de embarcações não tripuladas em levantamentos batimétricos**. Revista Brasileira de Cartografia**, v. 68, n. 10, p. 1885-1903, 2017a

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.**Coleta, processamento e análise de dados batimétricos**. 1ª ed. Saarbrucken: Novas Edições Acadêmicas, v. 1, 100p., 2015.

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.; ROSA, L. M. F. In bathymetric surfaces: IDW or Kriging? **Boletim de Ciências Geodésicas**, v. 23, n. 3, p. 493-508, 2017b.

FERREIRA, Í. O.; SANTOS, G. R.; RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. **Revista Brasileira de Cartografia**, Rio de Janeiro, v. 65, n. 5, p. 831-842, 2013.

FERREIRA, Í. O.; ZANETTI, J.; GRIPP, J. S.; MEDEIROS, N. G. Viabilidade do uso de imagens do sistema Rapideye na determinação da batimetria de águas rasas. **Revista Brasileira de Cartografia**, v. 68, n. 7, p. 1331-1340, 2016b.

GAO, J. Bathymetric mapping by means of remote sensing: methods, accuracy and limitations. **Physical Geography**, v. 33, n. 1, p. 103-116, 2009.

GUENTHER, G. C.; THOMAS, R. W. L. ; LAROCQUE, P. E. Design considerations for achieving high accuracy with the Shoals bathymetric Lidar system. In: CIS Selected Papers: Laser Remote Sensing of Natural Waters-From Theory to Practice. **International Society for Optics and Photonics**, p. 54-71, 1996.

HINKLEY, D. V. On power transformations to symmetry. **Biometrika**, v. 62, n. 1, p. 101–111, 1975.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. **Understanding robust and exploratory data analysis**. New York: Wiley, 433p., 1983.

HÖHLE, J. & HÖHLE, M. Accuracy assessment of digital elevation models by means of robust statistical methods. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 64, n. 4, p. 398-406, 2009.

HUBERT, M. & VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Journal of Computational statistics & data analysis**, v. 52, n. 12, p. 5186-5201, 2008.

HYPACK, Inc. **Hypack – Hydrographic Survey Software User Manual**. Middletown, USA, 1784p., 2012.

IGLEWICZ, B. & HOAGLIN, D. **How to detect and handle outliers**. Milwaukee, Wis.: ASQC Quality Press, 87p., 1993.

IHO – International Hydrographic Organization. **C-13: IHO Manual on Hydrography**. Mônaco: International Hydrographic Bureau, 540p., 2005.

IHO – International Hydrographic Organization. **S-44: IHO Standards for Hydrographic Surveys**. Special Publication n. 44–5th. Mônaco: International Hydrographic Bureau, 36p., 2008.

INSTITUTO HIDROGRÁFICO. **Especificação Técnica para Produção de cartografia hidrográfica**. Marinha Portuguesa, Lisboa, Portugal, v 0.0, 24p., 2009.

ISAAKS, E. H. **The application of Monte Carlo methods to the analysis of spatially correlated data**. PhD Thesis, Department of Applied Earth Sciences, Stanford University, USA, 213p., 1990.

JONG, C.D.; LACHAPELLE, G.; SKONE, S.; ELEMA, I. A. **Hydrography**. 2ª ed. Delft University Press: VSSD, 354p., 2010.

LINZ – Land Information New Zealand. **Contract Specifications for Hydrographic Surveys**. New Zealand Hydrographic Authority, v. 1.2, 111p., 2010.

LU, D.; LI, H.; WEI, Y.; ZHOU, T. Automatic outlier detection in multibeam bathymetric data using robust LTS estimation. In: 3rd International Congress on Image and Signal Processing (CISP), **IEEE**, v. 9, p. 4032-4036, 2010.

LURTON, X. An introduction to underwater acoustics: principles and applications. **Springer Science & Business Media**, 349p., 2002.

MALEIKA, W. The influence of the grid resolution on the accuracy of the digital terrain model used in seabed modeling. **Marine Geophysical Research**, v. 36, n. 1, p. 35-44, 2015.

MATHERON, G. **Les variables régionalisées et leur estimation**. Paris: Masson, 306p., 1965.

MIKHAIL, E. & ACKERMAN, F. **Observations and Least Squares**. University Press of America, 497p., 1976.

MOOD, A. M. **Introduction to the theory of statistics**. McGraw-Hi1l series in probability and statistics, 564p., 1913.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the Theory of Statistics**. McGraw-Hill International, 577p., 1974.

MORETTIN, P. A. & BUSSAB, W. O. **Estatística básica**. 5ª ed. São Paulo: Editora Saraiva, 526p., 2004.

MOTAO, H.; GUOJUN, Z.; RUI, W.; YONGZHONG, O.; ZHENG, G. Robust method for the detection of abnormal data in hydrography. **The International Hydrographic Review**, v. 76, n. 2, p. 93-102, 1999.

MOURA, A.; GUERREIRO, R.; MONTEIRO, C. As potencialidades da derivação de batimetria a partir de imagens de satélite multiespetrais na produção de cartografia náutica. **4as Jornadas de Engenharia Hidrográfica**. Instituto Hidrográfico Português, Lisboa, Portugal, 2016.

NOAA – National Oceanic and Atmospheric Administration. **Field Procedures Manual**. Office of Coast Survey, 2011.

PASTOL, Y. Use of Airborne lidar Bathymetry for Coastal Hydrographic Surveying: The French Experience. **Journal of Coastal Research,** n. 62, p. 6-18, 2011.

R CORE TEAM. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2017.

RIBEIRO JÚNIOR, P.J. & DIGGLE, P.J. **GeoR: a package for geostatistical analysis**. R-News. v. 1, p. 15-18, 2001.

ROUSSEEUW, P. J.; & CROUX, C. Alternatives to the median absolute deviation. **Journal of the American Statistical association**, v. 88, n. 424, p. 1273-1283, 1993.

SANTOS, A. M. R. T.; SANTOS, G. R.; EMILIANO, P. C.; MEDEIROS, N. G.; KALEITA, A. L.; PRUSKI, L. O. S. Detection of inconsistencies in geospatial data with geostatistics. **Boletim de Ciências Geodésicas,** v. 23, n. 2, p. 296-308, 2017.

SANTOS, A. P. **Controle de qualidade cartográfica: metodologias para avaliação da acurácia posicional em dados espaciais**. Tese (Doutorado). Programa de Pós-Graduação em Engenharia Civil, Departamento de Engenharia Civil, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 172p., 2015.

SANTOS, A. P.; RODRIGUES, D. D.; SANTOS, N. T.; GRIPP JUNIOR, J. Avaliação da acurácia posicional em dados espaciais utilizando técnicas de estatística espacial: proposta de método e exemplo utilizando a norma brasileira. **Boletim de Ciências Geodésicas**, v. 22, n. 4, p. 630-650, 2016.

SEO, S. **A review and comparison of methods for detecting outliers in univariate data sets**. Master Of Science, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, USA, 59p., 2006.

TUKEY, J.W. **Exploratory Data Analysis**. Princeton, Ed. Pearson (1977).

URICK, R. I. **Principles of Underwater Acoustics.** Toronto: McGraw-Hill, 1975.

USACE – U.S. Army Corps of Engineers. **Hydrographic Surveying**. Engineer Manual n. 1110-2-1003. Department of the Army. Washington, D. C. USA, 2013.

VANDERVIEREN, E. & HUBERT, M. An adjusted boxplot for skewed distributions. **Proceedings in Computational Statistics**, p. 1933-1940, 2004.

VICENTE, J. P. D. **Modelação de dados batimétricos com estimação de incerteza**. Dissertação (Mestrado). Programa de Pós-Graduação em Sistemas de Informação Geográfica Tecnologias e Aplicações, Departamento de Engenharia Geográfica, Geofísica e Energia, Universidade de Lisboa, Portugal, 158p., 2011.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. In. NOVAES, R. F.; ALVAREZ V.; V. H.; SCHAEFER, C. E G. R. **Tópicos em ciências do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, v.1. p. 2-54, 2000.

WARE, C; SLIPP, L.; WONG, K. W.; NICKERSON, B.; WELLS, D. E.; LEE, Y. C.; DODD, D.; COSTELLO, G. A System for Cleaning High Volume Bathymetry. **The International Hydrographic Review**, v. 69, n. 2, p. 77-94, 1992.

WARE, C.; KNIGHT, W.; WELLS, D. Memory intensive statistical algorithms for multibeam bathymetric data. **Computers & Geosciences**, v. 17, n. 7, p. 985-993, 1991.

WARRICK, A.W. & NIELSEN, D.R. Spatial variability of soil physical properties in the field. In: HILLEL, D. **Applications of soil physics**. New York: Academic Press, p.319-344, 1980.

WILLIAMSON, D. F.; PARKER, R. A.; KENDRICK, J. S. The box plot: a simple visual method to interpret data. **Annals of internal medicine**, v. 110, n. 11, p. 916-921, 1989.

CAPÍTULO 2. PROPOSTA METODOLÓGICA PARA AVALIAÇÃO DA QUALIDADE VERTICAL DE SONDAGENS BATIMÉTRICAS MONOFEIXE, COM ÊNFASE EM TESTES DE NORMALIDADE E INDEPENDÊNCIA

Resumo:
A destinação do produto resultante de um levantamento hidrográfico é altamente dependente da incerteza da profundidade coletada, sendo assim, o dado batimétrico deve ser sempre entregue com um nível de confiança estatisticamente comprovado. Todavia, devido à natureza das informações batimétricas, a avaliação da qualidade vertical das sondagens não é tarefa simples. É muito comum no levantamento hidrográfico estimar a qualidade vertical de uma sondagem batimétrica através de linhas de verificação, porém, quase sempre estas estimativas são realizadas sem qualquer critério estatístico. Diante disso, este trabalho propõe uma metodologia para avaliação estatística das profundidades coletadas por sistemas de sondagem batimétrica monofeixe através de amostras de discrepâncias, abordando normalidade e independência dos dados. Através dos resultados gerados neste trabalho pode-se perceber que numa análise estatística coerente, deve-se, primeiramente, verificar a presença de *outliers* e efetuar testes de independência e normalidade e só então, inferir quaisquer conclusões acerca das incertezas relacionadas ao produto analisado. Neste capítulo também é apresentado um novo estimador robusto para a avaliação da incerteza vertical amostral.

1. INTRODUÇÃO

Entende-se por levantamento hidrográfico o conjunto de atividades executadas com a finalidade de obtenção de dados batimétricos, da natureza física e configuração do fundo submerso, das alturas e variações do nível das águas, dentre outros (IHO, 2005). Embora o principal interesse desses levantamentos seja a navegação aquaviária, diversas outras finalidades são atendidas pelos dados coletados, tais como: o estabelecimento e manutenção de obras civis (pontes, portos, píeres), a locação de cabos e dutos e a prospecção de recursos minerais (FERREIRA et al., 2017a).

O termo levantamento batimétrico, é encontrado regularmente na literatura, constitui-se na principal tarefa de um levantamento hidrográfico, tendo por objetivo realizar medições de profundidades associadas a uma posição na superfície (FERREIRA et al., 2016b). Essas profundidades são utilizadas na construção dos Modelos Digitais de Elevação das superfícies submersas que servem de insumo a diversas análises (FERREIRA et al., 2017b). Em reservatórios de água, por exemplo, sejam aqueles destinados ao abastecimento ou a geração de energia, os dados

provenientes das sondagens batimétricas se mostram indispensáveis para a modelagem e gestão dos recursos hídricos, que são utilizados para estimar o grau de assoreamento, calcular volumes de armazenamento, atualizar as curvas de capacidade, modelagem hidrodinâmica, além de subsidiar informações aos órgãos competentes nas tomadas de decisões (WMO, 2003; FERREIRA et al., 2012; ANA, 2013).

No planejamento da batimetria é comum a divisão da superfície a ser mapeada em uma malha de linhas equidistantes, doravante denominadas linhas regulares de sondagem. Estas são percorridas pela plataforma de sondagem permitindo a coleta de dados de profundidade e posição. Os espaçamentos entre linhas, bem como a orientação, são altamente dependentes da tecnologia adotada na coleta dos dados (IHO, 2008; DHN, 2014).

Para medição da profundidade são utilizados prioritariamente sensores acústicos, como ecobatímetros monofeixe *(SBES - Single Beam Echo Sounders)* e multifeixe *(MBES – Multibeam Echo Sounders)*. Recentemente, tem crescido o número de usuários dos chamados sonares interferométricos, embora estes ainda não sejam utilizados no Brasil para fins de produção ou atualização da cartografia náutica. O MBES é um dos sistemas mais efetivos para medição da profundidade, pois proporciona uma cobertura (ensonificação) quase total do fundo submerso devido à elevada taxa de medição, com o consequente aumento da resolução e da capacidade de detecção de objetos (IHO, 2005; USACE, 2013; MALEIKA, 2015). No entanto, sistemas multifeixe ainda possuem um custo de aquisição elevado, em alguns casos, de até 10 vezes o valor de um sistema monofeixe, a coleta e o processamento de dados são mais complexos e, por isso, exigem profissionais devidamente capacitados. Embora forneçam maior detalhe e seu uso seja obrigatório em alguns casos, como por exemplo, em pesquisas para fins de navegação em áreas restritas (IHO, 2008; DHN, 2014), o seu emprego em reservatórios de armazenamento nem sempre é justificado.

Nesse sentido, o SBES ainda é utilizado, principalmente em reservatórios de baixa profundidade, onde o MBES perde a eficiência, e em pesquisas em que se faz necessário estimar a camada de lama fluída depositada no fundo do reservatório. Esses equipamentos, ao contrário do MBES, emitem apenas um feixe acústico *(ping)*, determinando assim uma única cota de profundidade por ciclo. As profundidades são georreferenciadas, preferencialmente, através de sistemas diferenciais de posicionamento GNSS (*Global Navigation Satellite System*) (FERREIRA et al., 2015).

No Brasil o uso do SBES é regulamentado por órgãos como: Agência Nacional de Águas (ANA, 2013), Agência Nacional de Energia Elétrica (CARVALHO et al., 2000) e pela Marinha do Brasil (DHN, 2014). Esta última, seguindo recomendações da IHO *(International Hydrographic Organization)*, permite o uso dos sistemas monofeixe em casos específicos, como em levantamentos de Ordem 1b e 2, conforme especificado na Publicação Especial S-44, 5ª edição (IHO, 2008). Outras organizações como *World Meteorological Organization* (WMO, 2003), *U.S. Geological Survey* (SEKELLICK & BANKS, 2010; ATHEARN et al., 2010) e *U.S. Army Corps of Engineers* (USACE, 2013) utilizam essa tecnologia para levantamentos batimétricos de reservatórios e águas interiores.

Independente da tecnologia empregada numa sondagem batimétrica, os dados coletados conterão incertezas, sejam elas de natureza grosseira, sistemática ou aleatória (FERREIRA et al., 2016a). Inúmeras fontes de incertezas são observadas na aquisição da profundidade e posição, afetando, consequentemente, o produto final (Modelo Batimétrico). Deve-se atentar que não é possível gerar modelos batimétricos acurados se os dados estiverem eivados de incertezas de magnitude maior que uma determinada tolerância definida por norma. A estimativa de volume em um reservatório, por exemplo, é altamente dependente da incerteza do dado batimétrico coletado, conforme observado por Veiga et al. (2010) e USACE (2013).

De acordo com USACE (2002) em uma sondagem batimétrica a incerteza de medição da profundidade possui muitos componentes em potencial. Sendo eles: o método de medição, a velocidade de propagação do som na água, a largura de feixe do transdutor, o tipo e formato de fundo, os movimentos da plataforma de sondagem *(heave-pitch-roll)* e a profundidade de imersão do transdutor *(draft).* Todas essas fontes contribuem com a incerteza da profundidade e consequentemente com a incerteza da modelagem batimétrica. Hare et al. (2011) ainda evidenciam que em uma sondagem batimétrica existem fontes de incertezas que contribuem apenas com a incerteza vertical, fontes de incerteza que contribuem apenas com a incerteza horizontal e aquelas que contribuem com ambas. Sendo assim, é necessário avaliar cada uma destas fontes de erros a fim de produzir dados confiáveis às diversas análises a que eles se propõem.

Se todas as fontes de incertezas individuais forem devidamente avaliadas, pode-se combiná-las através da aplicação da lei de propagação de covariâncias, desde que todos os pressupostos sejam atendidos, para fornecer uma estimativa da IPT

(Incerteza Propagada Total, do Inglês *TPU – Total Propagated Uncertainty*) do sistema de sondagem, o que fornece uma estimativa da possível incerteza do dado batimétrico coletado, tal como proposto por Ferreira et al. (2016a). A IPT calculada no plano horizontal origina a IHT (Incerteza Horizontal Total, do Inglês *THU – Total Horizontal Uncertainty*), de modo análogo, no plano vertical, têm-se a IVT (Incerteza Vertical Total, do Inglês *TVU – Total Vertical Uncertainty*)

No entanto, uma propagação de covariâncias, apesar de considerar incertezas obtidas em todas as etapas de um levantamento batimétrico, sejam elas sistemáticas ou aleatórias, estabelece apenas uma estimativa da qualidade do levantamento baseada nos possíveis desvios não correlacionados do sistema de sondagem (IHO, 2005; LINZ, 2010). Além disso, as incertezas utilizadas no cálculo da IPT são, em sua maioria, resultantes de testes de laboratório, isto é, não consideram as reais condições de operação (FERREIRA et al., 2016a).

Pode-se concluir que esta metodologia apenas demonstra a capacidade do sistema de levantamento e não leva em consideração a aleatoriedade de medidas obtidas naturalmente, como é o caso das sondagens batimétricas. Além disso, conforme afirma IHO (2008), simplesmente fazer uso de um equipamento teoricamente capaz de atingir a incerteza requerida, não é necessariamente o bastante. Fatores tais como: a maneira como o equipamento é montado, utilizado e o modo como ele interage com os demais componentes do sistema de levantamento interferem na incerteza do produto final. O hidrógrafo, nesse caso, é uma peça fundamental, pois precisa conhecer as técnicas de medição e os impactos das incertezas inerentes ao processo. Sendo assim, é preferível que a avaliação da qualidade do levantamento seja baseada em observações redundantes.

Devido à natureza dos levantamentos batimétricos, a coleta de profundidades redundantes não é tão simples e, na prática, é algo pouco provável de ocorrer. Portanto, estimar a acurácia de um levantamento hidrográfico torna-se algo impraticável e a precisão somente pode ser obtida através de suposições estatísticas. Conforme sugerido por IHO (2008), INMETRO (2012a, b) e Ferreira et al. (2016a), neste trabalho, será dada preferência ao termo incerteza em substituição a termos como: acurácia e erro.

Dado a dificuldade de reamostragem de determinadas feições submersas, em levantamentos hidrográficos realizam-se linhas de verificação, que cruzam as linhas regulares de sondagem. As linhas de verificação devem ser coletadas em momentos

distintos e em condições atmosféricas favoráveis. Supondo-se que os *outliers (spikes e tops)* tenham sido eliminados ou minimizados, procede-se a avaliação da qualidade vertical do levantamento a partir da comparação entre as profundidades próximas às interseções entre as linhas regulares de sondagem e as linhas de verificação. A partir das discrepâncias geradas são realizadas análises estatísticas objetivando avaliar a qualidade vertical do levantamento (IHO, 2008; DHN, 2014; FERREIRA et al., 2015).

As linhas de verificação indicam o nível de conformidade ou repetição das medidas, porém não indicam acurácia absoluta uma vez que os dados, geralmente, são coletados a partir da mesma plataforma de sondagem e, nesse caso, há um grande número de fontes de incertezas comuns em potencial entre os dados das linhas regulares e das linhas de verificação (IHO, 2008).

USACE (2002) afirma que esses métodos de avaliação somente fornecem uma estimativa da incerteza das medidas de profundidade, pois não são um teste independente. Todavia, as linhas de verificação fornecem um bom indicador de qualidade vertical do levantamento, e neste caso, seu uso é recomendável e exigido por normas como IHO (2008), Instituto Hidrográfico (2009), LINZ (2010), NOAA (2011), USACE (2013) e DHN (2014).

Embora as normativas exijam a execução de linhas de verificação, elas não apresentam procedimentos estatísticos para avaliação da incerteza vertical do levantamento, em outras palavras, não há recomendações acerca do tratamento estatístico das discrepâncias. No geral, estas normas apresentam apenas as tolerâncias para a incerteza vertical em um nível de confiança de 95%, o que sugere que a incerteza do levantamento seja estimada também em intervalos com 95% de confiança.

Segundo Höhle & Höhle (2009), a maioria das normas de avaliação cartográfica assumem que as discrepâncias são livres de *outliers*, seguem uma distribuição de probabilidade normal e são variáveis aleatórias independentes e identicamente distribuídas, suposições indispensáveis para um tratamento estatístico clássico (MORETTIN & BUSSAB, 2004). Tais pressupostos também são comumente assumidos por normativas de levantamentos hidrográficos (IHO, 2008; DHN, 2014). Entretanto, é sabido que essas hipóteses estatísticas dificilmente são atendidas e/ou verificadas e, caso negligenciadas, comprometem as análises (LI et al., 2005; MAUNE, 2007; SANTOS, 2015; SANTOS et al., 2017).

Dentre as três suposições, a mais difícil de garantir, no caso de análises espaciais de variáveis contínuas, é a independência entre os dados da amostra. Uma

vez que a natureza dos dados espaciais, tal como aqueles obtidos através dos levantamentos hidrográficos, é de possuir autocorrelação espacial, assim como sugere a 1ª lei da Geografia (TOBLER, 1970; FERREIRA et al., 2013, 2015). A distribuição normal é assumida por ser talvez a mais importante distribuição contínua. Sua importância se deve a vários fatores, entre eles podemos citar o Teorema Central do Limite (TCL), o qual é um resultado fundamental em aplicações práticas e teóricas. Na teoria das probabilidades, esse teorema afirma que quando o tamanho da amostra aumenta, a distribuição da média amostral aproxima-se cada vez mais de uma distribuição normal padrão. O Teorema Central do Limite também é utilizado na demonstração de diversos outros teoremas estatísticos (FISCHER, 2010).

Para quantificar a incerteza vertical nos levantamentos batimétricos, geralmente, é empregado a raiz do erro quadrático médio ou, do Inglês, *Root Mean Square Error (RMSE)* (SUSAN & WELLS, 2000; EEG, 2010; SEKELLICK & BANKS, 2010). Apesar de altamente influenciado pela presença de *outliers*, o $RMSE$ é considerado um estimador robusto e, por esse motivo, largamente utilizado também nas ciências geodésicas (MIKHAIL & ACKERMANN, 1976;). Todavia, é frequente a associação do $RMSE$ às distribuições de probabilidade para definição de intervalos de confiança, *a priori,* de forma equivocada (GREENWALT & SCHULTZ, 1962; FGDC, 1998; SUSAN & WELLS, 2000; EEG, 2010; SEKELLICK & BANKS, 2010).

Face ao exposto, é nítido que os indicadores estatísticos de incerteza dos levantamentos hidrográficos devem considerar que *outliers* possam existir e que a distribuição das discrepâncias pode não ser normal. Além dos mais, as metodologias empregadas devem considerar a estrutura de dependência espacial das variáveis analisadas e os intervalos de confiança devem ser teoricamente ótimos.

Sendo assim, o objetivo deste trabalho é propor uma metodologia para avaliar a qualidade vertical de levantamentos batimétricos monofeixe através de observações coletadas por meio de linhas de verificação, abordando independência e normalidade dos dados, bem como a presença, ou não, de dados discrepantes *(outliers).* A metodologia proposta é chamada, neste trabalho de tese, de MAIB (Metodologia para Avaliação da Incerteza de dados Batimétricos). As pesquisas realizadas neste capítulo permitiram também o desenvolvimento de um novo estimador para o cálculo da incerteza vertical amostral das profundidades coletadas num levantamento batimétrico.

2. PROPOSIÇÃO DO MÉTODO

A metodologia proposta neste estudo baseia-se no trabalho de Santos (2015) e é fundamentada em teoremas básicos da estatística clássica e Geoestatística, encontrados em vasta bibliografia, como: Mood (1913), Mood et al. (1974), Morettin & Bussab (2004), entre outros. Todo o método, incluindo a parte inovadora, foi implementado no *software* livre R (R Core Team, 2017) e o *script* pode ser consultado no Apêndice. O fluxograma ilustrado na Figura 1 resume a metodologia proposta, intitulada ***MAIB (Metodologia para Avaliação da Incerteza de dados Batimétricos)***.

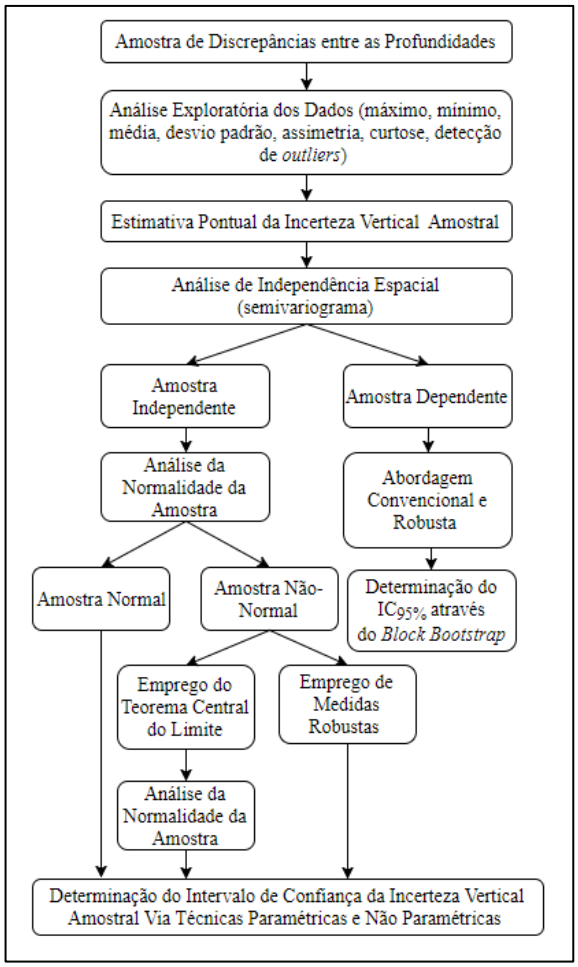

Figura 1 – Fluxograma da metodologia proposta para avaliação intervalar da incerteza vertical em dados de sondagem batimétrica monofeixe.

Num primeiro momento, conforme exposto na Figura 1, a MAIB é capaz de importar a amostra de discrepância georreferenciada. O arquivo deve estar no formato XYZdz, em que X e Y representam, respectivamente, as coordenadas posicionais, sejam elas locais, projetadas ou geodésicas, Z representa a profundidade e dz denota a discrepância entre "profundidades homólogas".

A próxima etapa consiste, assim como em qualquer análise estatística coerente, em efetuar a análise exploratória da amostra de discrepâncias, que pode fornecer informações importantes da distribuição dos dados e da presença de tendência e *outliers* (MORETTIN & BUSSAB, 2004; FERREIRA et al., 2013). Basicamente, a MAIB propõe a construção e interpretação de gráficos (histogramas, *boxplot*, *Q-Q Plot, etc.*) e de estatísticas (média, variância, mínimo, máximo, coeficientes de assimetria e curtose, *etc*.).

Para detecção de *outliers* na amostra de discrepâncias, a metodologia utiliza três técnicas, a saber: *Boxplot de Tukey* (TUKEY, 1977); o *Boxplot Ajustado* (VANDERVIEREN & HUBERT, 2004) e o *Z-Score Modificado* (IGLEWICZ & HOAGLIN, 1993). Cabe ao analista verificar e escolher qual metodologia se adequa melhor aos dados. Em todos os casos, o algoritmo da MAIB irá gerar três arquivos, cada um destes contendo os dados originais, excluindo-se os *outliers* detectados por cada técnica. Apesar dessa etapa estar implementada de forma automatizada, a exclusão de possíveis *outliers* deve ser realizada com cautela, uma vez que, dentre outros motivos, as técnicas estatísticas utilizadas supõem independência espacial. Destaca-se que a eliminação de dados deve sempre ser precedida de alto grau de confiabilidade, visto que, na mesma proporção que os *outliers* afetam a análise, a eliminação destes também pode omitir informações importantes (SANTOS et al., 2017).

Feito isso, pode-se aplicar medidas pontuais de acurácia teórica, tal como o *RMSE (Root Mean Square Error)*, para estimar a incerteza vertical amostral (MIKHAIL & ACKERMAN, 1976). Contudo, para construção de intervalos de confiança estatisticamente ótimos deve-se, primeiramente, avaliar a distribuição dos dados, bem como a autocorrelação espacial. Assim, no próximo passo, são realizadas análises de independência e normalidade.

Sabe-se que os testes de normalidade univariada pressupõe a independência estatística dos dados, sendo assim, deve-se primeiramente confirmar a presença de autocorrelação espacial e só então inferir sobre a normalidade da amostra de discrepâncias (MOOD, 1913; MOOD et al., 1974; MORETTIN & BUSSAB, 2004; SANTOS, 2015).

Nesse sentido, a próxima etapa constitui-se na verificação da presença de independência espacial entre os dados. Para isso, devido principalmente à sua eficiência, sugere-se o uso do semivariograma, ferramenta utilizada pela Geoestatística

para avaliar a autocorrelação espacial dos dados (MATHERON, 1965; FERREIRA et al., 2015).

Uma análise detalhada da construção do semivariograma podem ser consultados em: Vieira (2000), Ferreira et al. (2013) e Santos (2015) e, de forma resumida, no Capítulo 1 deste texto. A Figura 2 exemplifica dois semivariogramas experimentais, o primeiro para dados com dependência espacial (Figura 2a) e o segundo para dados independentes espacialmente (Figura 2b). Neste último caso, diz-se que o semivariograma apresentou efeito pepita puro.

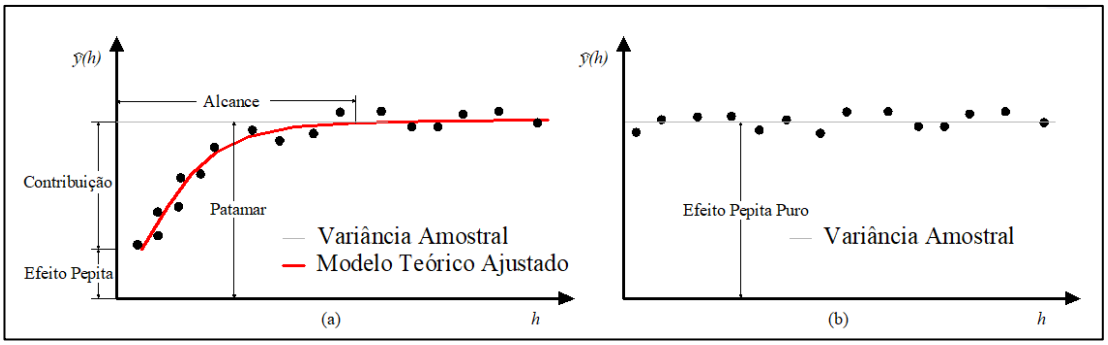

Figura 2 – Exemplo de semivariogramas para dados espacialmente dependentes (a) e espacialmente independentes (b).

Para construção do semivariograma experimental, deve-se definir uma distância de passo para que sejam selecionados os pares de discrepância, e uma distância limite para o crescimento dos passos. Sendo assim, o algoritmo desenvolvido calcula a distância máxima entre as discrepâncias e, baseado nesta informação, geram-se três semivariogramas, o primeiro com alcance igual a 75% da distância máxima; o segundo com 50% da distância máxima e o terceiro com 25%. Diante destes gráficos, o analista pode decidir sobre a existência, ou não, de dependência espacial. Nessa etapa, o algoritmo também é capaz de gerar o envelope de Monte Carlo (simulação de Monte Carlo) para confirmar, de forma explanatória, a existência de autocorrelação espacial (ISAAKS,1990).

Pode-se ainda recorrer a classificação proposta por Cambardella et al. (1994), que consideram uma dependência espacial forte quando o semivariograma apresentar efeito pepita menor ou igual a 25% do patamar, isto é, $\frac{Pepita}{Patamar} < 0{,}25$, moderada quando a relação entre o efeito pepita e o patamar estiver entre 25 e 75% e fraca quando a relação for maior que 75%.

Constatada a independência entre as amostras de discrepâncias pode-se recorrer a testes de normalidade como, por exemplo: *Anderson-Darling, Cramer-Von*

Mises, D'Agostino-Pearson, Jarque-Bera, Kolmogorov-Smirnov e *Shapiro-Wilk*. Segundo Machado et al. (2014), tem-se notado na literatura atual uma preferência pela aplicação do *Kolmogorov-Smirnov* (DOOB, 1949) e *Shapiro-Wilk* (SHAPIRO & WILK, 1965). Neste trabalho, propõe-se o uso do teste de *Kolmogorov-Smirnov* (KS), ao nível de confiança de 95%, uma vez que a aplicação do *Shapiro-Wilk* se limita a amostras com até 5.000 pontos (FILHO, 2013). Assim, aplica-se o teste, tendo como resposta o valor-p. Se o valor-p > 0,05, a amostra é dita normal ao nível de significância de 5%.

Nessa etapa, também podem-se utilizar ferramentas visuais, como, por exemplo, o histograma das discrepâncias ou gráfico Quantil-Quantil *(Q-Q Plot)*. A partir do histograma pode-se ter uma primeira impressão acerca da normalidade dos dados, que pode ser comprovada, ou não, através do *Q-Q Plot*. Esse gráfico permite checar a adequação da distribuição de frequência dos dados à uma distribuição de probabilidades qualquer, em suma, os quantis da função de distribuição empírica são plotados contra os quantis teóricos da distribuição de probabilidades, nesse caso, a distribuição normal. Se a distribuição empírica é normal, o gráfico será apresentado como uma linha reta (HÖHLE & HÖHLE, 2009). Tais gráficos, conforme citado, são construídos durante a análise exploratória. Todavia, evidencia-se que essas análises visuais devem ser sempre confirmadas através dos testes supracitados.

Assim sendo, dessa etapa em diante, a MAIB sugere a subdivisão da análise em três categorias: amostra independente e normal; amostra independente e não-normal e amostras dependentes.

2.1. Amostra independente e normal

A incerteza do levantamento hidrográfico pode ser estimada através do estimador $RMSE$, dado pela Equação 1.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(d_i)^2}{n}} \quad (1)$$

em que di corresponde as discrepâncias entre as "profundidades homólogas" observadas e n o número de discrepâncias. Segundo Santos (2015), o $RMSE$ é um estimador robusto, pois pode ser utilizado para descrever um conjunto de dados independente da sua distribuição de probabilidade. No entanto, esse estimador é altamente influenciado pela presença de *outliers* na amostra (LI et al., 2005).

Num contexto estatístico, é sempre preferível que uma estimativa seja apresentada conjuntamente com seu grau de confiabilidade. Isso pode ser realizado através da construção de intervalos de confiança (IC), quase sempre, baseados na distribuição amostral do estimador pontual (MORETTIN & BUSSAB, 2004).

Por exemplo, suponha que as discrepâncias resultem em uma distribuição gaussiana, são livres de dados anormais e independentes espacialmente. Diante disso, sabe-se da teoria dos erros que 68,3% dos dados avaliados estão no intervalo $\mu \pm \sigma$; 95% estão no intervalo $\mu \pm 1{,}96\sigma$ e 99,7% dos dados estão no intervalo $\mu \pm 3\sigma$ (MOOD, 1913; MOOD et al., 1974).

No caso das normas de levantamentos hidrográficos, tal como a NORMAM-25 (DHN, 2014), a incerteza vertical deve sempre ser estimada a um nível de confiança de 95%, ou seja, o $RMSE$ deve ser fornecido conjuntamente com seu $IC_{95\%}$. É comum na literatura encontrar o IC utilizado para o estimador $RMSE$ definido com base na constante 1,96, ou seja, $IC_{95\%} = [-1{,}96 \cdot RMSE, +1{,}96 \cdot RMSE]$ (GREENWALT & SCHULTZ, 1962; FGDC, 1998; SUSAN & WELLS, 2000; EEG, 2010; SEKELLICK & BANKS, 2010). No entanto, mesmo na presença de normalidade, o $IC_{95\%}$ apresentado acima não remete a medida de $RMSE$.

Neste trabalho, o intervalo de confiança do $RMSE$, para uma amostra normal e independente, será estimado a partir da distribuição qui-quadrado (χ^2) (STEIGER & LIND, 1980; COOPER, 1987).

Se $X_i \sim N(\mu, \sigma^2)$, então: $Y = \frac{RMSE^2 \cdot n}{\sigma^2} \sim \chi^2_{n-1}$. Após algumas manipulações matemáticas, têm-se que (Equação 2):

$$P\left(\sqrt{\frac{n}{\chi^2_{n-1,\,\alpha/2}}} \cdot RMSE \leq \sigma \leq \sqrt{\frac{n}{\chi^2_{n-1,\,1-\alpha/2}}} \cdot RMSE\right) = 1 - \alpha \qquad (2)$$

em que α é o nível de significância. Assim, pode-se estimar o intervalo de confiança para o $RMSE$ através da Equação 3.

$$IC_{95\%}(RMSE) = \left[\sqrt{\frac{n}{\chi^2_{n-1;\,0{,}025}}} \cdot RMSE,\ \sqrt{\frac{n}{\chi^2_{n-1\,;\,0{,}975}}} \cdot RMSE\right] \qquad (3)$$

Segundo Monico et al. (2009), o $RMSE$ reflete a acurácia teórica da amostra, que, para todos os casos, engloba efeitos aleatórios ($\sigma_{aleatório}$) e sistemáticos ($\sigma_{sistemático}$). Sendo assim, uma avaliação estatística mais fidedigna e conveniente seria em termos de dois parâmetros independentes, possibilitando que haja

discriminação entre efeitos aleatórios e sistemáticos, tal como proposto na Equação 4, em que Φ é um estimador para a incerteza vertical:

$$\Phi = \sqrt{\sigma^2_{aleatório} + \sigma^2_{sistemático}} \quad (4)$$

A Equação 4 indica que os dados estão distribuídos de forma aleatória e que, embora indesejáveis, ainda possuem efeitos sistemáticos. De acordo com USACE (2002) e Ferreira et al. (2015) em um levantamento hidrográfico os efeitos aleatórios variam com a profundidade observada e podem ser quantificados através do desvio padrão populacional (σ) das discrepâncias entre as profundidades, enquanto que os efeitos sistemáticos, parcela que não depende da profundidade, podem ser quantificados através da média populacional (μ) das discrepâncias entre as profundidades. Höhle & Höhle (2009) também sugerem o uso da média e do desvio padrão para quantificar, respectivamente, efeitos sistemáticos e aleatórios em amostras normais. Dado o exposto, pode-se deduzir que, teoricamente, as Equações (1) e (4) fornecem os mesmos resultados.

A Figura 3 ilustra, mesmo que de forma subjetiva, os conceitos supracitados.

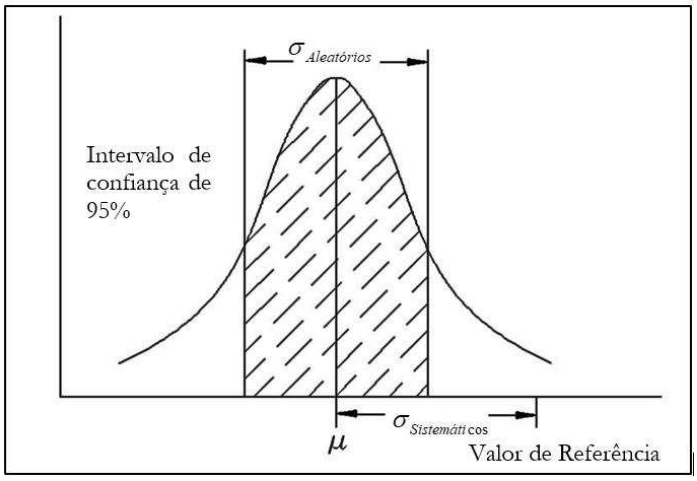

Figura 3 – Curva de dispersão padronizada para observações de profundidade.

Em uma amostra independente e normal contaminada por efeitos sistemáticos a média *(μ)* será diferente de zero, ou seja, a curva gaussiana estará deslocada, indicando um viés nos dados (tendência). O valor deste deslocamento fornece uma quantificação do efeito sistemático da amostra. Já o efeito aleatório, de acordo com a Figura 3, poderá ser quantificado pelo desvio padrão amostral. Evidencia-se que certo grau de aleatoriedade nas amostras é mais tolerável que um viés.

Para calcular o $IC_{95\%}(\Phi)$, sugere-se o emprego da técnica *Bootstrap,* dado que a distribuição de probabilidade deste estimador não pôde ser teoricamente definida e,

assim, uma abordagem paramétrica torna-se estatisticamente incoerente. Dado o exposto, neste trabalho pretende-se avaliar os parâmetros estimados pelas Equações (1) e (4), bem como os respectivos $IC_{95\%}$.

O *Bootstrap* consiste numa técnica de reamostragem que permite aproximar a distribuição de uma função das observações a partir da distribuição empírica dos dados baseado em uma amostra de tamanho finito (EFRON, 1979; EFRON & TIBSHIRANI, 1993). O método foi proposto por Efron (1979), inicialmente, como uma ferramenta para estimar o erro padrão da estimativa de um parâmetro. Atualmente esta metodologia possui uma série de aplicações, destacando-se, a obtenção de intervalos de confiança (HESTERBERG et al., 2003; FRANCO & REISEN, 2007).

Técnicas de reamostragem *(Bootstrap, Jackknife, etc.)*, descartam a distribuição amostral assumida de uma estatística e calculam uma distribuição empírica, ou seja, a real distribuição da estatística ao longo de centenas ou milhares de amostras. Com isso, não é necessário assumir que a distribuição de um determinado estimador segue normalidade (LEE & LAI, 2009).

Em outras palavras, a ideia básica do *Bootstrap* é efetuar uma reamostragem do conjunto de dados original, a partir das quais calculam-se a estatística do problema, que no caso deste trabalho, consiste na incerteza vertical. Essa metodologia não altera nenhum valor da amostra original, ela apenas reamostra de forma aleatória e com reposição, gerando réplicas de mesmo tamanho da amostra original. Evidencie-se que, frente as demais técnicas de reamostragem, o *Bootstrap* destaca-se por obter sua amostra via amostragem com reposição.

O *Bootstrap* pode ser implementado tanto de forma não paramétrica quanto de forma paramétrica. No primeiro caso, a amostragem é feita com reposição da amostra original. Nesse contexto, supõe-se que as observações são obtidas da função de distribuição empírica, que designa uma massa de probabilidade igual a $1/n$ para cada ponto amostral. Já no caso paramétrico, a amostragem é feita, com reposição, a partir da distribuição ajustada às observações amostrais (CARPENTER & BITHELL, 2000). Nesse caso, é nítido que a técnica proposta por Efron (1979) apresenta-se como uma solução para os casos em que a dedução dos intervalos de confiança mostra-se impossível ou demasiadamente complexo.

Através de testes iniciais realizados com a técnica *Bootstrap*, percebeu-se que as amostras geradas podem apresentar-se tanto simétricas quanto assimétricas. Diante

disso, na MAIB adotam-se duas técnicas *Bootstrap* para definição dos limites de confiança, o *Bootstrap-t*, paramétrico, e o método *BCa*, em uma forma não paramétrica (EFRON & TIBSHIRANI, 1993; CARPENTER & BITHELL, 2000).

O *Bootstrap-t* consiste em uma técnica paramétrica que assume que a distribuição *Bootstrap* é normalmente distribuída e, assim, o intervalo de confiança é calculado com base na distribuição *t-Student.* Desse modo, o IC dado pelo *Bootstrap-t* somente gera resultados confiáveis quando a distribuição da estatística na distribuição *Bootstrap* for aproximadamente normal e a estatística pouco viciada. Diz-se que a estatística é não viciada, quando a distribuição *Bootstrap* da incerteza vertical está centrada no valor esperado da incerteza vertical calculada através do conjunto de dados original (EFRON & TIBSHIRANI, 1993; HESTERBERG et al., 2003). O $IC_{bootstrap-t}(\Phi)$ é dado pela seguinte equação:

$$IC_{bootstrap-t}(\Phi) = [estatística \pm t \cdot SD_{boot}\,] \tag{5}$$

onde t é encontrado utilizando-se $n-1$ graus de liberdade e SD_{boot} é o desvio padrão *Bootstrap,* calculado através da Equação 6.

$$SD_{boot} = \sqrt{\frac{1}{\dot{n}} \cdot \sum \left(\theta_i - \frac{1}{\dot{n}} \cdot \sum \theta_i\right)^2} \tag{6}$$

em que $\dot{n}$ corresponde ao número de reamostras e θ_i é o valor da estatística calculada para cada reamostra (EFRON & TIBSHIRANI, 1993; HESTERBERG et al., 2003).

O método *BCa (Biased Corrected Accelerated)* é uma metodologia que permite definir intervalos de confiança quando os dados apresentarem assimetria muito forte. Nesses casos, os extremos dos intervalos, isto é, os percentis da distribuição *Bootstrap,* são ajustados por meio de uma constante de aceleração para corrigir o vício e assimetria da distribuição. A obtenção da constante de aceleração envolve estimativas não triviais, que necessitam de um esforço computacional maior. Por exemplo, para encontrar um intervalo de confiança de 95%, pode-se, tradicionalmente, calcular os percentis 2,5% e 97,5%, ou seja, os quantis 0,025 e 0,075 (Equação 7).

$$IC_{95\%}^{BCa}(\Phi) = \left[q_{0,025}, q_{0,975}\,\right] \tag{7}$$

Porém, nessa técnica, os intervalos são ajustados objetivando corrigir a assimetria e o vício (EFRON & TIBSHIRANI, 1993; HESTERBERG et al., 2003). Nesse método, *a priori,* não se presume nenhuma distribuição de probabilidade, logo, trata-se de uma técnica não paramétrica.

Para estimar o intervalo de confiança através do *Bootstrap* é necessário definir o número de replicações. Quando o número de reamostras tende ao infinito, isto é, $\dot{n} \rightarrow$

∞, as estimativas *Bootstrap* assemelham-se às estimativas de máxima verossimilhança. Segundo Efron & Tibshirani (1993), para se obter boas estimativas dos limites de confiança são necessárias mais de 500 replicações. Höhle & Höhle (2009), recomendam o uso de 1.000 replicações. Neste trabalho foram utilizadas 5.000 replicações para definir os intervalos de confiança.

2.2. Amostra independente e não-normal

Nos casos em que a independência espacial for constatada, porém a distribuição amostral não apresentar normalidade, sugere-se a seguir duas abordagens diferentes para estimar a incerteza amostral, embora, as equações apresentadas na seção anterior, com IC estimados por técnicas não paramétricas, tal como *Bootstrap*, possam ser utilizadas. Contudo, sabe-se que técnicas paramétricas possuem maior poder estatístico, sendo, portanto, sempre recomendadas (MOOD, 1913). Além do mais, a análise de não-normalidade deve ser refinada, visto que tal constatação pode ter sido causada pela presença de *outliers*, *a priori*, não detectados na análise exploratória e, nesses casos, medidas robustas são mais eficientes.

A primeira abordagem, consiste em aplicar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal (MORETTIN & BUSSAB, 2004). Neste trabalho, seguindo recomendações de Santos (2015), será aplicada uma transformação com base no seguinte teorema: para amostras aleatórias simples retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\bar{X}$ aproxima-se, para um grande conjunto de dados, de uma distribuição normal, com mesma média e variância proporcionalmente menor. Este teorema é conhecido como TCL (Teorema Central do Limite). A demonstração deste teorema, assim como exemplos teóricos, pode ser consultada em Mood (1913), Mood et al. (1974) e Morettin & Bussab (2004).

Assim, propõe-se agrupar alguns pontos próximos e calcular as médias das discrepâncias dos pontos. O resultado será uma nova amostra, chamada aqui de amostra TCL, contendo as médias de discrepâncias para cada agrupamento. Vale ressaltar que essa abordagem somente terá validade teórica e, consequentemente, prática, se a amostra original possuir um grande número de discrepâncias, o que conduzirá a um grande número de agrupamentos, isto é, uma amostra TCL com $Tamanho\ Amostral \rightarrow \infty$. Conforme visto, de acordo com o Teorema Central do

Limite, a distribuição das médias tende a uma distribuição normal, com a mesma média do conjunto original e com a variância dividida pelo tamanho amostral dos agrupamentos (n) (MORETTIN & BUSSAB, 2004; SANTOS, 2015).

A rotina desenvolvida neste trabalho aplica um processo de clusterização, baseado no algoritmo *k*-medoids, para agrupar as discrepâncias próximas e, então, calcular a média, gerando um novo conjunto de dados. A independência desse novo conjunto é garantida pelo seguinte teorema: Se $X_1, \cdots, X_k$ são variáveis aleatórias independentes e $g_1(\cdot), \cdots, g_s(\cdot)$ são s funções tal que $Y_j = g_j(X_j), j = 1, \cdots, k$ são variáveis aleatórias, então, $Y_1, \cdots, Y_s$ são independentes (MOOD, 1913; MOOD et al., 1974).

O *k*-medoids é uma técnica de clusterização mais robusta que o *k-médias*, pois não utiliza a média como centro do grupo *(cluster)*, mas sim, uma observação do conjunto original. Essa observação é chamada de objeto representativo ou *medoid*, que se localiza sempre mais ao centro do *cluster.* A dissimilaridade é quantificada pela distância "euclidiana" ou "manhattan" (REYNOLDS et al., 1992).

Conforme pressupostos do TCL, o número de *clusters*, coincidente com tamanho da amostra TCL, deve ser o maior possível. Neste trabalho, a quantidade de grupos é definida de tal forma que nenhum *cluster* contenha menos que 4 pontos, conforme sugerido por Santos (2015), ou seja, $n \geq 4$. Sabe-se que, teoricamente, o tamanho amostral dos agrupamentos (n) deve ser constante, todavia, dado a natureza dos dados batimétricos, é evidente que este pressuposto, pelo menos para levantamentos monofeixe, não é sempre atendido.

Após a obtenção do novo conjunto de dados, sugere-se que o pressuposto de normalidade seja novamente testado, uma vez que o teorema aplicado é valido apenas quando o tamanho da amostra TCL for razoavelmente grande, fato pouco comum em levantamentos monofeixe. Caso confirmado, determina-se a incerteza do levantamento batimétrico através de uma modificação sutil da Equação 4 (seção 2.1). Isto é, conforme citado, o TCL, teoricamente, transforma a amostra original, numa amostra normal com mesma média e variância σ^2/n. Assim sendo, pode-se quantificar a incerteza vertical através da Equação 8:

$$\Phi_{TCL} = \sqrt{\mu_{TCL}^2 + (\sigma_{TCL}^2 \cdot n)} \tag{8}$$

em que μ_{TCL} e σ_{TCL}, são, respectivamente, a média e o desvio padrão da amostra TCL e n é o tamanho amostral dos agrupamentos. Os intervalos de confiança de 95% são definidos conforme as técnicas descritas na seção anterior.

A segunda abordagem poderá ser utilizada naquelas ocasiões em que a amostra TCL não seguir normalidade. Nesse caso, sugere-se a aplicação de estimadores robustos e estatísticas não-paramétricas para estimar a incerteza vertical. Conforme afirma Morettin & Bussab (2004), estatísticas paramétricas são eficientes apenas quando se conhece a distribuição de probabilidade da amostra, uma vez que, esse ramo da estatística supõe que os dados são provenientes de um tipo de distribuição de probabilidade e faz inferências sobre os parâmetros da distribuição. Caso esta hipótese esteja incorreta, os métodos paramétricos poderão tornar-se incorretos e, por este motivo, são considerados menos robustos. Por outro lado, se a distribuição dos dados é conhecida, os métodos paramétricos produzem estimativas mais confiáveis e, por este motivo, possuem maior poder estatístico.

Neste trabalho, seguindo linhas propostas por Höhle & Höhle (2009), será adotado a mediana ($Q2$), como estimador para o possível efeito sistemático presente nos dados, e o $NMAD$ (Desvio Absoluto da Mediana Normalizado), para estimar o efeito aleatório. Assim, a incerteza vertical amostral pode ser calculada através da Equação 9, adaptada e proposta neste trabalho:

$$\Phi_{Robusta} = \sqrt{(Q2)^2 + (NMAD)^2} \tag{9}$$

A mediana é uma medida de tendência central mais adequada quando comparada à média amostral, pois indica exatamente o valor central de um conjunto de dados quando organizados em ordem crescente ou decrescente, além de ser pouco afetada por *outliers* (MOOD, 1913). Segundo Höhle & Höhle (2009), a mediana também é uma medida de localização mais apropriada para dados assimétricos.

O $NMAD$ corresponde a $1{,}4826 \cdot mediana\{|x_i - Q2)|\}$, em que x_i corresponde a discrepância i. É considerado uma estimativa para a dispersão dos dados mais resistente a *outliers* que o tradicional desvio padrão. Nos casos em que a distribuição normal for verificada, a $Q2$ e o $NMAD$, serão equivalentes, respectivamente, à média e ao desvio padrão (HOAGLIN et al., 1983; HÖHLE & HÖHLE, 2009). Sendo assim, o estimador proposto na Equação 9, torna-se mais robusto que aqueles apresentados até o momento.

Calculada a incerteza, os intervalos de confiança são construídos através da técnica *Bootstrap*, conforme esclarecido na seção anterior.

2.3. Amostra dependente

Até o momento, todas as estimativas assumiram que as discrepâncias são variáveis aleatórias independentes e identicamente distribuídas. Todavia, quando os dados são georreferenciados no espaço, é comum a ocorrência de dependência espacial (VIEIRA, 2000). Assim, torna-se necessário a utilização de metodologias que levem em conta a autocorrelação espacial, uma vez que, a desconsideração desta condição compromete as análises.

Para estimar a incerteza amostral, pode-se recorrer as equações apresentadas na seção 2.1 (Equações 1 e 4), uma vez que a presença de dependência espacial não afeta a estimativa pontual. Caso a análise gráfica (histogramas, *Q-Q Plot, etc.*) dos dados revele uma distribuição demasiadamente distorcida, sugere-se, utilizar o estimador robusto proposto na seção 2.2 (Equação 9).

Entretanto, a independência estatística é um pressuposto assumido pela maioria dos testes de aderência, tal como os testes de normalidade, o que impede que tais hipóteses sejam testadas e assim, os intervalos de confiança paramétricos não podem ser construídos (MORETTIN & BUSSAB, 2004). Uma alternativa seria a aplicação da metodologia *Bootstrap*, porém, na sua forma padrão, essa técnica também presume que a variável em estudo é independente e identicamente distribuída. Assim, a sua aplicação a dados autocorrelacionados, traduz-se em intervalos de confiança bastante estreitos e, portanto, inconsistentes (EFRON, 1979; EFRON & TIBSHIRANI, 1993; HÖHLE & HÖHLE, 2009).

Santos (2015) sugere o uso da Geoestatística para tratar dados dependentes espacialmente. Visto que na presença de autocorrelação, a variabilidade espacial da maioria dos fenômenos naturais não pode ser representada por simples funções matemáticas (FERREIRA et al., 2013). O semivariograma é a ferramenta básica de suporte às técnicas geoestatísticas, pois permite modelar a dependência espacial entre as amostras. Todavia, a modelagem do semivariograma deve ser realizada de forma minuciosa, visto que erros na modelagem comprometem as inferências estatísticas, como demonstrado por Ferreira et al. (2013).

Embora a técnica proposta por Santos (2015) possa ser utilizada, neste trabalho recomenda-se efetuar uma modificação do tradicional procedimento *Bootstrap* de tal forma que a dependência espacial seja levada em conta. Sendo assim, será implementada e testada uma variação não paramétrica do *Bootstrap*, conhecida como *Block Bootstrap*, conforme descrito, por exemplo, em Lahiri (1999), Lahiri (2003), Lee & Lai (2009) e Kreiss & Paparoditis (2011).

Segundo Lahiri (1999), nos últimos anos diferentes métodos do *Block Bootstrap* surgiram na literatura, dentre eles, destacam-se o *Moving Block Bootstrap*, o *Nonoverlaping Block Bootstrap,* o *Circular Block Bootstrap* e o *Stationary Bootstrap.* O mesmo autor afirma que o uso de blocos sobrepostos, como por exemplo o *Moving Block Bootstrap*, resulta em melhores estimativas quando comparado as técnicas que utilizam blocos sem sobreposição, tal como o *Nonoverlaping Block Bootstrap,* e que a escolha aleatória dos comprimentos dos blocos é sempre preferível.

Neste estudo, o *Block Bootstrap* foi implementado em ambiente R, assim como todo o restante da metodologia. No algoritmo desenvolvido o usuário precisa fornecer o tamanho da diagonal do bloco e o número de replicações *Bootstrap.* Propõe-se que a diagonal seja equivalente a distância dentro da qual os dados encontram-se correlacionados, isto é, o alcance. O número de replicações, conforme já discutido, deve ser maior que 500.

Posteriormente, o algoritmo da MAIB subdivide a área de estudo em blocos, selecionando, a cada iteração, 1 bloco e 1 ponto deste bloco aleatoriamente. Essa etapa é realizada *n* vezes, em que *n* corresponde ao número de dados da amostra. O resultado é um novo conjunto de dados com o mesmo tamanho do conjunto original. Evidencia-se que o mesmo bloco pode ser selecionado várias vezes e, consequentemente, o mesmo ponto dentro do bloco. Essa seleção é realizada, por exemplo, 500 vezes (número de replicações *Bootstrap)*, gerando, assim, 500 novos conjuntos de dados. De posse dessas amostras, calculam-se a estatística de interesse para cada conjunto gerando a amostra *Bootstrap.* Finalmente, estima-se o $IC_{95\%}$ a partir dos quantis da distribuição da amostra *Bootstrap*, isto é, $IC_{95\%} = \left[q_{0,025}, q_{0,975}\right]$.

A Tabela 1 resume os estimadores discutidos.

Tabela 1 – Resumos das estatísticas empregadas pela MAIB.

<table>
<tr><th colspan="2">Categoria</th><th>Estimativa Pontual da Incerteza</th><th>Intervalo de Confiança de 95%</th></tr>
<tr><td colspan="2" rowspan="3">Amostra Independente e Normal</td><td>$RMSE$</td><td>$IC_{95\%}(RMSE)$</td></tr>
<tr><td rowspan="2">Φ</td><td>$IC_{bootstrap-t}(\Phi)$</td></tr>
<tr><td>$IC_{95\%}^{BCa}(\Phi)$</td></tr>
<tr><td rowspan="4">Amostra Independente e Não-Normal</td><td rowspan="2">Aplicação do TCL</td><td rowspan="2">Φ_{TCL}</td><td>$IC_{bootstrap-t}(\Phi)$</td></tr>
<tr><td>$IC_{95\%}^{BCa}(\Phi)$</td></tr>
<tr><td rowspan="2">Abordagem Robusta</td><td rowspan="2">$\Phi_{Robusta}$</td><td>$IC_{bootstrap-t}(\Phi)$</td></tr>
<tr><td>$IC_{95\%}^{BCa}(\Phi)$</td></tr>
<tr><td rowspan="2">Amostra Dependente</td><td>Abordagem convencional (seção 2.1)</td><td>$RMSE$ ou Φ</td><td rowspan="2">Aplicação do Block Bootstrap $IC_{95\%} = [q_{0,025}, q_{0,975}]$</td></tr>
<tr><td>Abordagem Robusta (seção 2.2)</td><td>$\Phi_{Robusta}$</td></tr>
</table>

2.4. Classificação do Levantamento Hidrográfico

Computada a incerteza vertical amostral, bem como o intervalo de confiança, o levantamento hidrográfico pode ser classificado de acordo com a Publicação Especial nº 44 (S-44).

Nesse sentido, o intervalo de incerteza vertical máxima permitida, ao nível de confiança de 95%, é calculada pela seguinte equação:

$$IVT_{max} = \pm\sqrt{a^2 + (b \cdot P)^2} \tag{10}$$

em que as constantes a e b são dadas pela Tabela 2 e variam de acordo com a ordem dos levantamentos hidrográficos. O termo P é a profundidade, assim, verifica-se que cada profundidade possuirá uma estimativa de incerteza. Diante disso, é recomendável a adoção de uma profundidade média.

Pode-se entender o termo a como uma parcela da incerteza que não varia com a profundidade, enquanto $b \cdot P$ representa a parcela que varia com profundidade. Em

termos estatísticos, essas parcelas representam, respectivamente, os efeitos sistemáticos e aleatórios da profundidade reduzida.

Tabela 2 – Resumo dos padrões mínimos para Levantamentos Hidrográficos.

Ordem	Especial	1a	1b	2
IVT máxima permitida. Nível de confiança de 95%	a = 0,25 m b= 0,0075	a = 0,50 m b= 0,013	a = 0,50 m b= 0,013	a = 1,00 m b= 0,023

Fonte: Adaptado de IHO (2008).

Deve-se atentar que, para levantamentos batimétricos de ordem Especial e 1a, a norma prevê a adoção de sistemas de sondagem por faixa. Outrem, uma série de outros fatores condicionam a classificação do levantamento hidrográfico.

3. EXPERIMENTOS E RESULTADOS

Os dados que serviram de base para este estudo foram obtidos de Carmo (2014). A coleta dos dados ocorreu em setembro de 2013 através de um levantamento batimétrico realizado nas lagoas do município de Capitólio – centro oeste do Estado de Minas Gerais.

A Figura 4 ilustra a área de estudo. Para o bom andamento dos trabalhos, a área foi dividida em Reservatório 1 e Reservatório 2.

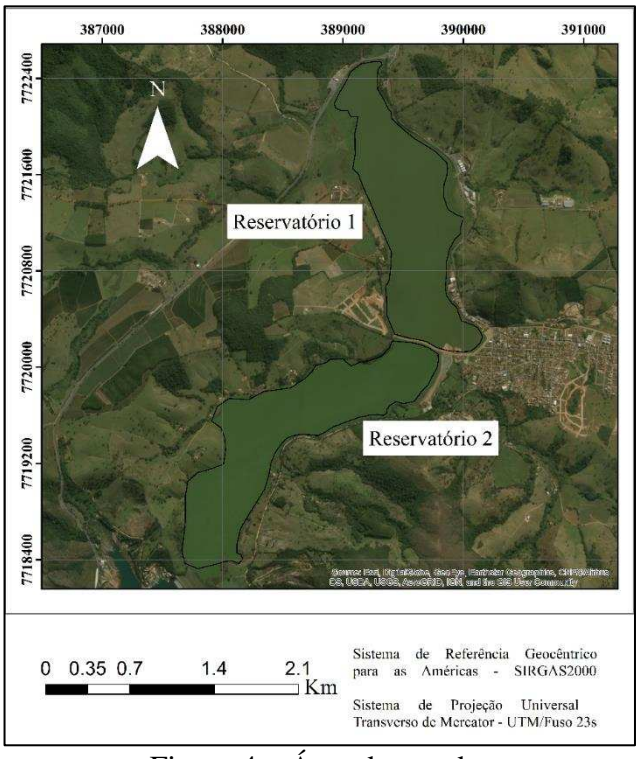

Figura 4 – Área de estudo.

Os levantamentos foram conduzidos a bordo da lancha EAM-01 do Departamento de Engenharia Civil da Universidade Federal de Viçosa, com o uso de um sistema de batimetria monofeixe (SBES) operando um transdutor de dupla frequência (33/210kHz). Para o posicionamento geodésico foi utilizado um par de receptores *GNSS* no modo *RTK (Real Time Kinematic)*. Segundo Ferreira et al. (2016a), esse sistema de batimetria monofeixe, teoricamente, é capaz de alcançar uma incerteza vertical abaixo de 30 centímetros para áreas com até 20 metros de profundidade. No entanto, dado as características do levantamento e da área de estudo, a incerteza vertical amostral deverá estabelecer-se em torno de $^1/_3$ da incerteza vertical total propagada.

Na fase de planejamento da batimetria estipulou-se um espaçamento de 20 metros para as linhas regulares de sondagem dispostas de modo perpendicular às linhas isobatimétricas da área e 100 metros para as linhas de verificação. Esse levantamento não adotou, a priori, espaçamentos baseados em normativas, por ele ter sido utilizado também em pesquisas com esta finalidade.

Entretanto, deve-se atentar que a escolha do espaçamento de linhas (bem como o seu direcionamento) em um levantamento hidrográfico é uma fase crítica, pois possui relação direta com o custo do levantamento, tempo de execução e resolução do produto final. Diversos autores sugerem formas de planejar estas linhas, como EAKIN (1939); CARVALHO et al. (2000); WMO (2003); FERRARI (2006); IHO (2008); INSTITUTO HIDROGRÁFICO (2009); LINZ (2010); NOAA (2011); ANA (2013); USACE (2013) e DHN (2014). Especificações para planejamento de levantamentos hidrográficos destinados a cartografia náutica são, em sua maioria, baseados na publicação S-44, 5ª edição (IHO, 2008).

Todavia, em um levantamento com um SBES, assume-se que o relevo submerso é uniforme entre duas linhas paralelas, nesse sentido, o espaçamento entre as linhas regulares de sondagem deve ser estabelecido considerando o objetivo do levantamento, atentando-se que a densidade de pontos é função não somente do espaçamento de linhas, mas também da taxa de medição do SBES e da velocidade da embarcação. Em rios e reservatórios, devido à dinâmica de fundo, pode-se adotar um espaçamento de linhas baseado na escala final da planta batimétrica, tal como sugerido por alguns autores supracitados. Já para as linhas de verificação, o espaçamento pode ser determinado baseado no número de intersecções requeridas para uma posterior

análise estatística do levantamento. USACE (2013) sugere um mínimo de 100 intersecções. Porém, na prática tem-se adotado, conforme recomendação de IHO (2008), um espaçamento não superior a 15 vezes o adotado para as linhas regulares de sondagem.

A Figura 5 apresenta o planejamento das linhas regulares de sondagem e de verificação para a área de estudo.

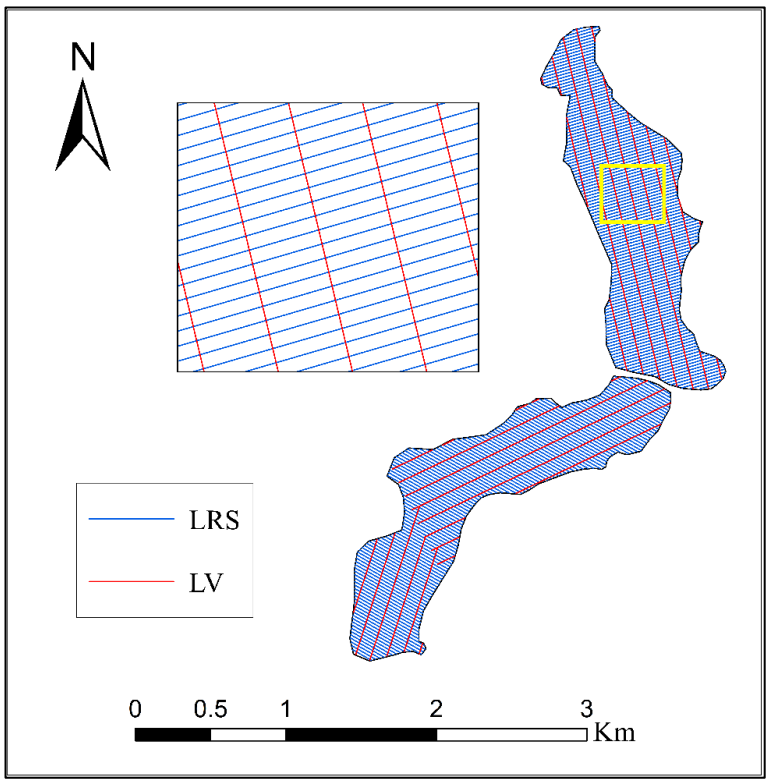

Figura 5 – Linhas Regulares de Sondagem (LRS) e Linhas de Verificação (LV).

A coleta, processamento e análise dos dados batimétricos seguiram as recomendações de IHO (2008), DHN (2014) e Ferreira et al. (2015). Durante o processamento foram eliminados todos os dados duvidosos, incluindo os efeitos de picos graves *(spikes)*, erros de posicionamento *(tops)* e interferências.

Feito isso, foi utilizada a ferramenta *Cross Check Statistcs* do *software Hypack 2012* (HYPACK, 2012) para criação dos arquivos de discrepâncias entre as profundidades provenientes das linhas de sondagem e de verificação. É sabido que a coleta de pontos homólogos em uma sondagem batimétrica é algo impraticável. Sendo assim, alternativamente, estabelece-se um raio de busca nas intersecções entre os perfis longitudinais e transversais, selecionando, através da menor distância euclidiana, duas profundidades, respectivamente, obtidas através das linhas regulares e de verificação. Estas profundidades são, então, consideradas homólogas. A escolha do raio é algo

subjetivo e irá variar, principalmente, com o tipo de relevo da área sondada, com o espaçamento entre linhas e com a qualidade da navegação em termos de coerência com as linhas planejadas. Neste estudo, após análises e considerações, adotou-se um raio de 3 metros.

Tendo em vista alcançar os objetivos propostos procedeu-se com a análise estatística das discrepâncias. Essa etapa foi executada através do algoritmo desenvolvido e implementado no *software* estatístico R (R Core Team, 2017). Toda a análise foi realizada sobre as profundidades sondadas na frequência de 210 kHz. A Figura a seguir ilustra as interseções obtidas para cada um dos reservatórios. Note que as interseções planejadas nem sempre são executadas, uma vez que podem ocorrer falhas de cobertura no levantamento, má navegação, exclusão de pontos batimétricos durante o processamento, dentre outros. Obviamente que o número interseções executadas tende a crescer com o valor do raio escolhido para obtenção dos pontos homólogos.

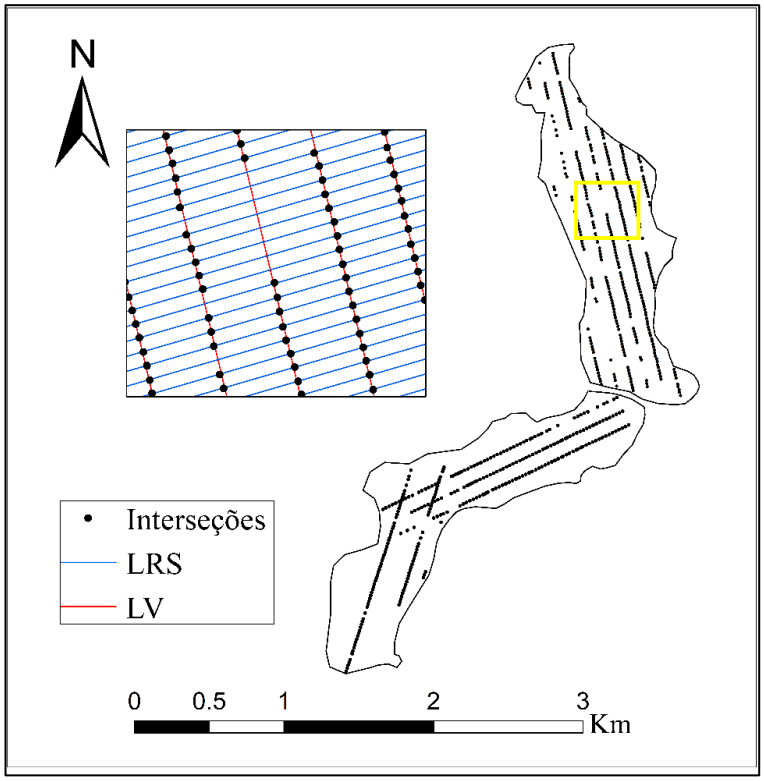

Figura 6 – Interseções previstas e executas para o Reservatório 1 e Reservatório 2.

3.1. Reservatório 1

Após o processamento dos dados referentes ao Reservatório 1, foram geradas 386 discrepâncias de 881 interseções planejadas. A primeira etapa, conforme exposto

na seção 2, consiste na importação do arquivo XYZdz, com posterior realização da análise exploratória (Tabela 3).

Tabela 3 – Estatística descritiva da área de estudo.

Número de Discrepâncias	386
Média (m)	0,091
Mínimo (m)	-1,773
Máximo (m)	2,150
Variância (m²)	0,1363
Coeficiente de Curtose	15,790
Coeficiente de Assimetria	1,150
Distância Mínima (m)	0,014
Distância Máxima (m)	2445,570

Analisando a Tabela 3, percebe-se que os dados apresentam uma variabilidade alta, considerando o valor da variância (WARRICK & NIELSEN, 1980). Os coeficientes de assimetria e curtose que quantificam, respectivamente, o desvio da distribuição das discrepâncias em relação a uma distribuição simétrica e o grau de achatamento da distribuição, indicam uma distribuição basicamente simétrica e leptocúrtica. Diante disso, conclui-se, inicialmente, que a amostra tende a possuir uma distribuição normal e, dada a alta variabilidade, está eivada de dados discrepantes, isto é, *outliers.*

A Figura 7 apresenta gráficos que auxiliam na análise exploratória e, assim sendo, são construídos e gerados pelo algoritmo desenvolvido.

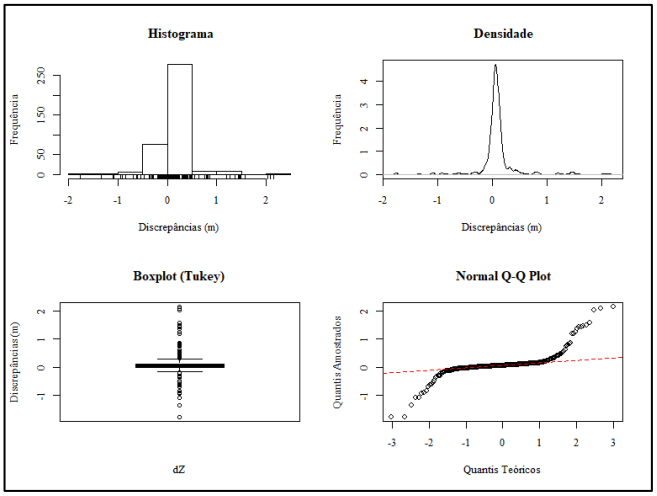

Figura 7 – Análise gráfica exploratória.

O histograma das discrepâncias, a curva de densidade e o gráfico *Q-Q Plot* são ferramentas capazes de fornecer uma impressão acerca da normalidade dos dados. Dentre os três, o *Q-Q Plot* possui características mais robustas, uma vez que permite checar a adequação da distribuição de frequência dos dados (empírica/real) à uma distribuição normal. Se a distribuição empírica é normal, o gráfico será apresentado como uma linha reta (HÖHLE & HÖHLE, 2009). Após a análise gráfica, em desacordo com a conclusão anterior, verifica-se a não normalidade dos dados que deverá posteriormente, caso seja constatada a independência espacial, ser confirmada por testes de normalidade univariada.

Para detecção de *outliers* a metodologia sugere a aplicação de 3 métodos, a saber: *Boxplot de Tukey*, o *Boxplot Ajustado* e o *Z-Score Modificado*. Através da aplicação destas técnicas foram detectados 55 *outliers* pelos métodos *Boxplot de Tukey* e *Boxplot Ajustado* e 52 pelo *Z-Score Modificado*. Ressalta-se que esta é uma fase importante e, por esse motivo, deve ser realizada com muita atenção. Diante disso, os dados processados, correspondentes as discrepâncias localizadas como *outliers,* foram reavaliados, buscando-se falhas no processamento durante as fases de eliminação de *spikes* e *tops*, redução de maré e aplicação da velocidade de propagação do som às sondagens. Feito isso, pôde-se concluir que para a área de estudo os métodos *Boxplot* mostraram-se, de certo modo, eficientes. Os resultados idênticos obtidos pelas técnicas *Boxplot* reafirmam a simetria da base de dados analisada.

Após a eliminação dos *outliers* a amostra apresentou graficamente, um maior grau de normalidade, conforme pode ser visto na Figura 8. Todavia, a comprovação sobre a normalidade do conjunto de discrepâncias apenas pode ser obtida através da aplicação de testes de normalidade, que por sua vez, supõem independência amostral.

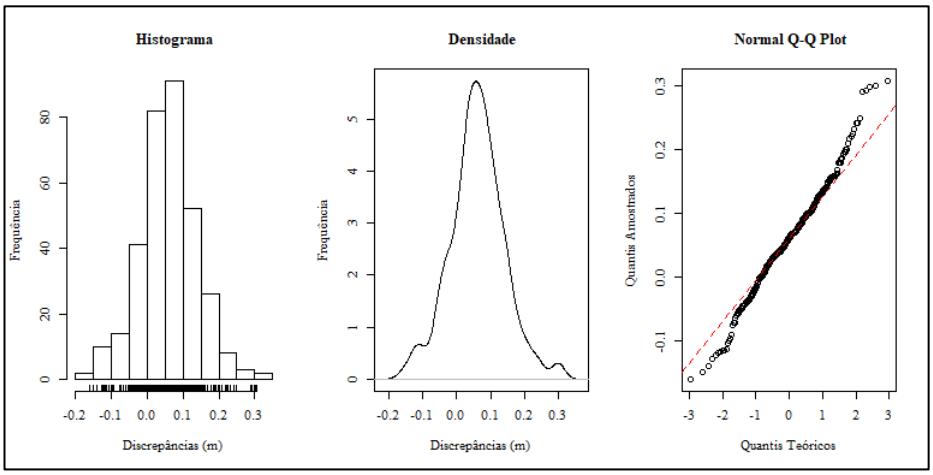

Figura 8 – Análise gráfica exploratória após a eliminação dos *outliers*.

A próxima etapa, consiste no cálculo da incerteza amostral que pode ser realizada, conforme exposto na seção 2, pelos estimadores: $RMSE$, Φ e $\Phi_{Robusta}$. A Tabela 4 apresenta as estimativas para os dados sem *outliers* e com *outliers.*

Tabela 4 – Estimativa pontual da incerteza vertical amostral.

Estimador	**Dados com *outliers***	**Dados sem *outliers***
$RMSE\ (m)$	0,380	0,100
$\Phi\ (m)$	0,380	0,100
$\Phi_{Robusta}\ (m)$	0,109	0,091

Os estimadores $RMSE$ e Φ, como esperado, apresentaram-se coerentes. Porém, uma problemática notável destes estimadores é a alta influência sofrida pela presença de *outliers.* Em contrapartida, o estimador $\Phi_{Robusta}$ mostrou-se bastante eficiente, visto que a presença de *outliers* na base de dados interferiu minimamente na estimativa pontual da incerteza. Desse modo, confirma-se o alto grau de robustez da estatística $\Phi_{Robusta}$, no que concerne o tratamento de dados, sabidamente ou duvidosamente, eivado de *outliers*.

Comparando as incertezas obtidas através dos dois diferentes conjuntos de dados, observa-se que a computação da incerteza vertical sem uma análise prévia da presença de *outliers*, subestimaria a qualidade do levantamento. Tal fato pode colocar em dúvida a classificação da sondagem batimétrica de acordo com a norma utilizada e/ou requerida para o projeto em questão, levando, inclusive, a uma rejeição da batimetria pelo executor ou contratante.

Em concordância com a seção 2, num contexto estatístico é sempre preferível que o estimador seja sempre apresentado conjuntamente com seu grau de incerteza, isto é, o intervalo de confiança. Comumente adotam-se níveis de significância de 5%, sendo necessário, nesses casos, estimar a incerteza vertical associada ao $IC_{95\%}$. De posse destas quantidades, pode verificar se o levantamento hidrográfico atende os requisitos de incerteza previstos na S-44 (IHO, 2008; DHN, 2014).

Assim, seguindo o fluxograma da metodologia proposta, passa-se a análise de independência dos dados a partir da construção do semivariograma. Nessa etapa, o algoritmo confecciona três semivariogramas, o primeiro com alcance igual a 75% da distância máxima; o segundo com 50% e o terceiro com 25%. Por meio da análise desses gráficos concluiu-se que a variável em questão é espacialmente dependente.

Contudo, como pode ser visto na Figura 9, a dependência é moderada ($\frac{Pepita}{Patamar} \sim 0{,}55$) (CAMBARDELLA et al., 1994).

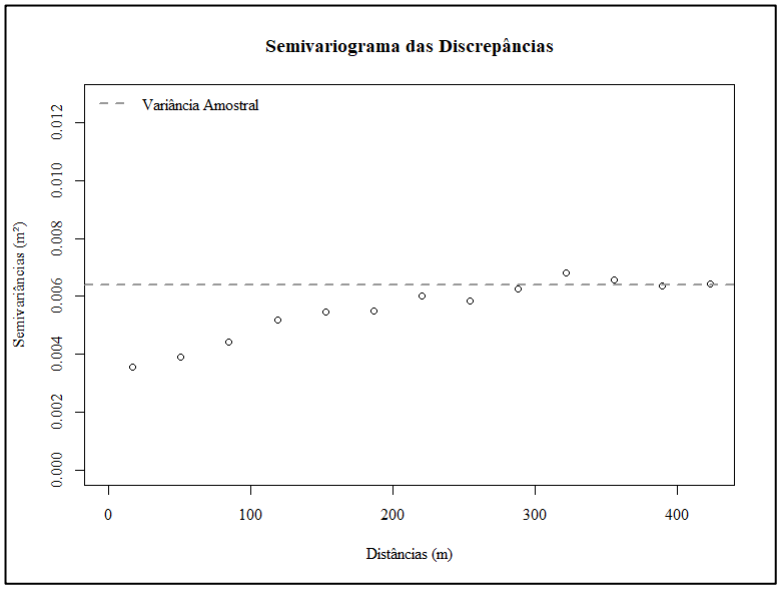

Figura 9 – Semivariograma das discrepâncias para distância de 440m (18% da distância máxima).

Constatada a dependência espacial, torna-se necessária a utilização de metodologias que levem em conta a autocorrelação espacial, uma vez que a desconsideração desta condição compromete a definição dos intervalos de confiança. Sendo assim, de acordo com o método proposto, aplica-se o *Block Bootstrap* para a estimativa dos níveis de confiança. Deve-se atentar que, na presença de autocorrelação espacial, a aplicação de testes de normalidade é, teoricamente, inconsistente.

A diagonal do bloco foi configurada com o valor aproximado do alcance, ou seja, 300 metros (Figura 9), e o número de replicações *Bootstrap* adotado foi de 1.000. Assim, o *Block Bootstrap* estabeleceu 1.000 novos conjuntos de dados, todos com 331 observações (tamanho do conjunto original, após a exclusão dos *outliers*). A partir destes, obteve-se a amostra *Bootstrap*, contendo 1.000 valores da estatística de interesse. A Figura 10 ilustra os histogramas e gráficos *Q-Q Plot* das amostras *Bootstrap* geradas para os estimadores: Φ, $RMSE$ e $\Phi_{Robusta}$.

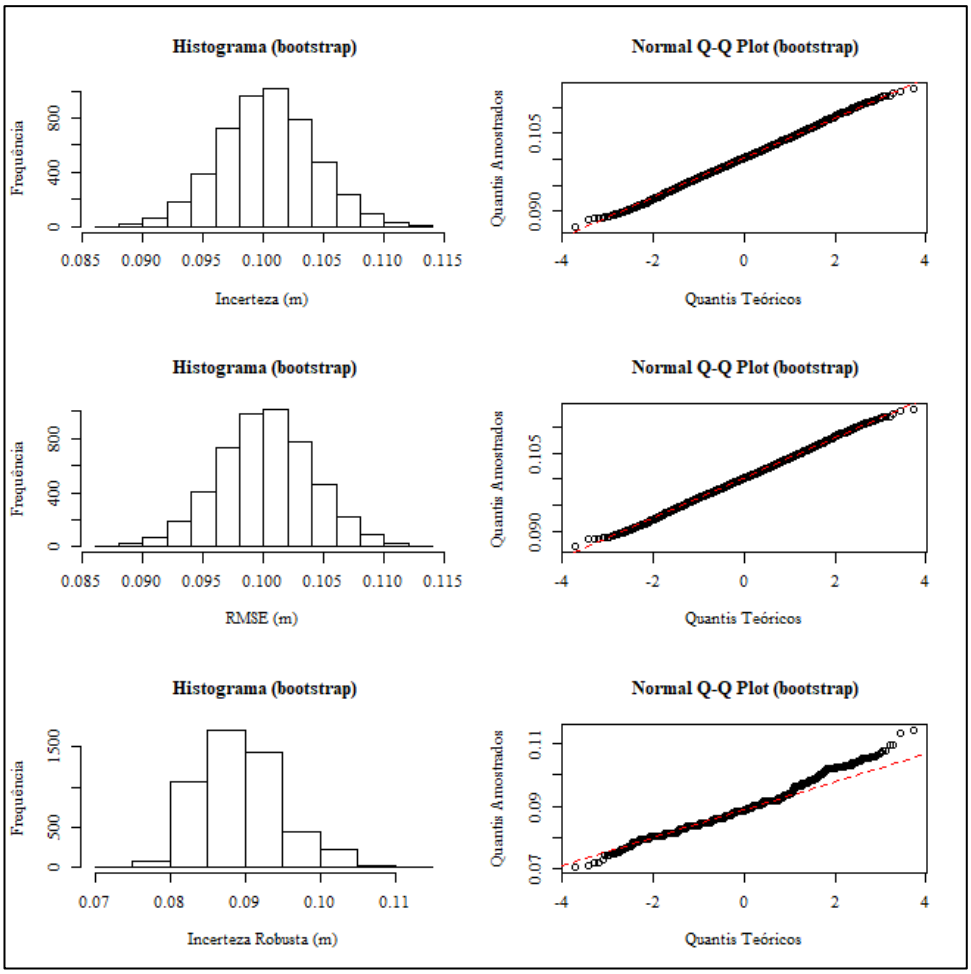

Figura 10 – Histograma e gráficos *Q-Q Plot* da amostra *Bootstrap* dos estimadores: Φ, $RMSE$ e $\Phi_{Robusta}$.

Analisando a Figura 10, percebe-se que as amostras *Bootstrap,* principalmente, as concernentes aos estimadores Φ e $RMSE$, aparentam seguir normalidade. Através dos cálculos da variância e dos coeficientes de assimetria e curtose, verifica-se que ambas as amostras apresentam baixa variabilidade (WARRICK & NIELSEN, 1980), com variância na casa do centésimo de milímetro, são basicamente simétricas e possuem funções de distribuição muito próximas de uma distribuição mesocúrtica. Dentre os três estimadores, a amostra referente a $\Phi_{Robusta}$ possui distribuição com menor grau de normalidade. Em todos os casos, os intervalos de confiança são extraídos das amostras *Bootstrap* através dos cálculos dos quantis 0,025 e 0,975, ou seja, $IC_{95\%} = [q_{0,025}, q_{0,975}]$.

Com o objetivo de avaliar as estimativas do algoritmo *Block Bootstrap* desenvolvido, foi implementado o cálculo do viés da amostra *Bootstrap*, que é definido como a diferença entre a incerteza estimada através da amostra original e a mediana da amostra *Bootstrap*. A Tabela 5 apresenta uma síntese das estatísticas calculadas.

Tabela 5 – Incerteza vertical amostral ao nível de confiança de 95% e viés das amostras *Bootstrap.*

Estimador	**Incerteza Vertical**	$IC_{95\%}$	**Viés da Amostra *Bootstrap***
$RMSE\ (m)$	0,100	[0,090; 0,107]	0
$\Phi\ (m)$	0,100	[0,090; 0,108]	0
$\Phi_{Robusta}\ (m)$	0,091	[0,080; 0,102]	0,002

Os estimadores apresentaram intervalos de confiança bastante estreitos, com amplitude em torno de 2 centímetros, o que mostra que os dados analisados possuem uma boa confiabilidade, isto é, tais amostras representam com fidelidade a população de origem e dessa forma, pode-se confiar no julgamento acerca da qualidade vertical do levantamento hidrográfico analisado. No que tange as amostras *Bootstrap*, geradas a partir do método *Block Bootstrap* implementado neste trabalho, nota-se que o viés calculado possui valor nulo para os estimadores $RMSE$ e Φ e apenas 2 milímetros para o estimador $\Phi_{Robusta}$. Tais fatos apenas compravam a eficiência da metodologia proposta.

De posse das estatísticas apresentadas na Tabela 5, conclui-se que ambos os estimadores exibem resultados coerentes. Todavia, destaca-se que a presença de *outliers* na base de dados pode, dependendo do estimador utilizado, mascarar os resultados. Sendo assim, quando forem utilizados os estimadores $RMSE$ ou Φ, deve-se ter certeza que a amostra não possui *outliers.* Nos casos em que há dúvidas acerca da presença de valores anômalos, uma opção é utilizar o estimador $\Phi_{Robusta}$. Na verdade, em todos os casos, a escolha desta última estatística sempre trará resultados mais confiáveis.

Computada a incerteza vertical, ao nível de confiança de 95%, pode-se proceder com a classificação do levantamento hidrográfico de acordo com as tolerâncias estipuladas na Publicação Especial nº 44 (S-44) ou demais normativas que a batimetria deva atender. A Tabela 6 exibe as tolerâncias definidas pela S-44 para o levantamento analisado, bem como a classificação alcançada através da análise tradicional (seção 2.4).

Tabela 6 – Tolerâncias estipuladas para o Levantamento Hidrográfico da área de estudo e classificação via exame tradicional (profundidade média: 3,315 metros).

Ordem	Intervalo de 95% de Tolerância (m)	Classificação
Especial	[-0,251; 0,251]	85,23%
1A/1B	[-0,502; 0,502]	91,71%
2	[-1,003; 1,003]	95,60%

Primeiramente deve-se esclarecer que o enquadramento do levantamento hidrográfico em determinada ordem (S-44) ou categoria (NORMAM-25) é condicionado a uma série de fatores e não somente ao intervalo de incerteza vertical amostral alcançado. Nesse sentido, ao focar apenas na incerteza vertical, a Tabela 6 sugere que o levantamento analisado seria classificado na Ordem 2, uma vez que pouco mais de 95% das discrepâncias possuem magnitude dentro do intervalo [-1,003; 1,003]. Segundo IHO (2008), esta ordem é a menos restritiva e destina-se a áreas onde a profundidade da água é tal que uma descrição geral do leito submerso é adequada. Note que, se menos de 95% das discrepâncias estivesse fora da tolerância definida para a Ordem 2, o levantamento não encontraria classificação junto a S-44.

Por outro lado, através da aplicação do método proposto, obteve-se uma Incerteza vertical em torno de 10 centímetros e $IC_{95\%}$ com amplitude máxima de 2 centímetros. Sendo assim, nitidamente, a classificação tradicional mostra-se pouco eficiente, subestimando a qualidade dos dados coletados. Vale ressaltar que a normativa citada neste texto concerne a avaliação de dados batimétricos destinados a produção de cartas náuticas que serão utilizadas com vistas à segurança da navegação de superfície e à proteção de ambiente marítimo. Assim sendo, uma avaliação mais fidedigna e confiável é sempre preferível.

3.2. Reservatório 2

No planejamento do batimetria do Reservatório 2 programou-se 669 interseções. Entretanto, após a sondagem e processamento, apenas 301 discrepâncias foram efetivamente obtidas. A Tabela a seguir resume o resultado da análise exploratória.

Tabela 7 – Estatística descritiva da área de estudo.

Número de Discrepâncias	301
Média (m)	-0,011
Mínimo (m)	-3,930
Máximo (m)	4,779
Variância (m²)	0, 5636
Coeficiente de Curtose	18,190
Coeficiente de Assimetria	-0,050
Distância Mínima (m)	2,365
Distância Máxima (m)	2576,151

A distribuição dos dados está centrada no valor -0,011 metros e possui discrepâncias variando de -3,930 a 4,779 metros. A variância indica uma variabilidade alta (WARRICK & NIELSEN, 1980). Por fim, os coeficientes de assimetria e curtose, definem uma distribuição com baixo grau de assimetria e leptocúrtica. A Figura 11 apresenta alguns gráficos que auxiliam na análise exploratória.

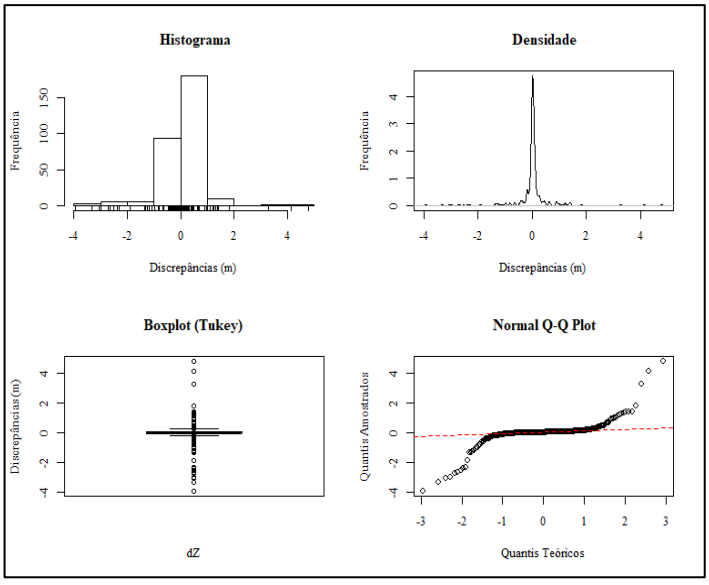

Figura 11 – Análise gráfica exploratória.

Diante destas informações, pode-se concluir que a amostra de discrepâncias analisada está eivada de *outliers* e é, nitidamente, não-normal. Todavia, a confirmação da presença de valores anômalos deve ser assistida por técnicas de detecção de *outliers*, enquanto a não normalidade comprovada através de testes de normalidade univariada.

As técnicas *Boxplot de Tukey*, *Boxplot Ajustado* e o *Z-Score Modificado*, detectaram, respetivamente, 67, 69 e 61 *outliers*. Todos foram minuciosamente

analisados antes da eliminação, uma vez que na mesma proporção que a presença de *outliers* compromete as análises, a exclusão indiscriminada de discrepâncias pode omitir informações importantes. Após a inspeção, contatou-se que para a área de estudo o método *Boxplot Ajustado* apresentou melhores resultados, embora apenas 58 *outliers* tenham sido realmente confirmados.

A Figura abaixo mostra o histograma, curva de densidade e o gráfico *Q-Q plot* da base de dados após a eliminação dos 58 *outliers*. Observando, especialmente, o gráfico *Q-Q plot*, percebe-se que a amostra não apresenta normalidade, conforme será comprovado adiante.

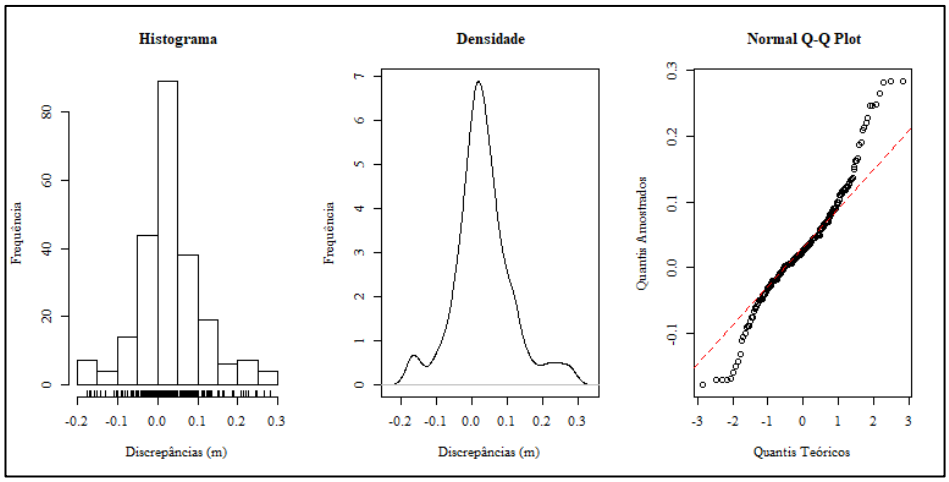

Figura 12 – Análise gráfica exploratória após a eliminação dos *outliers.*

Seguindo a metodologia proposta, a próxima etapa consiste no cálculo da incerteza amostral. A Tabela 8 apresenta as estimativas para o conjunto de dados sem *outliers* e com *outliers*, apenas a título de exemplificação, uma vez que o cálculo na presença de dados anômalos é teoricamente equivocado.

Tabela 8 – Estimativa pontual da incerteza vertical amostral.

Estimador	**Dados com *outliers***	**Dados sem *outliers***
$RMSE\ (m)$	0,750	0,088
$\Phi\ (m)$	0,751	0,088
$\Phi_{Robusta}\ (m)$	0,091	0,065

Os estimadores $RMSE$ e Φ, como esperado, apresentaram-se congruentes em todos os casos. Entretanto, conforme exposto nas seções anteriores, as estatísticas falseiam os resultados na presença de *outliers*. O estimador $\Phi_{Robusta}$ apontou uma

pequena divergência quando aplicado às bases de dados. Essa diferença, de cerca de 26 milímetros, deve-se, principalmente, a distribuição dos dados ou a exclusão demasiada de possíveis *outliers*, embora este último fato seja menos provável.

A Figura 13 ilustra o histograma da amostra de discrepâncias bruta, o mesmo observado na Figura 11, porém com uma riqueza maior de detalhes, isto é, um número maior de classes.

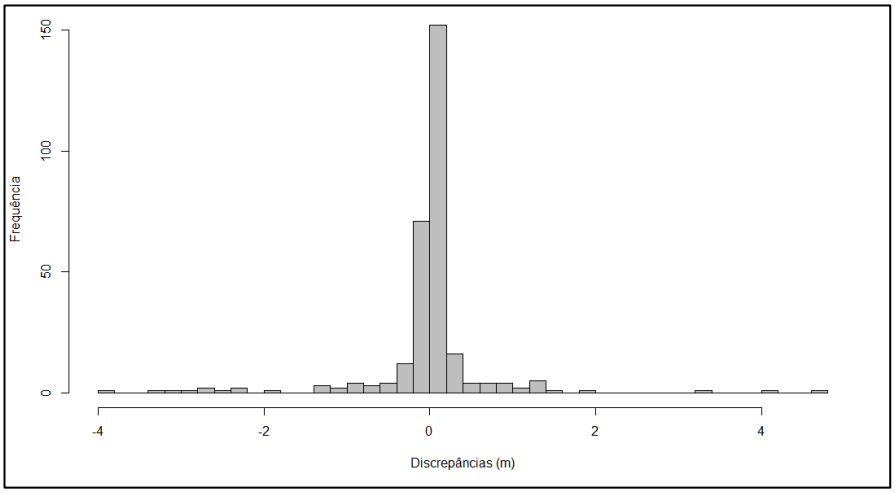

Figura 13 – Histograma dos dados brutos com 45 classes.

É notório que a distribuição dos dados é bastante distorcida, o que se deve, entre outros, a presença de *outliers*, alguns com magnitudes acima de 4 metros. Tal afirmação é comprovada através da simples visualização da Tabela 7 e Figura 13. Inevitavelmente, a natureza da distribuição das discrepâncias conduziu a uma variabilidade excessivamente alta, refletida no estimador $NMAD$ e consequentemente, no estimador $\Phi_{Robusta}$. Diante disso, pode-se concluir que o estimador incerteza robusta trata de forma adequada dados eivados de *outliers*, todavia, quando a amostra apresenta um maior grau de variabilidade, deve-se atentar que uma sutil subestimação da qualidade da batimetria pode ocorrer. Em todos os casos, o método proposto neste capítulo sugere um refinamento das análises.

De acordo com o fluxograma apresentado na seção 2, a próxima etapa constitui-se na verificação de independência dos dados. Essa etapa é realizada através da construção de três semivariogramas (Figura 14).

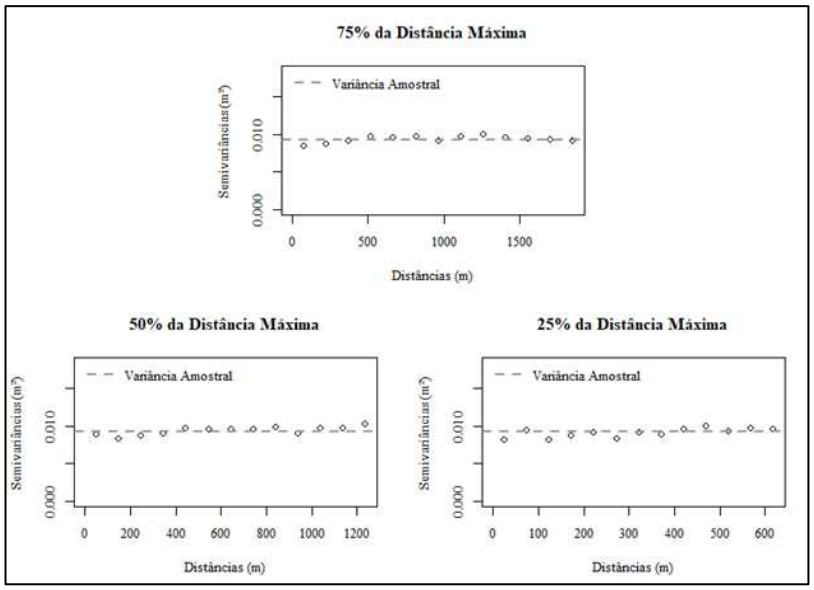

Figura 14 – Semivariogramas experimentais com 75%, 50% e 25% da distância máxima.

Conforme ilustrado na Figura 14, os semivariogramas apresentam efeito pepita puro, isto é, a variável analisada é espacialmente independente. Nessa etapa, pode-se ainda gerar o envelope de Morte Carlo com vistas a sanar quaisquer dúvidas que possam surgir. Assim sendo, a Figura 15 apresenta o semivariograma omnidirecional para 50% da distância máxima sobreposto ao envelope de Monte Carlo. Como nenhuma semivariância apresenta-se fora dos limites da simulação de Monte Carlo, conclui-se, portanto, que a variável analisada não apresenta autocorrelação espacial (ISAAKS,1990).

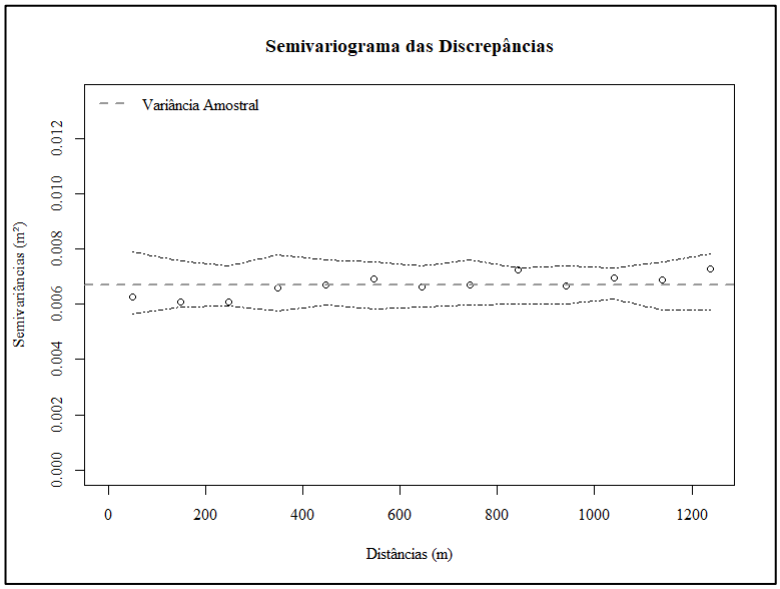

Figura 15 – Semivariograma das discrepâncias para distância de 1288m sobreposto ao envelope de Monte Carlo (50% da distância máxima).

Uma vez comprovada a independência da base de dados, pode-se recorrer ao teste *Kolmogorov-Smirnov* (KS), ao nível de confiança de 95%, para investigação acerca da normalidade dos dados. Conforme esperado, após a aplicação do teste KS constatou-se a não normalidade da amostra (valor-p = 0,017).

Diante disso, o método proposto sugere duas linhas de tratamento (seção 2.2): aplicação do TCL (Teorema Central do Limite) ou aplicação da abordagem robusta. Em termos teóricos, recomenda-se dar preferência a utilização do TCL visto que este conduzirá a uma análise paramétrica, que como sabido, possui maior poder estatístico. Conquanto, poderá haver casos em que a aplicação do TCL não acarrete a uma análise consistente, isto é, a amostra transformada não siga normalidade. Dado o exposto, por mais experiente que o analista seja, sugere-se invariavelmente aplicar ambas as linhas de tratamento.

Para aplicação do TCL foi implementada uma rotina que aplica um processo de clusterização, baseado no algoritmo *k*-medoids. Para esta área de estudo, dado a quantidade e distribuição espacial das discrepâncias, o tamanho amostral dos agrupamentos foi definido com o valor mínimo sugerido, isto é, 4, e dissimilaridade foi quantificada pela distância "euclidiana". A partir do conjunto de dados originais foram gerados 60 *clusters*, conforme pode ser visto na Figura 16. De posse desses agrupamentos, calculam-se a média das discrepâncias para cada *cluster,* obtendo a amostra TCL (Figura 16). Conforme afirma o TCL, a distribuição das médias tende a uma distribuição normal, com a mesma média da amostra original e com a variância dividida pelo tamanho amostral dos agrupamentos, ou seja, $\sigma^2_{TCL} = {\sigma^2}/{4}$.

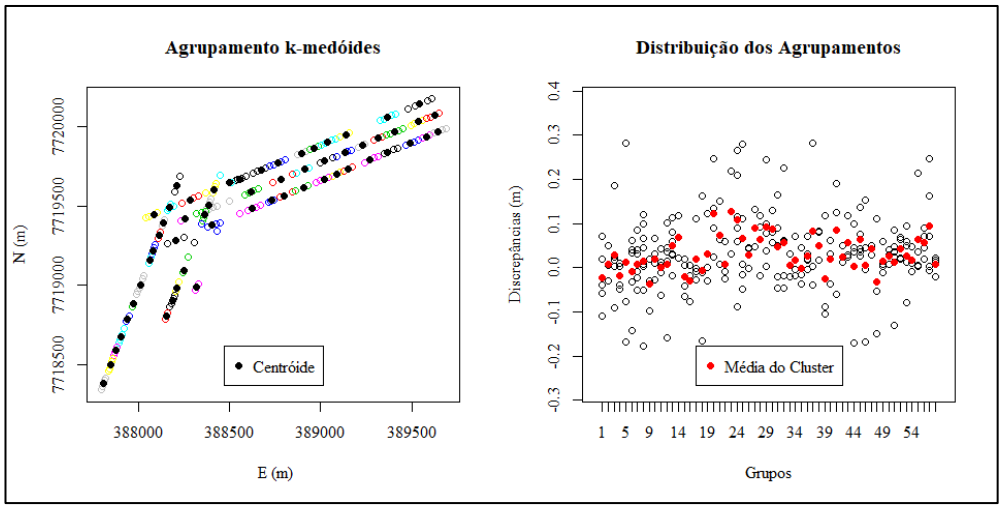

Figura 16 – Agrupamentos e distribuição dos agrupamentos obtidos pelo algoritmo *k*-medoids.

Obtida a amostra de médias do TCL, aplica-se o teste KS com objetivo de confirmar o teorema, isto é, a normalidade do novo conjunto de dados. Através do teste KS, pôde-se constatar que a nova amostra é normal (valor-p = 0,3373). Diante disso, recorre-se a Equação 8 para obter uma estimativa da Incerteza Vertical, enquanto o $IC_{95\%}$ é obtido através das técnicas apresentadas na seção 2.1. A Tabela 9 sumariza os resultados obtidos.

Tabela 9 – Incerteza vertical amostral (TCL) ao nível de confiança de 95%.

Estimador		$IC_{bootstrap-t}$	$IC_{95\%}^{BCa}$
$\Phi_{TCL}\ (m)$	0,086	[0,071; 0,108]	[0,074; 0,101]
$\mu_{TCL}\ (m)$	0,032	-	-
$\sigma_{TCL}^2\ (m^2)$	0,0016	-	-

Nota-se que a média da amostra TCL é igual a média da amostra original, à medida que a σ_{TCL}^2 corresponde a exatamente $1/4$ da variância da base de dados (sem *outliers*), conforme afirmativa do Teorema Central do limite.

O estimador Φ_{TCL} mostrou coerência com as estimativas da Tabela 8, confirmando a robustez da metodologia proposta. Mesmo assim, optou-se por aplicar também a abordagem robusta, embora o valor pontual da $\Phi_{Robusta}$ já tenha sido apontado na Tabela 8. Objetiva-se com isso obter parâmetros de comparação, bem como alcançar uma maior confiabilidade no processo. Os resultados são apresentados a seguir (Tabela 10).

Tabela 10 – Incerteza vertical amostral (robusta) ao nível de confiança de 95%.

$\Phi_{Robusta}\ (m)$	$IC_{bootstrap-t}\ (m)$	$IC_{95\%}^{BCa}\ (m)$
0,065	[0,056; 0,077]	[0,057; 0,080]

Em ambos os casos, os intervalos de confiança expressam uma boa confiabilidade, com amplitude mínima de 2,1 centímetros e máxima de 3,7 centímetros. Isso garante que a amostra de discrepâncias analisada representa com fidelidade a população, e assim, a estimativa da qualidade do levantamento batimétrico através deste conjunto de dados torna-se confiável. Quanto a estimativas *Bootstrap* para os $IC_{95\%}$, ambas se mostraram coerentes, sem diferenças significativas.

De posse das estatísticas apresentadas, conclui-se que ambas as abordagens expressaram resultados coesos e, nesse caso, podem ser aplicadas a critério do analista.

Entretanto, destaca-se que a presença de *outliers* na base de dados pode, dependendo do estimador utilizado, mascarar os resultados. Por outro lado, apesar da abordagem TCL exibir-se como uma alternativa válida nos casos em que a amostra de discrepâncias for independente e não-normal, a sua aplicação requer uma análise mais cuidadosa e, portanto, morosa. Nesse sentido, o uso da abordagem robusta mostra-se uma escolha mais adequada na maioria das vezes.

Computada a incerteza vertical, ao nível de confiança de 95%, pode-se proceder com a classificação do levantamento hidrográfico de acordo com as tolerâncias estipuladas na Publicação Especial nº 44 (S-44). A Tabela 11 exibe as tolerâncias estipuladas pela S-44 para o levantamento analisado, bem como a classificação alcançada através da análise tradicional (seção 2.4).

Tabela 11 – Tolerâncias estipuladas para o Levantamento Hidrográfico da área de estudo e classificação via exame tradicional (profundidade média: 5,246 metros).

Ordem	**Intervalo de 95% de Tolerância (m)**	**Classificação**
Especial	[-0,253; 0,253]	78,41%
1A/1B	[-0,505; 0,505]	86,05%
2	[-1,007; 1,007]	91,36%

De acordo com a análise tradicional, o levantamento hidrográfico do Reservatório 2 não apresenta classificação junto a S-44. No entanto, através da aplicação do método proposto, obteve-se uma Incerteza vertical menor que 10 centímetros e $IC_{95\%}$ com amplitude máxima em torno de 2 centímetros. Sendo assim, claramente, a classificação tradicional mostra-se pouco eficiente, subestimando a qualidade dos dados coletados.

4. CONCLUSÕES

Nota-se que a avaliação da qualidade de uma sondagem batimétrica não é tarefa fácil, principalmente pela inexistência de observações redundantes. Diferentemente de medidas realizadas em terra, onde observações repetidas permitem uma avaliação estatística fácil e rápida, a análise das profundidades coletadas por sistemas de sondagem batimétrica exige acomodações teóricas.

Tal fato, nitidamente, conduz a erros nas inferências estatísticas, principalmente, quando as análises são realizadas de maneira equivocada e sem qualquer cuidado. Outrem, o cálculo de medidas de incerteza, em qualquer área, deve se adaptar ao fato de que valores anômalos possam existir e que os dados analisados podem apresentar-se dependentes espacialmente e não-normais.

Diante da carência por medidas confiáveis de incerteza, este trabalho teve como objetivo principal o desenvolvimento de uma metodologia para avaliação da qualidade vertical de sondagens batimétricas, prioritariamente, de levantamentos monofeixe, denominada MAIB.

Através dos resultados pode-se concluir que numa análise estatística coerente, deve-se, primeiramente, verificar a presença de *outliers* e efetuar testes de independência e normalidade e só então, inferir quaisquer conclusões acerca das incertezas relacionadas ao produto analisado. Pôde-se perceber também a importância da apresentação das estimativas de incerteza sempre em conjunto com os respectivos intervalos de confiança, uma vez que, o simples fornecimento de uma medida pontual, não é capaz de descrever com clareza a qualidade dos dados.

Deve-se destacar que as metodologias de avaliação da qualidade de dados batimétricos devem primar pelo rigor estatístico e, como comprovado neste estudo, as técnicas de avaliação tradicionais mostraram-se pouco eficientes. Sendo assim, recomenda-se que a metodologia proposta seja utilizada na prática hidrográfica.

Por fim, apesar da MAIB ter sido desenvolvida com ênfase no tratamento de dados batimétricos, o seu uso em outras áreas da ciência, principalmente, na engenharia de posição, é possível e, inclusive, recomendado. O algoritmo da MAIB encontra-se disponível (vide Apêndice) e é bastante eficaz também em termos de tempo de processamento. Para processar os dados do Reservatório 1 e 2, utilizando uma máquina com sistema operacional Windows 10, memória RAM de 8GB (parcialmente dedicada ao *software R*) e processador Intel® Core™ i7-4500U CPU @ 1,80GHz 2,40 GHZ, foram gastos aproximadamente 8 minutos (excluindo-se os tempos de análise).

Através dos estudos realizados neste trabalho, em paralelo ao desenvolvimento da MAIB, foi concebido uma nova medida estatística para estimativa da incerteza vertical amostral de levantamentos hidrográficos. Tal estimador é simbolizado por $\Phi_{Robusta}$ e possui como principal característica, ser resistente, isto é, o estimador pontual é pouco afetado por mudanças de uma pequena porção das observações. Em

outras palavras, pode-se dizer que as estimativas providas pela $\Phi_{Robusta}$ não são influenciadas por possíveis *outliers* presentes na base de dados.

Para trabalhos futuros recomenda-se que a MAIB, bem como a estatística $\Phi_{Robusta}$, sejam aplicadas em levantamento batimétricos realizados através de sistemas de sondagem por faixa, bem como em outras áreas da engenharia de posição. Melhorias nos algoritmos do *Block Bootstrap* e Teorema Central do Limite também são desejáveis, principalmente, no que tange o tratamento de grandes massas de dados.

REFERÊNCIAS BIBLIOGRÁFICAS

ANA - Agência Nacional de Águas. **Orientações para atualização das curvas cota x área x volume**. Superintendência de Gestão da Rede Hidrometeorológica. Brasília: ANA, SGH, 2013.

ATHEARN, N.; TAKEKAWA, J.; JAFFE, B.; HATTENBACH, B. Mapping elevations of tidal wetlands restoration sites in San Francisco Bay: comparing accuracy of aerial Lidar with a singlebeam echosounder. **Journal of Coastal Research**, v. 26, n. 2, p. 312–319, 2010.

CAMBARDELLA, C. A.; MOORMAN, T. B.; NOVAK, J. M.; PARKIN, T. B.; KARLEN, D. L.; TURCO, R. F.; KONOPKA, A. E. Field scale variability of soil properties in Central Iowa soils. **Soil Science Society of America Journal**, Madison, v. 58, n. 5, p. 1501-1511, 1994.

CARMO, E. J. **Avaliação dos interpoladores krigagem e topo to raster na geração de modelos digitais de elevação a partir de dados batimétricos.** Dissertação (Mestrado). Programa de Pós-Graduação em Engenharia Civil, Departamento de Engenharia Civil, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 95p., 2014.

CARPENTER, J. & BITHELL, J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. **Statistics in medicine,** v. 19, n.9, p. 1141-1164, 2000.

CARVALHO, N. O.; FILIZOLA JÚNIOR, N. P.; SANTOS, P. M. C.; LIMA, J. E. F. W. **Guia de avaliação de assoreamento de reservatórios.** Superintendência de Estudos e Informações Hidrológicas. Brasília: ANEEL. 140p., 2000.

COOPER, M. A. R. **Control surveys in civil engineering**. Nichols Pub Co, 381p., 1987.

DHN – Diretoria de Hidrografia e Navegação. **NORMAM 25 – Normas da Autoridade Marítima para Levantamentos Hidrográficos**. Marinha do Brasil. Brazil. 2014.

DOOB, J. L. Heuristic approach to the Kolmogorov-Smirnov theorems. **The Annals of Mathematical Statistics**, v. 20, n. 3, p. 393-403, 1949.

EAKIN, H. M. **Silting of reservoirs**. Technical bulletin n° 524. Department of Agriculture. Revised by Brown, C.B., U. S. Government Printing Office, Washington, United States, 167p., 1939.

EEG, J. Multibeam Crosscheck Analysis: A Case Study. **The International Hydrographic Review**, n. 4, p. 25-33, 2010.

EFRON, B. & TIBISHIRANI, R. J. **An Introduction to the Bootstrap**. New York: Chapman & Hall, 436p., 1993.

EFRON, B. Computers and the theory of statistics: thinking the unthinkable. **SIAM review**, v. 21, n. 4, p. 460-480, 1979.

FERRARI, R. L. **Reconnaissance Techniques for Reservoir Surveys**. Denver, Colorado, U.S. Department of the Interior, Bureal of Reclamation, Technical Service Center, 2006.

FERREIRA, Í. O. ; RODRIGUES, D. D. ; SANTOS, A. P. Levantamento batimétrico automatizado aplicado à gestão de recursos hídricos. Estudo de caso: represamento do ribeirão São Bartolomeu, Viçosa-MG. **In: IV Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação**, 2012, Recife. Geotecnologias para o Planejamento e a Gestão Eficiente do território, 2012.

FERREIRA, Í. O.; RODRIGUES, D. D.; NETO, A. A.; MONTEIRO, C. S. Modelo de incerteza para sondadores de feixe simples**. Revista Brasileira de Cartografia**, v. 68, n. 5, p. 863-881, 2016a.

FERREIRA, Í. O.; NETO, A. A.; MONTEIRO, C. S. O uso de embarcações não tripuladas em levantamentos batimétricos**. Revista Brasileira de Cartografia**, v. 68, n. 10, p. 1885-1903, 2017a.

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.; **Coleta, processamento e análise de dados batimétricos**. 1ª ed. Saarbrucken: Novas Edições Acadêmicas, v. 1, 100p., 2015.

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.; ROSA, L. M. F. In bathymetric surfaces: IDW or Kriging? **Boletim de Ciências Geodésicas**, v. 23, n. 3, p. 493-508, 2017b.

FERREIRA, Í. O.; SANTOS, G. R.; RODRIGUES, D. D. Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas. **Revista Brasileira de Cartografia**, Rio de Janeiro, v. 65, n. 5, p. 831-842, 2013.

FERREIRA, Í. O.; ZANETTI, J.; GRIPP, J. S.; MEDEIROS, N. G. Viabilidade do uso de imagens do sistema Rapideye na determinação da batimetria de águas rasas. **Revista Brasileira de Cartografia**, v. 68, n. 7, p. 1331-1340, 2016b.

FGDC – Federal Geographic Data Committee. **National Standard for Spatial Data Accuracy**, **Part 3: National Standard for Spatial Data Accuracy.** Federal Geographic Data Committee: Reston, USA, 25p., 1998.

FILHO, N. A. P. **Teste Monte Carlo de Normalidade Univariado**. Tese (Doutorado). Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, Minas gerais, 56p., 2013.

FISCHER, H. **A history of the central limit theorem: From classical to modern probability theory**. Springer Science & Business Media, 399p., 2010.

FRANCO, G. C. & REISEN, V. A. Bootstrap approaches and confidence intervals for stationary and non-stationary long-range dependence processes. **Physica A: Statistical Mechanics and its Applications,** v. 375, n. 2, p. 546-562, 2007.

GREENWALT, C. R. & SCHULTZ, M.E. Principles of Error Theory and Cartographic Applications. **Aeronautical Chart and Information Center**: St. Louis, MO, USA, 98p., 1962.

HARE, R.; EAKINS, B.; AMANTE, C. Modelling bathymetric uncertainty. **The International Hydrographic Review,** n. 6, p. 31-42, 2011.

HESTERBERG, T.; MOORE, D. S.; MONAGHAN, S.; CLIPSON, A.; EPSTEIN, R. Bootstrap methods and permutation tests. **In: The practice of business statistics: using data for decisions**. New York: W.H. Freeman, 2003.

HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. **Understanding robust and exploratory data analysis**. New York: Wiley, 433p., 1983.

HÖHLE, J. & HÖHLE, M. Accuracy assessment of digital elevation models by means of robust statistical methods. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 64, n. 4, p. 398-406, 2009.

HUBERT, M. & VANDERVIEREN, E. An adjusted boxplot for skewed distributions. **Journal of Computational statistics & data analysis**, v. 52, n. 12, p. 5186-5201, 2008.

HYPACK, Inc. **Hypack – Hydrographic Survey Software User Manual**. Middletown, USA, 1784p., 2012.

IGLEWICZ, B. & HOAGLIN, D. **How to detect and handle outliers**. Milwaukee, Wis.: ASQC Quality Press, 87p., 1993.

IHO – International Hydrographic Organization. **C-13: IHO Manual on Hydrography**. Mônaco: International Hydrographic Bureau, 540p., 2005.

IHO – International Hydrographic Organization. **S-44: IHO Standards for Hydrographic Surveys**. Special Publication n. 44–5th. Mônaco: International Hydrographic Bureau, 36p., 2008.

INMETRO – Instituto Nacional de Metrologia Normalização, Qualidade e Tecnologia. **Avaliação de dados de medição: guia para a expressão de incerteza de medição (GUM 2008)**. Duque de Caxias, RJ: INMETRO/CICMA/SEPIN, 141 p., 2012a

INMETRO – Instituto Nacional de Metrologia Normalização, Qualidade e Tecnologia. **Vocabulário Internacional de Metrologia: conceitos fundamentais e gerais de termos associados (VIM 2012)**. Duque de Caxias, RJ : INMETRO, 94 p., 2012b.

INSTITUTO HIDROGRÁFICO. **Especificação Técnica para Produção de cartografia hidrográfica**. Marinha Portuguesa, Lisboa, Portugal, v 0.0, 24p., 2009.

ISAAKS, E. H. **The application of Monte Carlo methods to the analysis of spatially correlated data**. PhD Thesis, Department of Applied Earth Sciences, Stanford University, USA, 213p., 1990.

KREISS, J. P. & PAPARODITIS, E. Bootstrap methods for dependent data: A review. **Journal of the Korean Statistical Society**, v. 40, n.4, p. 357-378, 2011.

LAHIRI, S. N. Resampling methods for dependent data. **Springer Science & Business Media**, 374p., 2003.

LAHIRI, S. N. Theorical Comparisons of block bootstrap methods. **The Annals Of Statistics,** v. 27, n. 1, p. 386-404, 1999.

LEE, S. M. S. & LAI, P. Y. Double block bootstrap confidence intervals for dependent data. **Biometrika**, v. 96, n. 2, p. 427-443, 2009.

LI, Z.; ZHU, Q.; GOLD, C. M. **Digital terrain modelling. Principles and methodology**. New York: CRC Press, 319p., 2005.

LINZ – Land Information New Zealand. **Contract Specifications for Hydrographic Surveys**. New Zealand Hydrographic Authority, V. 1.2, 111p., 2010.

MACHADO, A. F.; DE ALMEIDA, A. C.; ARAÚJO, A. C.; FERRARI, D.; LEMES, Í. R.; FARIA, N. C. S.; LIMA, T. S.; FERNANDES, R. A. Aplicação de testes de normalidade em publicações nacionais: levantamento bibliográfico. **Colloquium Vitae**; v. 6, n.1, p. 01-10, 2014.

MALEIKA, W. The influence of the grid resolution on the accuracy of the digital terrain model used in seabed modeling. **Marine Geophysical Research**, v. 36, n. 1, p. 35-44, 2015.

MAUNE, D. F. Digital Elevation Model Technologies and Applications: The DEM Users Manual. **American Society for Photogrammetry and Remote Sensing**, 2007.

MATHERON, G. **Les variables régionalisées et leur estimation**. Paris: Masson, 306p., 1965.

MIKHAIL, E. & ACKERMAN, F. **Observations and Least Squares**. University Press of America, 497p., 1976.

MONICO, J. F. M.; DAL POZ, A. P.; GALO, M.; SANTOS, M.C.; OLIVEIRA, L. C. Acurácia e Precisão: Revendo os conceitos de forma acurada. **Boletim de Ciências Geodésicas**, v. 15, n. 3, p. 469-483, 2009.

MOOD, A. M. **Introduction to the theory of statistics**. McGraw-Hi1l series in probability and statistics, 564p., 1913.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the Theory of Statistics**. McGraw-Hill International, 577p.1974.

MORETTIN, P. A. & BUSSAB, W. O. **Estatística básica**. 5ª ed. São Paulo: Editora Saraiva, 526p., 2004.

NOAA – National Oceanic and Atmospheric Administration. **Field Procedures Manual**. Office of Coast Survey, 2011.

R Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. 2017.

REYNOLDS, A. P.; RICHARDS, G.; DE LA IGLESIA, B.; RAYWARD-SMITH, V. J. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. **Journal of Mathematical Modelling and Algorithms**, v. 5, n.4, p. 475–504, 1992.

SANTOS, A. M. R. T.; SANTOS, G. R.; EMILIANO, P. C.; MEDEIROS, N. G.; KALEITA, A. L.; PRUSKI, L. O. S. Detection of inconsistencies in geospatial data with geostatistics. **Boletim de Ciências Geodésicas,** v. 23, n. 2, p. 296-308, 2017.

SANTOS, A. P. **Controle de qualidade cartográfica: metodologias para avaliação da acurácia posicional em dados espaciais**. Tese (Doutorado). Programa de Pós-Graduação em Engenharia Civil, Departamento de Engenharia Civil, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 172p., 2015.

SEKELLICK, A. J., BANKS, W. S. L. **Water Volume and Sediment Accumulation in Lake Linganore, Frederick County, Maryland**. Scientific Investigations Report 2010–5174. USGS - U.S. Geological Survey. 2010.

SHAPIRO, S. S. & WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3/4, p. 591-611, 1965.

STEIGER, J. H., & LIND, J. C. Statistically-based tests for the number of common factors. **Annual Spring meeting of the Psychometric Society**, Iowa City, IA, 1980.

SUSAN, S. & WELLS, D. Analysis of Multibeam Crosschecks Using Automated Methods. **In: US Hydro 2000 Conference paper**, Biloxi, Mississippi. 2000.

TOBLER, W. R. A computer movie simulating urban growth in the Detroit region. Proceedings in International Geographical Union. Commission on Quantitative Methods, **Economic Geography**, v. 46, p. 234-240, 1970.

TUKEY, J.W. **Exploratory Data Analysis**. Princeton, Ed. Pearson (1977).

USACE - U.S. Army Corps of Engineers. **HIDROGRAPHIC SURVEYING**. Engineer Manual No 1110-2-1003. Department of the Army. Washington, D. C., 30 Nov. 2013.

USACE – U.S. Army Corps of Engineers. **HYDROGRAPHIC SURVEYING**. Engineer Manual n. 1110-2-1003. Department of the Army. Washington, D. C., USA, 2002.

VANDERVIEREN, E. & HUBERT, M. An adjusted boxplot for skewed distributions. **Proceedings in Computational Statistics**, p. 1933-1940, 2004.

VEIGA, L.; SILVA, R.; ARTILHEIRO, F. Levantamentos para fins especiais: dragagens. **Primeiras Jornadas de Engenharia Hidrográfica**, Instituto Hidrográfico. 2010.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. **Tópicos em ciências do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, v.1. p. 2-54, 2000.

WARRICK, A.W. & NIELSEN, D.R. Spatial variability of soil physical properties in the field. In: HILLEL, D. **Applications of soil physics**. New York: Academic Press, p.319-344, 1980.

WMO - WORLD METEOROLOGICAL ORGANIZATION**. Manual on Sediment Management and Measurement**. Operational Hydrology Report No. 47. Geneva, Switzerland, Secretariat of the World Meteorological Organization, 2003.

CAPÍTULO 3. PROPOSTA METODOLÓGICA PARA AVALIAÇÃO DA QUALIDADE VERTICAL DE DADOS BATIMÉTRICOS COLETADOS A PARTIR DE SISTEMAS DE SONDAGEM POR FAIXA

Resumo:
A carta náutica é o principal produto resultante de um levantamento hidrográfico. Nesse sentido, destaca-se a importância de avaliar a qualidade dos dados coletados, principalmente, por sistemas de sondagem por faixa. Devido à natureza das informações batimétricas, a avaliação da qualidade vertical das sondagens não é tarefa simples. É muito comum no meio hidrográfico estimar a qualidade vertical de uma sondagem realizada com ecobatímetros monofeixe através de linhas de verificação. Todavia, o volume de dados gerados numa sondagem, através de sistemas de varredura, impôs adaptações a essa metodologia. Em todos os casos, obtém-se um arquivo de discrepância que é analisado com o propósito de fornecer a qualidade vertical do levantamento hidrográfico. Entretanto, quase sempre essas estimativas são realizadas sem qualquer critério estatístico. Face ao exposto, o objetivo deste trabalho é propor um método que permita obter um arquivo de discrepâncias para posterior avaliação estatística das profundidades coletadas por sistemas de sondagem por varrimento, abordando normalidade e independência, bem como a detecção de *outliers* na base de dados. Também é apresentada uma alternativa à realização de varreduras de verificação para a estimativa da incerteza vertical amostral.

1. INTRODUÇÃO

Devido a eficiência do transporte aquaviário (hidroviário e marítimo), o uso de rotas marítimas e fluviais tem crescido nos últimos tempos. Estima-se que mais de 80% do comércio internacional é realizado por vias aquáticas (IHO, 2005). O Brasil possui mais de 40.000 quilômetros de vias interiores navegáveis, uma área jurisdicional superior a 4,5 milhões de km² e quase 9.000 quilômetros de costa marítima, características propícias à navegação que tornam o transporte aquaviário um meio bastante promissor (OLIVA, 2008; POMPERMAYER et al., 2014).

O transporte marítimo, seja de cabotagem (navegação costeira) ou de longo curso (navegação entre portos brasileiros e estrangeiros), é o mais importante, respondendo por quase 75% do comércio internacional brasileiro. Por outro lado, o transporte hidroviário, apesar de ser o meio mais econômico e limpo, ainda é pouco desenvolvido no Brasil. Entretanto, existem regiões, como é o caso da Amazônia, que dependem quase que exclusivamente deste meio de transporte, uma vez que estradas

e ferrovias são escassas e bastante precárias (OLIVA, 2008; POMPERMAYER et al., 2014).

Nesse sentido, destaca-se a importância do desenvolvimento de uma cartografia náutica de qualidade. Conforme afirma IHO (2005), os navegantes possuem uma fé inquestionável nas cartas e publicações náuticas, de tal forma que, na ausência da representação cartográfica de um perigo, eles acreditam fielmente na sua inexistência. A carta náutica é o produto final de um levantamento hidrográfico, sendo assim, a sua acurácia é dependente da qualidade dos dados adquiridos. Enfatiza-se que, embora o principal interesse desses levantamentos sejam a navegação, diversas outras finalidades são atendidas pelos dados coletados (IHO, 2005; FERREIRA et al., 2012).

Segundo Miguens (1996), cartas náuticas são documentos cartográficos resultantes de levantamentos de áreas oceânicas, mares, baías, rios, canais, lagos, lagoas, represamentos ou qualquer outra massa d'água navegável e que se destinam a servir de base à navegação. O objetivo das cartas e publicações náuticas é representar acidentes terrestres e submarinos, fornecendo informações sobre profundidades e a natureza do fundo; objetos que ofereçam perigos à navegação, tais como: bancos de areia, pedras submersas, cascos soçobrados, *etc*.; áreas de fundeio; altitudes e pontos notáveis; linha de costa; ilhas; elementos de maré; auxílios à navegação, como por exemplo: faróis, faroletes, boias, balizas, luzes de alinhamento, radiofaróis, *etc*.; além de outras indicações necessárias à segurança da navegação.

Compete à DHN (Diretoria de Hidrografia e Navegação), por meio do CHM (Centro de Hidrografia da Marinha), executar os levantamentos hidrográficos com vista a manter todas as cartas náuticas em Águas Jurisdicionais Brasileiras atualizadas. Além das cartas náuticas tradicionais (em papel), a DHN, seguindo tendências mundiais, também reproduz cartas em formato *raster* e cartas náuticas eletrônicas *(ENC – Eletronic Navigation Chart),* em conformidade com os padrões estabelecidos pela IHO (DHN, 2014). As cartas náuticas são, sem dúvida, o documento mais importante de auxílio ao navegante. Contudo, além das cartas náuticas, os navegantes fazem uso das publicações de auxílio à navegação, cujas informações complementam ou ampliam os elementos fornecidos pelas Cartas Náuticas. Nesse âmbito, as publicações náuticas também devem ser mantidas atualizadas[7].

[7] http://www.mar.mil.br/dhn/chm/box-levantamento-hidrografico/levantamento_relacao.html

Além dos levantamentos executados pela Marinha, o CHM fiscaliza, por força de determinação legal, a execução dos levantamentos hidrográficos executados por entidades extra Marinha com vistas ao aproveitamento dos dados para confecção ou atualização da cartografia náutica. Sendo assim, a Marinha do Brasil também é responsável por estabelecer normas e procedimentos específicos referentes à coleta, processamento e envio dos dados, bem como à confecção dos relatórios finais. Essas especificações técnicas são elaboradas segundo padrões internacionais de qualidade recomendados pela IHO *(International Hydrographic Organization)*.

As normas e os procedimentos para autorização e controle dos levantamentos hidrográficos estão, atualmente, definidos na NORMAM – 25 (DHN, 2014). A NORMAM-25 classifica os levantamentos hidrográficos em duas categorias de natureza administrativa, a saber: CATEGORIA ALFA: levantamentos hidrográficos que devem seguir especificações técnicas que permitam que os dados obtidos sejam aproveitados na atualização de cartas náuticas ou para as demais finalidades descritas no item 0206 da NORMAM-25; e a CATEGORIA BRAVO: levantamentos hidrográficos executados sem o propósito de produzir elementos que sirvam para atualização de cartas náuticas.

Os levantamentos hidrográficos Categoria ALFA ou A devem cumprir integralmente as especificações previstas na Publicação Especial S-44, 5ª edição (IHO, 2008; DHN, 2014). Nesse sentido, a S-44 especifica quatro Ordens de Levantamentos: Ordem Especial, Ordem 1a, Ordem 1b e Ordem 2 (IHO, 2008). É importante enfatizar que a S-44 define os procedimentos gerais para os levantamentos hidrográficos cujo objetivo seja a segurança da navegação, especificadamente, os requisitos mínimos em termos de incerteza das profundidades coletadas. A S-44 também esclarece que é competência dos serviços hidrográficos nacionais o estabelecimento de especificações técnicas que melhor se adequem a dinâmica dos levantamentos executados.

Embora a NORMAM – 25 não estabeleça procedimentos técnicos específicos para os levantamentos hidrográficos categoria BRAVO ou B, ela recomenda a adoção dos mesmos procedimentos técnicos dos levantamentos ALFA, visando uma eventual alteração de categoria, com vista ao aproveitamento dos dados.

Na atualidade, a realização de levantamentos hidrográficos, principalmente aqueles destinados a atualização cartográfica, concentra-se no uso de ecobatímetros multifeixe e sonares interferométricos para medição de profundidade. Em comparação com os antecessores ecobatímetros monofeixe, estes sistemas apresentam um elevado

ganho em resolução e acurácia, tanto em termos planimétricos quanto altimétricos (profundidade). Devido ao grande adensamento de dados, houve uma enorme melhora na capacidade de detecção de objetos (CRUZ et al., 2014; MALEIKA, 2015). Sistemas menos eficientes já são capazes de coletar em águas rasas mais de 30 milhões de pontos por hora (BJØRKE & NILSEN, 2009).

Os sonares interferométricos são uma tecnologia relativamente nova, porém passível de alcançar resultados semelhantes ou superiores aos da batimetria multifeixe, com ganhos, especialmente, na cobertura de fundo em águas rasas (CRUZ et al., 2014). A cobertura de fundo de um sistema multifeixe convencional é cerca de 4 vezes a profundidade, enquanto que um sonar interferométrico é capaz de cobrir uma faixa de até 12 vezes (CHS, 2013).

O processo de determinação da profundidade baseia-se na integração de diversas medições individuais além daquelas efetivamente realizadas pelos sonares, tais como: a profundidade de imersão do transdutor, a altura de maré, a atitude da embarcação de sondagem, a posição da embarcação de sondagem, o perfil de velocidade do som na água, *etc*. Todas essas medições são realizadas com um certo grau de incerteza, sendo, portanto, propagadas para as profundidades reduzidas, dando origem a IPT (Incerteza Propagada Total) (HARE, 1995; HARE et al., 2011; FERREIRA et al., 2016). A IPT é composta pelas componentes horizontal ou IHT (Incerteza Horizontal Total) e vertical ou IVT (Incerteza Vertical Total).

Para cada ordem de levantamento, a S-44 estipula valores máximo permitidos para a IHT e a IVT (Tabela 1). Em relação a IVT, o valor máximo permitido, ao nível de confiança de 95%, é dado por:

$$IVT_{max} = \pm\sqrt{a^2 + (b \cdot P)^2} \quad (1)$$

em que as constantes a e b são dadas pela Tabela 1 e variam de acordo com a ordem dos levantamentos hidrográficos. O termo P é a profundidade, assim, verifica-se que cada profundidade possuirá uma estimativa de incerteza.

Pode-se entender o termo a como uma parcela da incerteza que não varia com a profundidade, enquanto $b \cdot P$ representa a parcela que varia com profundidade. Em termos estatísticos, essas parcelas configuram-se, respectivamente, como efeitos sistemáticos e aleatórios da profundidade reduzida.

Tabela 1 – Resumo dos padrões mínimos para Levantamentos Hidrográficos.

Ordem	Especial	1a	1b	2
IHT máxima permitida. Nível de confiança de 95%	2 m	5 m + 5% da profundidade	5 m + 5% da profundidade	20 m + 10% da profundidade
IVT máxima permitida. Nível de confiança de 95%	a = 0,25 m b= 0,0075	a = 0,50 m b= 0,013	a = 0,50 m b= 0,013	a = 1,00 m b= 0,023

Fonte: Adaptado de IHO (2008).

Após computadas a IHT e a IVT de cada sondagem e se seguidos todos os procedimentos previstos em norma, pode-se classificar a ordem do levantamento de acordo com as especificações da S-44. Nesse caso, todas as profundidades do levantamento em questão devem ter IHT e IVT, expressas ao nível de confiança de 95%, iguais ou inferiores aos valores máximos permitidos. De modo contrário, o levantamento deve ser reclassificado, refeito ou as profundidades com incertezas superiores as tolerâncias estabelecidas em norma devem ser desconsideradas.

No entanto, uma propagação de covariâncias, apesar de considerar incertezas obtidas em todas as etapas de um levantamento hidrográfico, sejam elas sistemáticas ou aleatórias, estabelece apenas uma estimativa da qualidade do levantamento baseada nos possíveis desvios não correlacionados do sistema de sondagem (IHO, 2005; LINZ, 2010; Ferreira et al., 2015). Além disso, as incertezas utilizadas na computação da IPT são, em sua maioria, resultantes de testes de laboratório que não consideram as reais condições de operação (FERREIRA et al., 2016). Face ao exposto, uma problemática muito comum desse tipo de avaliação está ilustrada na Figura 1, onde, considerando uma profundidade de 10 metros, a sondagem, mesmo apresentando uma IVT tolerável, é, nitidamente, um *spike.*

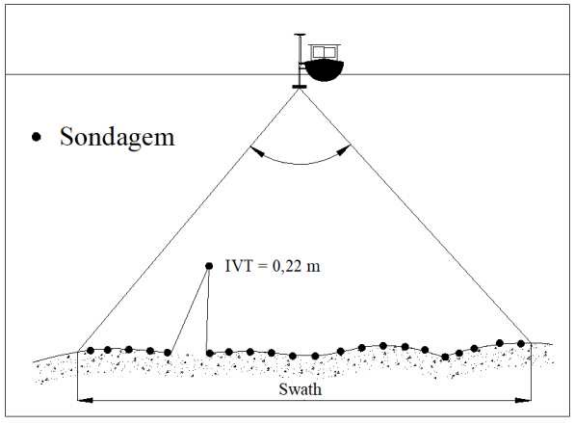

Figura 1 – Problemática da utilização apenas do modelo preditivo para determinação da incerteza.

Diante disso, é preferível que a avaliação da qualidade do dado batimétrico esteja fundamentada em informações redundantes, que permitam o cálculo da incerteza amostral. Entretanto, na prática é muito pouco provável que a mesma feição submersa seja reamostrada, uma vez que a coleta de dados é realizada com a embarcação em movimento. Mesmo na ocorrência de varreduras sobrepostas, tal como recomenda a NORMAN-25, o espaçamento entre feixes e o intervalo entre transmissões sucessivas do feixe sonoro tornam essa redundância praticamente impossível (CALDER & MAYER, 2003; CLAKE, 2014).

Diante dessa dificuldade, em levantamentos hidrográficos comumente realizam-se linhas de verificação planejadas de forma aproximadamente perpendicular às linhas regulares de sondagem. Nesse caso, adotam-se um espaçamento máximo não superior a 15 vezes o adotado para as linhas regulares de sondagem (DHN, 2014). As linhas de verificação devem ser sondadas de preferência em momentos distintos e em condições atmosféricas favoráveis. Se possível, utilizando embarcação e sistema de sondagem diferentes, o que reduz os efeitos sistemáticos. Considerando-se que os *spikes* e *tops* tenham sido eliminados ou minimizados durante a fase de limpeza dos dados, procede-se a avaliação da qualidade vertical do levantamento a partir da comparação entre as profundidades das linhas regulares de sondagem e das linhas de verificação (IHO, 2008; DHN, 2014; FERREIRA et al., 2015).

O uso de linhas de verificação é uma metodologia bastante utilizada para avaliar as profundidades coletadas por ecobatímetros monofeixe. Contudo, devido à enorme massa de dados gerada numa varredura multifeixe a fase de obtenção da amostra de discrepâncias necessita de uma abordagem diferente. É comum a criação de modelos digitais de profundidade das faixas sondadas através das linhas regulares e das linhas de verificação e, assim, efetuar a comparação entre os modelos batimétricos, *pixel* a *pixel,* visando obter um arquivo de discrepâncias (SUSAN & WELLS, 2000; SOUZA & KRUEGER, 2009; EEG, 2010). Porém, modelos digitais são resultantes de interpolações matemáticas e/ou estatísticas que, como sabido, possuem incertezas em suas estimativas (FERREIRA et al., 2013, 2015). Sendo assim, é notório que a análise da qualidade do levantamento hidrográfico ficaria comprometida.

Souza & Krueger (2009), apresentam um estudo em que a estimativa da incerteza amostral é realizada através da comparação entre modelos batimétricos. Nesse trabalho foi utilizado um sistema de batimetria multifeixe capaz de criar uma

expectativa de incerteza vertical de ±0,24 metros, mesmo assim, o menor intervalo da incerteza vertical amostral do levantamento hidrográfico, ao nível de confiança de 95%, variou entre ±0,305 metros, o que sugere a existência de componentes de incerteza não quantificados anteriormente, isto é, provavelmente o processo de interpolação.

Embora as normativas recomendem e exijam a execução de linhas de verificação, procedimentos para avaliação acerca da incerteza, de forma robusta, são raros. Diversos pacotes comerciais, tais como: *Caris-Hips, Hysweep (Hypack), QPS* e *PDS2000* possuem ferramentas para a comparação estatística das varreduras de verificação com a sondagem regular. Basicamente, os algoritmos realizam uma confrontação entre as profundidades provenientes das linhas de verificação com a superfície batimétrica gerada a partir das linhas de sondagem regular, calculando estatísticas como média, desvio-padrão, máximo e mínimo. Porém, sabe-se que essas medidas de localização possuem baixo grau de robustez, pois, dentre outros, não são capazes de quantificar a simetria ou assimetria da distribuição dos dados e são afetados, de forma exagerada, por valores extremos (MORETTIN & BUSSAB, 2004). Soma-se a isso, o fato das medidas serem tratadas como estimativas de quantidades populacionais e, nesse caso, a incerteza do estimador também deve ser fornecida, isto é, seria necessário construir intervalos de confiança.

Na prática, a principal estatística utilizada para classificar a ordem do levantamento é a diferença entre as profundidades ou discrepância. Na hipótese de 95% destas discrepâncias estar dentro da tolerância prevista na S-44 para a ordem requerida, é comum, entre a comunidade hidrográfica, considerar que determinado levantamento cumpre os requisitos de incerteza para ser classificado naquela ordem. Contudo, a avaliação baseada apenas na percentagem é insuficiente para validar e classificar o levantamento hidrográfico, sendo necessário desenvolver metodologias mais robustas. Além do mais, a distribuição das discrepâncias pode não ser simétrica.

É frequente, em vários ramos da ciência, assumir que as discrepâncias são livres de *outliers*, seguem uma distribuição normal e são variáveis aleatórias independentes e identicamente distribuídas (Höhle & Höhle, 2009). Tais pressupostos são assumidos, quase sempre, para justificar o emprego da estatística clássica (MORETTIN & BUSSAB, 2004). No entanto, sabe-se que essas hipóteses dificilmente são atendidas e/ou verificadas e, quando negligenciadas, comprometem as análises (LI et al., 2005; MAUNE, 2007; SANTOS, 2015; SANTOS et al., 2017).

Esses pressupostos também são comumente assumidos por normativas de levantamentos hidrográficos (IHO, 2008; DHN, 2014).

Susan & Wells (2000) propuseram um método para avaliar a incerteza vertical baseado na sobreposição de superfícies batimétricas gerados através de duas faixas sondadas. A partir das discrepâncias calculadas determina-se o valor do *RMSE (Root Mean Square Error)* que é, então, multiplicado por 1,96 visando obter um indicador de qualidade a um nível de confiança de 95%, tal como exigido pela S-44. Eeg (2010) utilizou um procedimento semelhante para estimar a incerteza do levantamento e então classificá-lo de acordo com as ordens previstas em norma. Porém, conforme exposto no Capítulo 2 deste texto, a definição dos intervalos de confiança descritos acima é equivocada e teoricamente inconsistente, mesmo assim, é largamente utilizada (GREENWALT & SCHULTZ, 1962; FGDC, 1998; SUSAN & WELLS, 2000; EEG, 2010; SEKELLICK & BANKS, 2010).

Conforme discutido, levantamentos hidrográficos realizados para fins de construção ou atualização de cartas náuticas, devem adotar como espaçamento entre as linhas regulares de sondagem a metade da largura de varredura. Nesses moldes, varreduras adjacentes irão se sobrepor. Assim, a geração de amostras de discrepâncias através dessas áreas de sobreposição, com posterior avaliação da qualidade vertical das profundidades coletadas, pode apresentar-se vantajosa, principalmente por permitir uma considerável redução da campanha batimétrica.

Deve-se atentar que linhas de verificação ou áreas de varredura sobrepostas não indicam acurácia absoluta, uma vez que os dados são coletados, quase sempre, a partir da mesma plataforma de sondagem e, neste caso, há um grande número de fontes de incertezas comuns em potencial entre os dados das linhas regulares e das linhas de verificação. Porém, quando uma faixa submersa é sondada novamente, seja por uma varredura de verificação ou varreduras adjacentes, é esperado que as profundidades reduzidas, calculadas a partir de observações eivadas apenas de efeitos aleatórios, distribuam-se em torno da profundidade real (desconhecida), assim, essas informações "redundantes" podem ser capazes de fornecer um bom indicador de qualidade vertical do levantamento, desde que os dados coletados recebam um tratamento estatístico coerente.

Face ao exposto, este trabalho objetiva, em um primeiro momento, propor e validar um novo método para extração de pontos homólogos de levantamentos hidrográficos realizados a partir de sistemas de sondagem por faixa, sem a necessidade

de recorrer a interpolações matemáticas e/ou estatísticas. Este método, denominado PP *(Point to Point)* foi submetido a uma minuciosa análise comparativa com as técnicas comumente utilizados entre a comunidade hidrográfica. As discrepâncias obtidas foram, então, tratadas através da metodologia MAIB (Metodologia para Avaliação da Incerteza de dados Batimétricos), apresentada no Capítulo 2, bem como por meio das análises tradicionais. Em todas as fases que demandaram uma investigação por *outliers*, recorreu-se ao AEDO (Algoritmo Espacial para Detecção de *Outliers*), metodologia proposta no Capítulo 1. Por último, vislumbrando eliminar a necessidade da realização de varreduras de verificação, o Método PP foi aplicado a sobreposições de sucessivas varreduras regulares de sondagem, gerando discrepâncias para posterior estimativa intervalar da incerteza vertical.

2. PROPOSIÇÃO DO MÉTODO

De acordo com a NORMAM-25, os levantamentos hidrográficos Categoria A devem cumprir os requisitos de incerteza estipulados pela S-44. Visando comprovar o cumprimento desses requisitos é comum entre a comunidade hidrográfica efetuar a análise da incerteza vertical do levantamento hidrográfico a partir de informações redundantes advindas das varreduras de verificação. Embora essa metodologia não forneça a acurácia absoluta do levantamento, ela é capaz de fornecer um bom indicador da qualidade das profundidades coletadas. Salienta-se que mesmo que o levantamento hidrográfico não necessite ter uma classificação junto à Marinha do Brasil, é de extrema importância que os dados sejam fornecidos sempre com atributos de incertezas.

Devido à enorme massa de dados gerada numa sondagem multifeixe ou através de sonares interferométricos, a fase de obtenção de amostras de discrepâncias necessita de uma abordagem diferente daquela utilizada nos levantamentos com ecobatímetro monofeixe. Neste estudo, são aplicadas três rotinas com objetivo de comparar as profundidades levantadas através das linhas regulares de sondagem e das linhas de verificação, a saber: ***Método SS (Surface to Surface)***, ***Método SP (Surface to Point)*** e ***Método PP (Point to Point).*** As duas primeiras abordagens são bastante conhecidas entre a comunidade hidrográfica e, por isso, de larga aplicação. O método PP, desenvolvido e proposto neste estudo, oferece diversas vantagens frente aos demais, a

principal delas reside no fato da sua aplicação não depender da utilização de interpoladores matemáticos e/ou geoestatísticos.

A Figura 2 ilustra o fluxograma empregado neste estudo para a geração das amostras de discrepâncias.

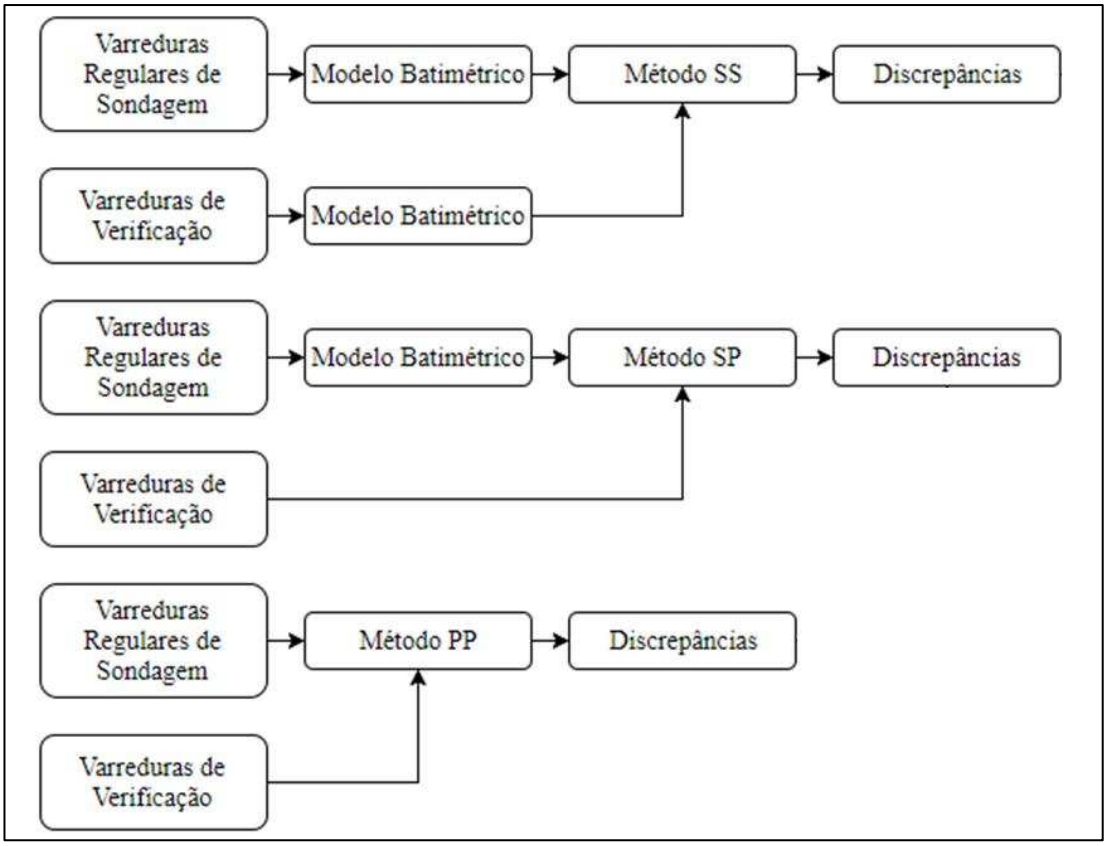

Figura 2 – Técnicas para obtenção das amostras de discrepâncias de levantamentos hidrográficos realizados a partir de sistemas de sondagem por faixa.

No método SS, são criados modelos batimétricos das varreduras regulares de sondagem e das varreduras de verificação. As profundidades armazenadas nas grades são comparadas, *pixel* a *pixel*, visando gerar o arquivo de discrepâncias, tal como realizado, por exemplo, por Susan & Wells (2000), Souza & Krueger (2009) e Eeg (2010). Nesse método, a quantidade de discrepâncias, bem como a qualidade das análises, é altamente dependente do tamanho do *pixel* utilizado na interpolação, isto é, da resolução do modelo batimétrico. Uma vez que os modelos digitais são resultantes de interpolações, inevitavelmente, as estimativas da qualidade vertical do levantamento hidrográfico são prejudicadas. Dessa forma, priorizando robustecer as análises, sugere-se que os modelos batimétricos sejam construídos com base na análise geoestatística, conforme recomendações de Ferreira et al. (2017).

Alternativamente ao método SS, pode-se empregar o método SP. Nesta técnica constrói-se uma superfície batimétrica apenas para as varreduras regulares de

sondagem, com isso, reduz-se as incertezas inerentes aos processos de interpolação. O modelo batimétrico gerado é comparado com as profundidades calculadas das linhas de verificação, gerando os arquivos de discrepâncias. Sendo assim, o tamanho amostral do arquivo de discrepâncias é proporcional à densidade da nuvem de pontos da varredura de verificação, enquanto a qualidade das análises permanece dependente da resolução do modelo batimétrico e, logicamente, dos dados coletados. A metodologia SP é utilizada, pela maioria dos pacotes comerciais de processamento de dados hidrográficos, como principal ferramenta de controle de qualidade vertical. Todavia, deve-se destacar que todos esses *softwares* geram a superfície batimétrica através da aplicação de interpoladores determinísticos, o que pode reduzir ainda mais eficiência das análises, visto que técnicas determinísticas são menos robustas que o interpolador Krigagem, conforme comprovado por Ferreira et al. (2017).

Em todos os casos, recomenda-se que as análises geoestatísticas sejam realizadas preferencialmente através do pacote geoR, desenvolvido por Ribeiro Júnior & Diggle (2001). Nos casos em que o volume de dados conduzir a um excessivo esforço computacional, pode-se recorrer ao *software* ArcGIS (ESRI, 2014).

Na terceira e última abordagem, o arquivo de discrepâncias é obtido através da comparação das varreduras regulares de sondagem e de verificação, sem recorrer a interpolações. Essa técnica, proposta neste trabalho, é chamada de método PP e oferece vantagens frente as demais, apresentando-se como um algoritmo bastante robusto e de fácil aplicação. Toda a parte inovadora foi implementada no *software* livre R (R Core Team, 2017) e pode ser consultada no Apêndice.

Num primeiro momento o algoritmo desenvolvido importa a faixa regular de sondagem (LRS) e a faixa de verificação (LV). Preferencialmente, os arquivos devem estar no formato *Shapefile* (ESRI, 2014), embora o algoritmo seja capaz de importar arquivos em formato de texto, convertendo-os, posteriormente, para o formato nativo do código implementado. Para isso, o usuário deve fornecer o sistema de projeção adotado. Em todos os casos, os arquivos devem conter, pelo menos, as coordenadas posicionais (locais, projetadas ou geodésicas) e as profundidades reduzidas.

Destaca-se que nessa fase, propositalmente, são lidos apenas dois arquivos por vez, o arquivo LRS e o arquivo LV, sugerindo que as faixas regulares de sondagem devam ser fornecidas no mesmo conjunto de dados (Arquivo *Shapefile* ou Arquivo de texto) e as análises das varreduras de verificação sejam realizadas separadamente, ou seja, um exame para cada varredura de verificação. Seguidamente, o usuário deve

fornecer duas informações: a ***Distância Limite*** e o ***Buffer***. A distância limite consiste na distância máxima na qual os pontos, pertencentes, respectivamente, a LRS e a LV, possam ser considerados homólogos. Idealmente, $Distância\ Limite = 0\ un$, ou seja, as sondagens somente são consideradas homólogas quando possuírem coordenadas planimétricas idênticas. O *buffer* aplica uma extrapolação nos limites da área de interseção entre as varreduras, o que permite que os pontos de borda não sejam, acidentalmente, excluídos durante as análises. O valor *default* é $Buffer = 0\ un.\ Un$ representa a unidade de medida das coordenadas, geralmente, metros.

Realizadas as devidas configurações, o código implementado seleciona todos os pontos considerados homólogos, dentro da distância limite, calculando as discrepâncias (dp_i) através da seguinte Equação:

$$dp_i = Z_i^{LV} - Z_i^{LRS} \quad (1)$$

em que Z_i^{LRS} e Z_i^{LV} são as profundidades coletadas nas varreduras regulares de sondagem e nas varreduras de verificação, respectivamente. De posse dos arquivos de discrepâncias pode-se proceder com a análise da qualidade vertical do levantamento hidrográfico.

É sabido que nas batimetrias multifeixe, para fins de construção ou atualização de cartas náuticas no âmbito levantamentos hidrográficos regulamentados pela NORMAM – 25, deve-se adotar como espaçamento entre as linhas regulares de sondagem a metade da largura de varredura (*swath* ou cobertura de fundo). Esse planejamento visa obter 100% de sobreposição, implicando na prática em uma ensonificação de 200% (Figura 3). Esse valor de sobreposição é o recomendado para um correto processamento dos dados coletados. Nessas situações, varreduras adjacentes irão se sobrepor em 50%. Diante disso, pode-se vislumbrar a utilização deas informações para avaliar a qualidade da sondagem batimétrica, seja por verificação do encaixe vertical e horizontal dos perfis gerados a partir das sucessivas varreduras, seja através da geração de amostras de discrepâncias.

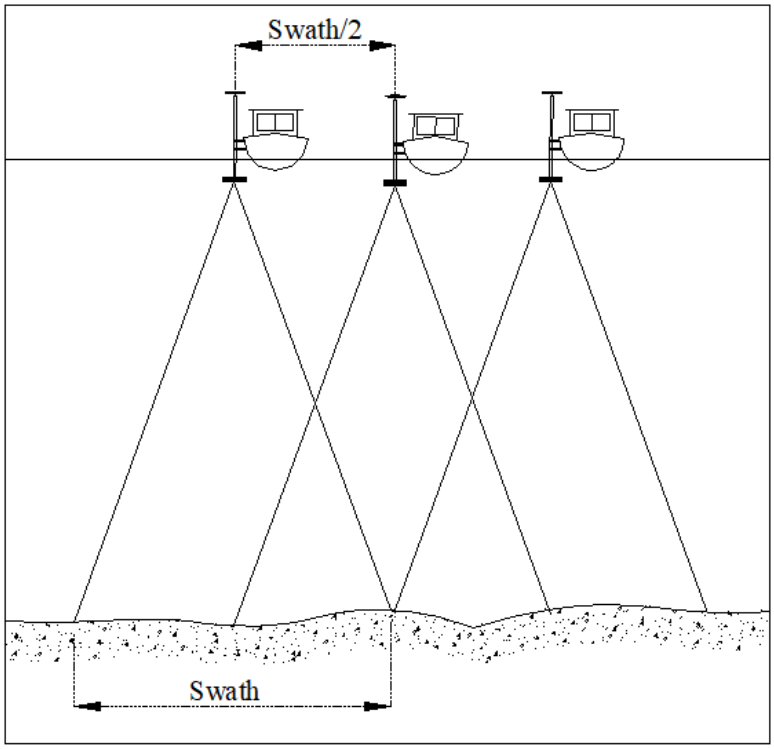

Figura 3 – Esquema do planejamento de um levantamento multifeixe com sobreposição entre linhas adjacentes de 100%.

A verificação do encaixe vertical e horizontal possui eficácia comprovada na avaliação da qualidade da sondagem. Sendo assim, ela é parte integrante do controle de qualidade de dados batimétricos coletados por sistemas de varredura (Ecobatímetros Multifeixe, Sonares Interferométricos, *LiDAR* batimétricos). Por outro lado, a estimativa da incerteza amostral por meio das sobreposições de faixas regulares de sondagem ainda não é explorada. Tal fato é justificado pela natureza do processo de medição de profundidade, em que as profundidades calculadas a partir de feixes externos possuem uma maior incerteza quando comparadas com profundidades advindas dos feixes nadirais. Cita-se, nesses casos, a forma cônica do feixe acústico, o que resulta em pegadas maiores para os feixes externos, diminuindo a resolução angular; a refração sofrida pela onda sonora, muito mais aparente para os feixes externos; as incertezas na medição do *roll*, que irão introduzir incertezas na profundidade e na posição, sendo maiores para os feixes externos, dentre outras. Devido a isso, as varreduras de verificação são, geralmente, sondadas com um *swath angle* reduzido. A Figura 4 ilustra o processo de formação de feixes de um sistema multifeixe.

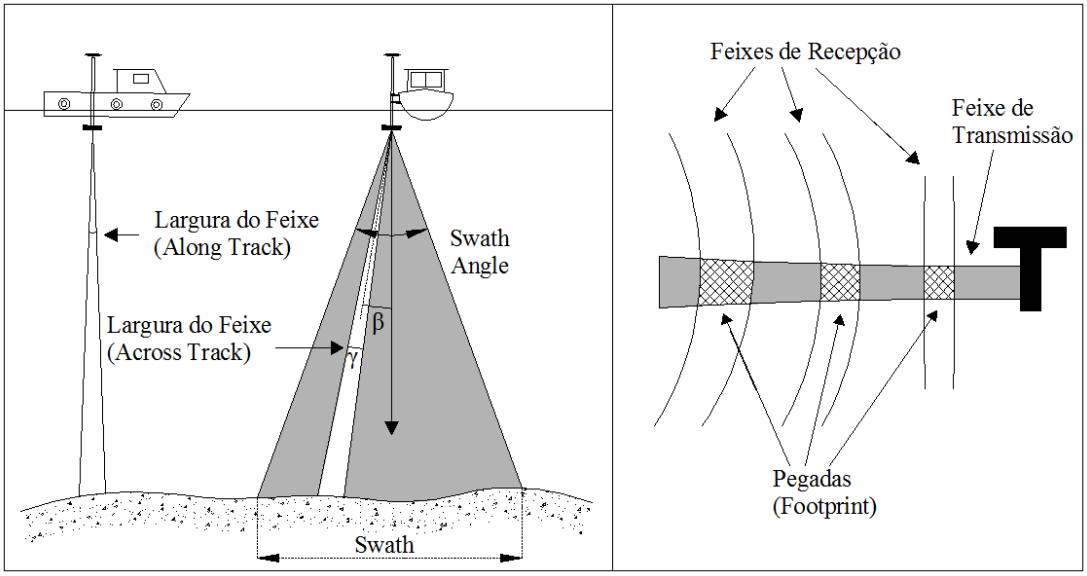

Figura 4 – Geometria do processo de formação de feixes de um sistema de sondagem multifeixe.

Contudo, não se verifica na literatura nenhum estudo que comprove a ineficiência da avaliação estatística a partir das sobreposições de faixas adjacentes. Diante disso, neste estudo são realizados testes objetivando quantificar a magnitude da diferença entre a avaliação através de varreduras de verificação e a estimativa da incerteza amostral através da sobreposição das varreduras regulares de sondagem. Nessa etapa será utilizado o método PP.

Obtido os arquivos de discrepâncias, são realizadas análises estatísticas com a finalidade de estimar o intervalo, com 95% de confiança, da incerteza vertical do levantamento batimétrico. Tais inferências estatísticas são baseadas na MAIB, com destaque para a verificação da presença de *outliers*, realizada através de uma adaptação da metodologia AEDO, isto é, o arquivo XYZ (profundidade georreferenciada), será substituído pelo arquivo XYdz (discrepância georreferenciada). Nessa fase, recomenda-se adotar o *Método* δ.

O fluxograma a seguir resume o funcionamento lógico da ferramenta desenvolvida. Destaca-se que a metodologia é fundamentada em teoremas básicos da estatística clássica e Geoestatística e encontra-se implementado no *software* livre R, incluindo toda a parte inovadora (R Core Team, 2017).

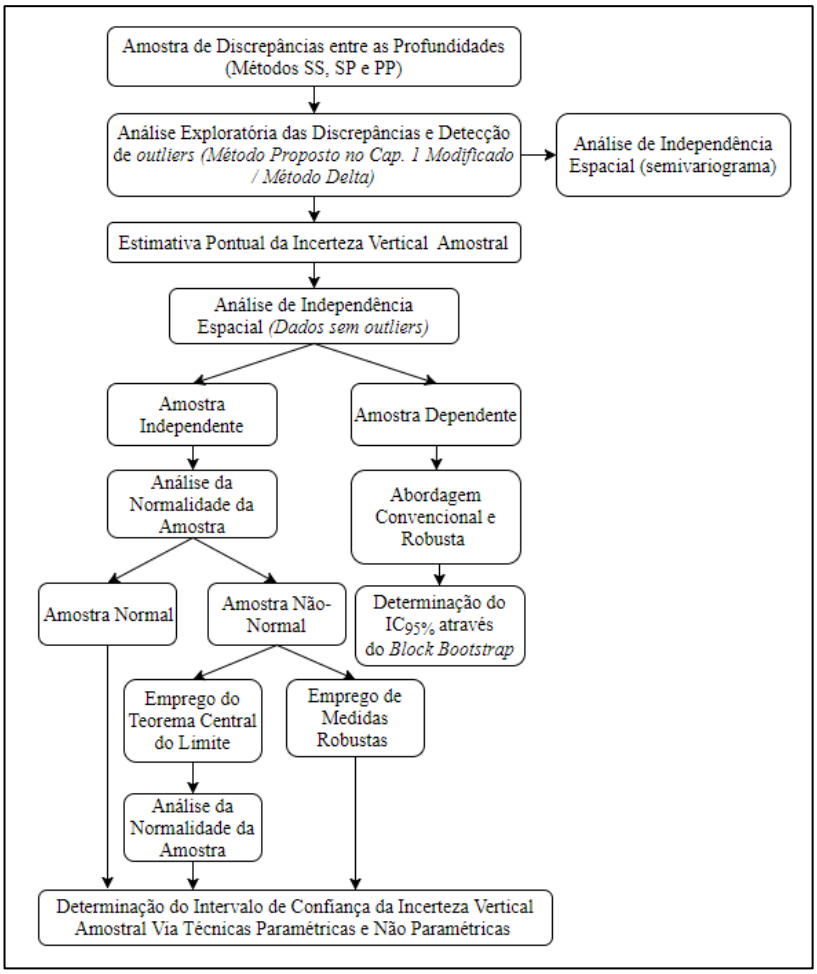

Figura 5 – Método proposto para avaliação intervalar da incerteza vertical em dados de sondagem por varrimento.

Apesar da MAIB apresentar-se bastante eficiente, é importante destacar que a análise estatística proposta não deve ser utilizada como única ferramenta para avaliação de sondagens batimétricas, principalmente, aquelas destinadas a confecção de cartas náuticas, uma vez que, a não existência de encaixe vertical dos perfis batimétricos gerados entre sucessivas varreduras regulares, desqualifica qualquer análise posterior.

3. EXPERIMENTOS E RESULTADOS

Os dados batimétricos que serviram de base para a aplicação das metodologias propostas foram obtidos a partir de uma parceria com a empresa *A2 Marine Solution*. O levantamento hidrográfico foi realizado em 18 de maio de 2017, na região do Porto do Rio de Janeiro, entre a Ilha das Enxadas e a Ilha das Cobras.

Os dados foram coletados por um sistema de sondagem por faixa, composto pelo ecobatímetro multifeixe modelo *Sonic 2022* da marca *R2* Sonic, integrado com o sistema inercial, modelo *I2NS (Integrated Inertial Navigation System)* da marca *Applanix*. Destaca-se que o planejamento, execução, bem como o processamento dos dados seguiram recomendações da NORMAM-25 e S-44 para a categoria A e ordem Especial, respectivamente. Os dados utilizados integram um levantamento hidrográfico que ainda se encontra em avaliação junto à DHN, assim, apenas parte dos dados sabidamente classificados, na categoria e ordem pretendidas, foram utilizados neste estudo.

Para o desenvolvimento desta pesquisa foram empregadas três varreduras regulares de sondagem adjacentes e quatro varreduras de verificação. As linhas foram primeiramente submetidas a um pré-processamento no *software Hysweep* (Hypack, 2012), que consistiu nas seguintes etapas:

- Conversão dos dados coletados pelos diversos sensores para o formato do *Hysweep*;
- Análise dos dados dos sensores auxiliares (atitude, latência, velocidade do som, maré, *etc*.), objetivando a identificação de possíveis falhas. Se necessário, interpolação ou rejeição de dados anômalos;
- Junção dos datagramas[8];
- Cálculo da *Total Propagated Uncertainty* (Horizontal e Vertical);
- Cálculo das coordenadas tridimensionais no formato XYZ (profundidades reduzidas georreferenciadas), e
- Detecção e eliminação de *spikes* através da metodologia proposta no Capítulo 1 *(Método* $\delta / c = 3$ *)*.

Posteriormente, as varreduras regulares de sondagem foram combinadas num único arquivo de pontos (XYZ), chamado, simplesmente, de LRS. Esse arquivo foi utilizado nas análises posteriores. A Tabela a seguir resume as informações gerais.

[8] Entidade de dados completa e independente. Nesse caso, refere-se aos dados gerados pelos diversos sensores.

Tabela 2 – Informações gerais dos dados da área de estudo.

ID da Linha	Nome	Tipo de Varredura	Quantidade de Pontos	AEDO – *Método* $\delta / C = 3$	
				% Spikes	Tempo Médio de Processamento
007_1733.HSX	LRS1	Regular	2.342.330	-	-
006_1723.HSX	LRS2	Regular	2.263.971	-	-
005_1713.HSX	LRS3	Regular	2.305.298	-	-
-	LRS	Regular	6.911.599	0,265	339 horas
001_1948.HSX	LV1	Verificação	777.998	0,001	89 horas
002_1954.HSX	LV2	Verificação	952.299	0,002	106 horas
003_2001.HSX	LV3	Verificação	809.704	0,004	97 horas
004_2007.HSX	LV4	Verificação	802.954	0,091	91 horas

A área de estudo pode ser visualizada na Figura 6.

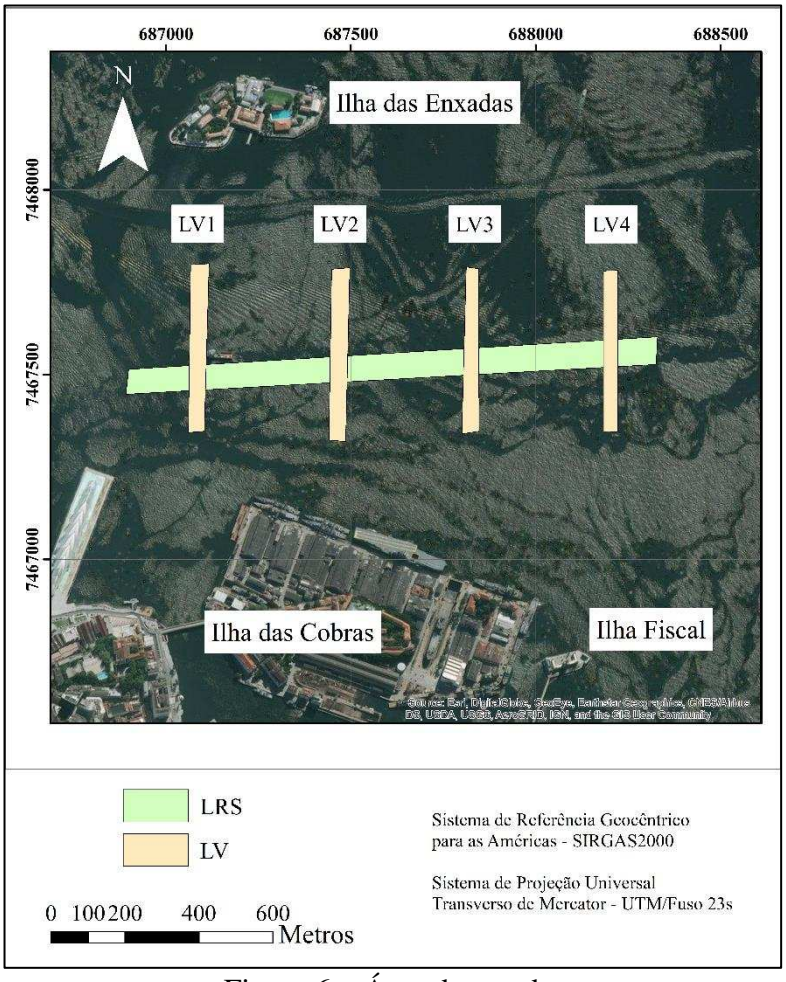

Figura 6 – Área de estudo.

De posse das coordenadas tridimensionais, aplicaram-se os métodos para obtenção dos arquivos de discrepâncias (Figura 2). Nesse sentido, num primeiro momento foi realizada a análise geoestatística dos conjuntos de dados visando gerar os modelos batimétricos. Buscando uma padronização, em todos os casos empregou-se a Krigagem Simples. O tamanho do *pixel* adotado foi de 50 centímetros, visto que, idealmente a resolução de um modelo batimétrico deve ser metade do tamanho do menor objeto que se pretenda detectar/representar. Nesse caso, para levantamentos hidrográficos de ordem Especial, em que é obrigatório a detecção de estruturas cúbicas com 1 metro de aresta, a resolução da superfície batimétrica deve ser superior a 0,5 metros (IHO, 2008; VICENTE, 2011).

Conforme recomendações de Viera (2000) e Ferreira et al. (2013), após a modelagem geoestatística, efetuou-se o processo de validação cruzada (autovalidação *leave-one-out*). A Tabela 3 sumariza os resultados obtidos. Os semivariogramas omnidirecionais experimentais e os modelos ajustados, bem como os relatórios de processamento, encontram-se no Apêndice. A variável profundidade apresentou dependência espacial para todos os conjuntos de dados, informação, inclusive, utilizada na aplicação do método para localização de *spikes* proposto no Capítulo 1 (Tabela 2).

Tabela 3 – Resultados da análise geoestatística.

Modelo Batimétrico	**Tamanho Amostral**	**Modelo Teórico**	**Parâmetro do Modelo**	**Validação Cruzada**			
				RMSE **(m)**	**Erro Médio (m)**	**a (m)**	**b (m)**
LRS	6.893.289	*Stable*	1,237	0,027	0	0,993	0,095
LV1	777.991	*Gaussiano*	-	0,018	0	0,999	0,001
LV2	952.281	*Stable*	2,000	0,017	0	0,999	0,001
LV3	809.672	*K-Bessel*	9,999	0,012	0	0,999	0,000
LV4	802.227	*Gaussiano*	-	0,013	0	0,999	0,008

Analisando a Tabela 3, pode-se concluir que os modelos batimétricos gerados, como esperado, possuem baixo grau de incerteza, uma vez que os valores de $RMSE$, Erro Médio e coeficiente "b" da regressão linear entre os valores observados e preditos, apresentaram valor nulo ou muito próximo de zero, enquanto o coeficiente "a" foi maior que 0,99 metros em todos os casos. Tais fatos contribuirão qualitativamente com

as análises posteriores. Nessa etapa, incluindo as intervenções e verificações do analista, foram gastos aproximadamente 26 horas.

Com os modelos batimétricos gerados, pôde-se dar continuidade a aplicação dos métodos. Conforme é ilustrado na Figura 6, a área de estudo possui quatro áreas de sobreposição, assim, em todos os casos, serão formados quatro arquivos de discrepâncias. No método SS, as grades batimétricas foram comparadas, *pixel* a *pixel,* gerando ao final os arquivos de discrepâncias, denominados dp1_ss, dp2_ss, dp3_ss e dp4_ss. Posteriormente, comparou-se o modelo batimétrico construído com base nas varreduras regulares com as profundidades calculadas através das sondagens das linhas de verificação, ao final obteve-se os arquivos de discrepâncias dp1_sp, dp2_sp, dp3_sp e dp4_sp. Para a aplicação do método PP, configurou-se a $Distância\ Limite = 0{,}01\ metros$ e o $Buffer = 0\ metros$. O método PP também foi aplicado as sobreposições das varreduras regulares de sondagem, resultando, ao final, nos arquivos dp1, dp2, dp3, dp4, dp5 e dp6 (Tabela 4).

Tabela 4 – Síntese dos resultados obtidos por meio da aplicação dos métodos PP, SS e SP.

Método	Varreduras	Nome do Arquivo	Tempo Médio de Processamento	Tamanho Amostral
PP	LRS X LV1	dp1	12,31 minutos	1.954
	LRS X LV2	dp2	13,37 minutos	2.058
	LRS X LV3	dp3	12,51 minutos	2.213
	LRS X LV4	dp4	12,39 minutos	2.187
	LRS1 X LRS2	dp5	21,45 minutos	18.597
	LRS2 X LRS3	dp6	21,33 minutos	18.822
SS	LRS X LV1	dp1_ss	-	29.792
	LRS X LV2	dp2_ss	-	30.400
	LRS X LV3	dp3_ss	-	30.400
	LRS X LV4	dp4_ss	-	28.576
SP	LRS X LV1	dp1_sp	-	229.594
	LRS X LV2	dp2_sp	-	253.126
	LRS X LV3	dp3_sp	-	259.435
	LRS X LV4	dp4_sp	-	261.204

Em seguida, os arquivos de discrepâncias foram submetidos a metodologia AEDO, proposta no Capítulo 1. Para isso, a metodologia desenvolvida foi modificada para investigar a variável discrepância, ao invés da profundidade reduzida. Evidencia-se que, na presença de autocorrelação espacial, efetuou-se análises geoestatísticas com vista a gerar os resíduos padronizados (RPs). Nesses casos, a pesquisa por discrepâncias anômalas foi realizada sobre os RPs, conforme descrito na metodologia AEDO. Durante a aplicação do AEDO, especialmente para o método SP, enfrentou-se problemas com tempo de processamento de máquina, talvez devido aos computadores serem utilizados simultaneamente também em outras tarefas. Por esse motivo, o tratamento dos pontos, em partes, foi realizado em blocos estratificados espacialmente. Em vista disto, a pesquisa por *outliers* nas bordas dos blocos foi realizada manualmente. Em todos os casos, adotou-se o limiar δ com constante $c = 3$. O raio de busca foi escolhido, para o método PP, como 3 vezes o valor da distância mínima e nos demais casos, adotou-se 3 vezes a resolução do modelo batimétrico, isto é, 1,5 metros. Os resultados dessa etapa estão sumarizados na Tabela 5. Detalhes podem ser consultados no Apêndice.

Tabela 5 – Resultados da detecção de *outliers* via metodologia AEDO.

Método	**Nome do Arquivo**	**Tamanho Amostral**	**Análise de Independência**	**Raio de Busca (m)**	**Tempo Médio de Processamento**	***% Outliers***
PP	dp1	1.954	Independente	1,712	3,53 minutos	0,921
	dp2	2.058	Independente	1,677	5,97 minutos	1,020
	dp3	2.213	Dependente	1,623	3,13 minutos	0,497
	dp4	2.187	Dependente	1,649	3,16 minutos	0,457
	dp5	18.597	Independente	1,517	16,21 horas	1,608
	dp6	18.822	Dependente	1,643	15,36 horas	0,462
SS	dp1_ss	29.792	Dependente	1,500	41,63 horas	0,117
	dp2_ss	30.400	Dependente	1,500	60,89 horas	5,296
	dp3_ss	30.400	Dependente	1,500	44,23 horas	9,365
	dp4_ss	28.576	Dependente	1,500	36,65 horas	0,511
SP	dp1_sp	229.594	Dependente	1,500	76,34 horas	3,111
	dp2_sp	253.126	Dependente	1,500	77,45 horas	10,629
	dp3_sp	259.435	Dependente	1,500	65,89 horas	20,085
	dp4_sp	261.204	Dependente	1,500	83,79 horas	9,328

A amostra dp1_ss apresentou a menor proporção de *outliers.* Das 29.792 discrepâncias, apenas 35 foram detectadas como anômalas. Por outro lado, a amostra dp3_sp, obteve uma porcentagem de contaminação por *outliers* acima dos 20%. No geral, o método PP refletiu em média uma menor proporção de *outliers*, seguido dos métodos SS e SP. Assim, nota-se que a quantidade de *outliers* é diretamente proporcional ao tamanho amostral do conjunto de dados, como esperado.

Após a eliminação das discrepâncias duvidosas, procedeu-se com o exame estatístico da qualidade vertical do levantamento hidrográfico, conforme método proposto (Figura 5). A primeira etapa consiste na análise exploratória das bases de dados, sintetizada na Tabela 6.

Tabela 6 – Estatística descritiva da área de estudo.

	Estatísticas	**dp1**	**dp2**	**dp3**	**dp4**	**dp5**	**dp6**
PP	Média (m)	-0,007	-0,026	-0,015	-0,006	0,022	0,016
	Mínimo (m)	-0,170	-0,230	-0,150	-0,140	-0,150	-0,170
	Máximo (m)	0,160	0,210	0,100	0,110	0,330	0,220
	Variância (m²)	0,0020	0,0018	0,0012	0,0011	0,0014	0,0017
	Coef. de Curtose	3,600	3,860	3,170	3,130	3,640	3,520
	Coef. de Assimetria	-0,010	-0,110	-0,060	-0,160	0,030	0,060
SS	Média (m)	0,051	0,025	0,563	0,538	-	-
	Mínimo (m)	-0,313	-0,382	-0,951	-0,322	-	-
	Máximo (m)	0,457	2,739	4,981	2,215	-	-
	Variância (m²)	0,0099	0,0958	2,1124	0,6001	-	-
	Coef. de Curtose	3,420	29,850	5,540	2,040	-	-
	Coef. de Assimetria	0,790	5,130	2,070	0,880	-	-
SP	Média (m)	-0,048	0,024	-0,173	-0,475	-	-
	Mínimo (m)	-0,263	-0,255	-3,725	-2,219	-	-
	Máximo (m)	0,213	0,293	0,226	0,361	-	-
	Variância (m²)	0,0080	0,0055	0,5554	0,5682	-	-
	Coef. de Curtose	2,500	5,330	15,350	2,550	-	-
	Coef. de Assimetria	-0,580	-0,640	-3,610	-1,120	-	-

Através da análise exploratória, obtém-se uma visão geral da distribuição amostral do conjunto de dados, podendo-se, ainda, verificar a presença de tendência

e/ou *outliers* não detectados, características que influenciam a aplicação dos métodos propostos, sobretudo a identificação e interpretação da dependência espacial através do semivariograma. Durante a análise exploratória, deve-se atentar para três aspectos principais: Tendência Central, Variabilidade e a Forma da distribuição.

A tendência central é caracterizada pelo valor típico da variável estudada, é utilizada para descrever o valor central da distribuição através de estatísticas como média, mediana ou moda. A variabilidade dos dados, geralmente quantificada pela variância amostral, fornece uma ideia da dispersão dos dados em relação à média. Existem diversas outras medidas de dispersão, como por exemplo, desvio padrão, amplitude, coeficiente de variação e Desvio Absoluto da Mediana Normalizado. Por fim, a distribuição de frequências pode assumir inúmeras formas, destacando-se três formas básicas que favorecem as análises propostas nesta tese, a saber: simétrica, assimétrica positiva e assimétrica negativa (MOOD et al., 1974; MORETTIN & BUSSAB, 2004).

Em termos de tamanho amostral, na área analisada o método SP proveu uma maior quantidade de discrepâncias, seguido dos métodos, SS e PP. Em ambos os casos, a quantidade de amostras apresentou-se adequada para os exames subsequentes. Embora, quanto maior a massa de dados, maiores são os tempos de processamento e análise requeridos. Observa-se na Tabela 6 que o método PP, no geral, exibiu os menores efeitos sistemáticos, ou seja, médias amostrais absolutas, seguido dos métodos SP e SS. Nota-se, ainda, que os métodos PP e SP apresentaram efeitos sistemáticos, com exceção da amostra dp2_sp, negativos, à medida que todas as amostras do método SS tiveram médias positivas. Ao comparar as amostras por método (PP, SS e SP), observa-se inconsistências estatísticas em todas as medidas de tendência central, exceto para as amostras dp2_ss e dp2_sp. Destaca-se que, na maior parte dos casos, o método SS apresentou as maiores discordâncias.

No que concerne a variabilidade dos dados, todas as amostras revelaram possuir uma alta variabilidade, principalmente, ao examinar as medidas de variância amostral (WARRICK & NIELSEN, 1980). O método PP apresentou, num quadro geral, a menor variabilidade dos dados, seguido, novamente, dos métodos SP e SS. Confrontando as bases de dados por método, analogamente ao observado para a medida de tendência central, nota-se inconsistências estatísticas, isto é, bases de dados que deveriam apresentar características, pelo menos, semelhantes, mostraram-se demasiadamente diferentes. Nesse caso, destaca-se o método SS, que proveu conjuntos

de dados com atributos bastante dessemelhantes daqueles apresentados pelos métodos PP e SP.

Quanto à forma das distribuições de frequências, as estatísticas descritivas revelam que o método PP conduziu a distribuições leptocúrticas e com maior grau de simetria. Por outro lado, os demais métodos apresentaram, com exceção para as amostras dp1_ss e dp2_sp, distribuições extremamente distorcidas. Isso pode, em alguns casos, prejudicar as análises posteriores. Mais uma vez, o método SS expôs, via de regra, resultados com propriedades diferentes daquelas apresentadas pelos métodos PP e SP. Diante dos fatos apresentados, pode-se concluir que o método SS pode não se apresentar confiável para os fins pretendidos deste trabalho, conforme será demonstrado adiante.

As informações discutidas podem ser visualmente confirmadas através do exame da Figura 7. Maiores detalhes são apresentados no Apêndice.

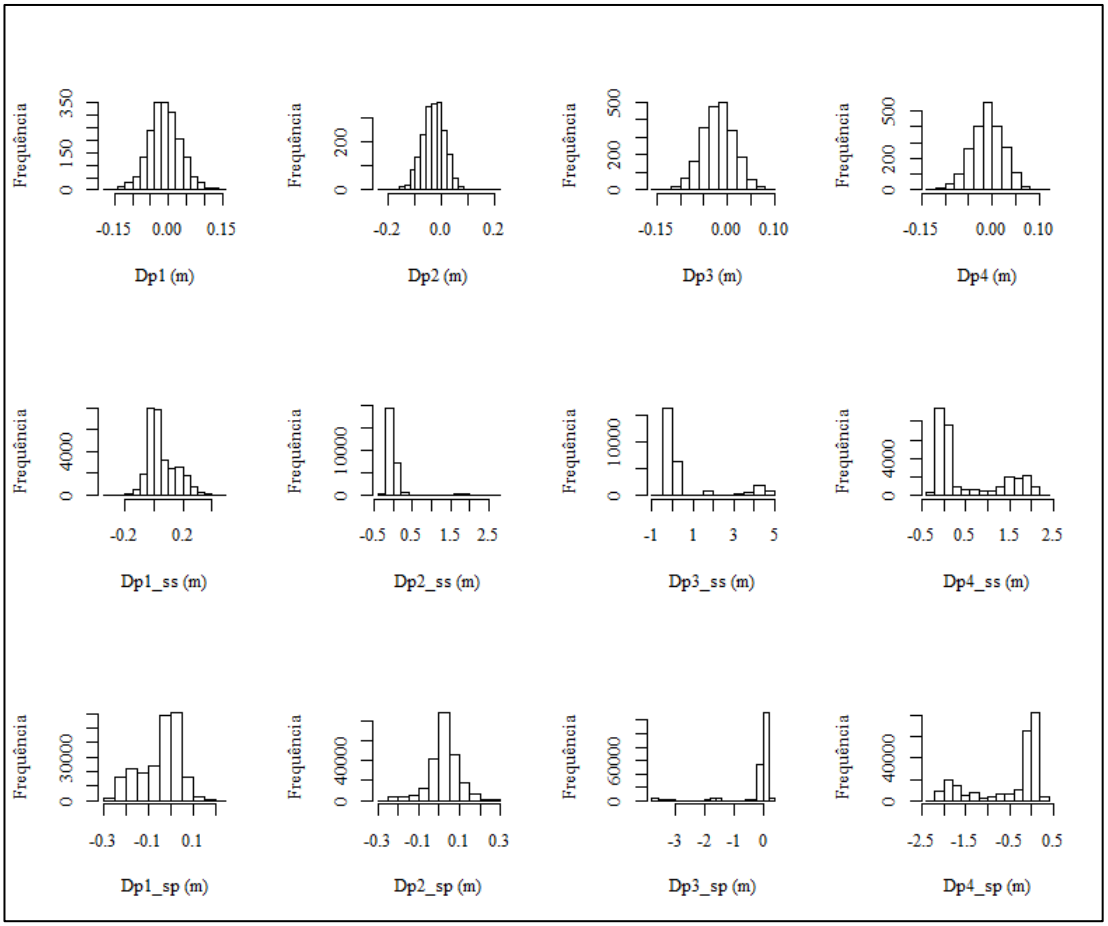

Figura 7 – Histograma das bases de dados analisadas.

Em tese, esperava-se que as amostras dp5 e dp6 apresentassem características semelhantes aos conjuntos de dados dp1, dp2, dp3 e dp4, uma vez que um dos objetivos deste estudo é validar o uso daquelas em alternativa a estas. Todavia, o efeito

sistemático apresentado pelas bases dp5 e dp6, diferentemente das demais, foi positivo. Embora, as magnitudes tenham sido equivalentes. Quanto ao tamanho amostral, as bases de dados dp5 e dp6 são, em média, 9 vezes maiores. Por outro lado, as medidas de variabilidade dos dados, bem como o formato das distribuições de frequência, mostraram-se, de certo modo, semelhantes (Figura 8).

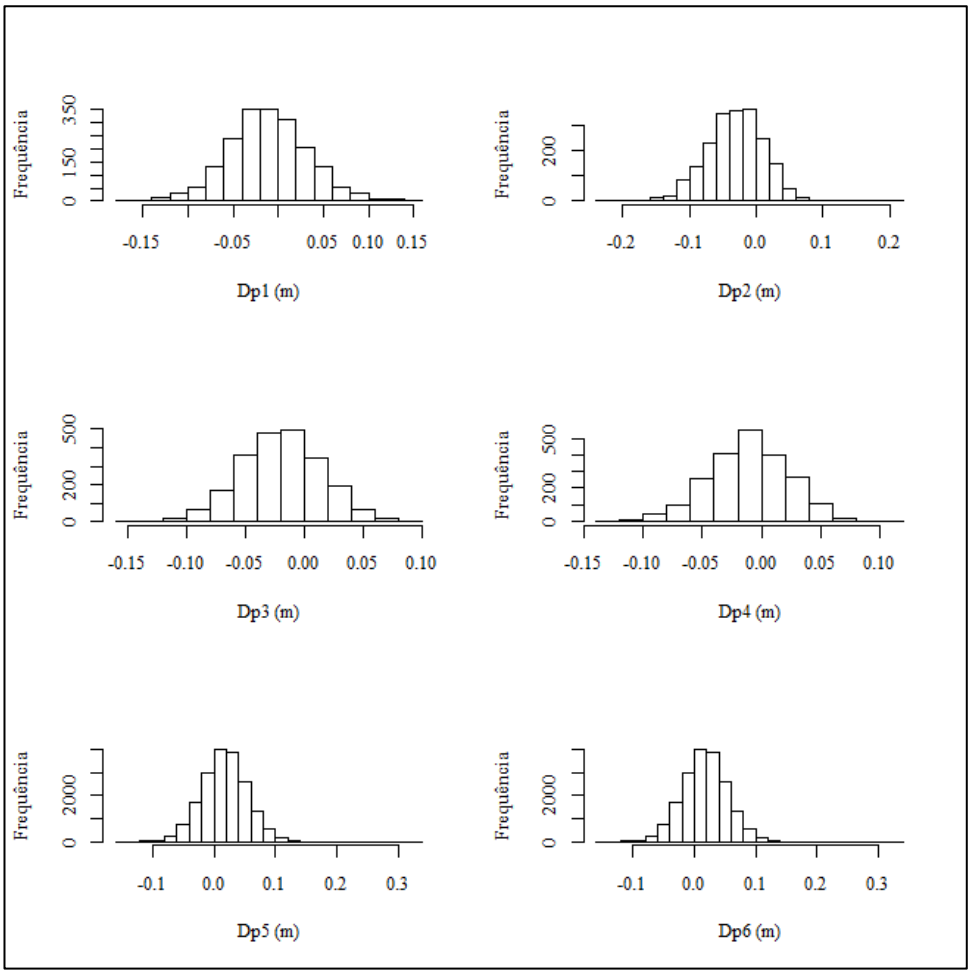

Figura 8 – Histograma das bases de dados geradas a partir do método PP.

Posteriormente a análise exploratória, pôde-se recorrer aos estimadores de incerteza amostral apresentados no Capítulo 2. Visando uma posterior comparação das metodologias propostas com aquelas tradicionalmente aplicadas, a Tabela 7 exibe um resumo da análise tradicional com base nas tolerâncias definidas pela S-44 para o levantamento em questão. Também é apresentado os valores de incerteza vertical amostral computadas com base nos estimadores $RMSE$ e $\Phi_{Robusta}$. Todas as estatísticas foram aplicadas aos dados com *outliers*, conforme comumente é praticado entre a comunidade hidrográfica.

Tabela 7 - Análise tradicional do Levantamento Hidrográfico e estimativa da $\Phi_{Robusta}$ (profundidade média: 15,600 metros).

Método	Nome do Arquivo	Intervalo de 95% de Tolerância para Ordem Especial (m)	Atende em %	$RMSE\ (m)$ (dados brutos)	$\Phi_{Robusta}\ (m)$ (dados brutos)
PP	dp1	[-0,276; 0,276]	99,95	0,048	0,046
	dp2		100,00	0,052	0,045
	dp3		100,00	0,038	0,031
	dp4		100,00	0,036	0,031
	dp5		99,98	0,045	0,036
	dp6		99,99	0,045	0,049
SS	dp1_ss	[-0,276; 0,276]	97,53	0,113	0,076
	dp2_ss		91,31	0,538	0,064
	dp3_ss		74,43	1,680	0,105
	dp4_ss		61,82	0,944	0,123
SP	dp1_sp	[-0,276; 0,276]	97,55	0,114	0,083
	dp2_sp		89,76	0,599	0,064
	dp3_sp		73,55	1,937	0,100
	dp4_sp		60,46	0,972	0,143

*O termo dados brutos refere-se a amostra de discrepâncias original.

Avaliando os resultados apresentados na Tabela 7, nitidamente verifica-se que através da aplicação do método PP (amostras dp1, dp2, dp3 e dp4), o levantamento seria classificado na ordem pretendida, isto é, Ordem Especial, visto que na média, 99,99% das discrepâncias estão abaixo da tolerância estipulada na S-44. Por meio do emprego do estimador $RMSE$ obteve-se, em média, uma incerteza vertical amostral de 0,043 metros. Como esta estatística é altamente influenciada por *outliers*, conforme sugerido no Capítulo 2, é preferível a utilização do estimador $\Phi_{Robusta}$. Nesse caso, a incerteza amostral média foi de 0,038 metros. Observa-se que a magnitude da diferença das incertezas estimadas é bastante sutil, fato que apenas comprova a qualidade dos dados coletados e do método PP. Embora as estimativas pontuais da incerteza amostral encontram-se dentro do intervalo de 95% de tolerância para a ordem especial, não é possível garantir a confiabilidade da estimativa, dado que os intervalos de confiança não foram construídos.

Ao aplicar a análise tradicional sobre os resultados gerados pelo método SS, contata-se que o levantamento hidrográfico não seria, *a priori,* classificado na ordem pretendida, já que apenas a amostra dp1_ss apresentou mais de 95% de discrepâncias dentro do intervalo de tolerância estipulado pela S-44. Em média, obteve-se somente 81,27% de aproveitamento. No que concerne a estimativa pontual da incerteza amostral, o $RMSE$ e o $\Phi_{Robusta}$, foram, na média, iguais a 0,819 metros e 0,092

metros, respectivamente. As magnitudes das diferenças entre as estimativas de incerteza ($RMSE$ e $\Phi_{Robusta}$) indicam que os *outliers* presentes nas amostras deste método estão mascarando a análise da qualidade vertical da batimetria. A maior diferença, 1,575 metros, ocorreu na amostra dp3_ss, seguido das amostras dp4_ss, dp2_ss e dp1_ss. Note que a análise baseada na porcentagem não reflete tais resultados, posto que naquele caso, a amostra dp4_ss apresentou a menor porcentagem de discrepâncias dentro do intervalo tolerável (61,82%). Evidencia-se que as estimativas pontuais médias, bem como todas as estimativas isoladas do $\Phi_{Robusta}$, isto é, por base de dados, enquadram o levantamento hidrográfico na Ordem Especial, em contrapartida, se o estimador $RMSE$ fosse utilizado como único indicador da qualidade vertical da batimetria, o levantamento não seria enquadrado na classe pretendida, uma vez que apenas a base dp1_ss obteve $RMSE$ dentro do intervalo tolerável. No entanto, essas afirmações tornam-se equivocadas, visto que não foram gerados intervalos de confiança para as estimativas pontuais.

Resultados com magnitudes semelhantes como as apontadas acima, foram encontrados a partir da avaliação das discrepâncias obtidas pelo método SP. De modo idêntico ao ocorrido para o método SS, apenas a amostra dp1_ss apresentou resultados, em termos de porcentagem, dentro do intervalo tolerável para Ordem Especial. Como, em média, somente pouco mais de 80% das discrepâncias encontram-se dentro do intervalo, de 95% de confiança, sugerido pela S-44, o levantamento não seria classificado na ordem pretendida. Referente as estimativas da incerteza pontual, novamente, os resultados foram equivalentes aqueles alcançados pelo método SS. O $RMSE$ e o $\Phi_{Robusta}$, foram, na média, iguais a 0,906 metros e 0,098 metros, respectivamente. Mais uma vez, a maior diferença ocorreu na amostra dp3_ss, seguido das amostras dp4_ss, dp2_ss e dp1_ss. As considerações acerca das estimativas pontuais são idênticas as apresentadas para o método SS.

Dada a natureza das análises realizadas e sintetizadas na Tabela 7, a única justificativa para os resultados divergentes entre os métodos utilizados para obtenção das amostras de discrepâncias, consiste no fato dos métodos SS e SP serem baseados no uso de interpoladores matemáticos, na prática, e geoestatísticos, nesta tese (Figura 2). Sendo assim, o método PP destaca-se e, por isso, é recomendado nas análises tradicionais de levantamentos hidrográficos. Nos casos em que os métodos SS e SP forem utilizados no exame tradicional, sugere-se que o estimador $\Phi_{Robusta}$, desenvolvido nesta tese, seja utilizado como indicador da qualidade vertical das

profundidades coletadas, em substituição a estatística $RMSE$ e/ou às análises baseadas simplesmente na porcentagem. A despeito do método PP, ambas as ferramentas de análise se mostraram, no âmbito geral, equivalentes.

Para a avaliação e, posterior, validação da estimativa da incerteza amostral por meio das sobreposições de faixas regulares de sondagem, foram estabelecidas as amostras dp5 e dp6. Em suma, pode-se afirmar que os resultados alcançados são significativos e bastante promissores. As amostras dp1, dp2, dp3 e dp4 apresentaram, em média, 99,99% das discrepâncias abaixo da tolerância estipulada na S-44, valor idêntico ao atingido pelas amostras dp5 e dp6. Examinando as amostras isoladamente, obtêm-se as mesmas conclusões. Em termos de estimativa pontual da incerteza amostral, o $RMSE$ e o $\Phi_{Robusta}$, produzidos pelas amostras dp5 e dp6, foram iguais a 0,045 metros e 0,043 metros, respectivamente. Resultados matematicamente equivalente àqueles obtidos através das sondagens de linhas de verificação. Ressalta-se que a amostra dp6, em particular, teve a estimativa do $\Phi_{Robusta}$ maior que o valor do $RMSE$. Tal fato, apesar de incomum, é possível de ocorrer e deve-se, principalmente, a natureza da distribuição de frequências do conjunto de dados.

Findada essa etapa, procedeu-se com continuação da aplicação da metodologia proposta (Figura 5). Assim, efetuou-se a construção de novos semivariogramas por meio dos dados sem *outliers*, constatando-se que os *outliers* que estavam eivando as amostras afetaram, nesta área de estudo, minimamente a análise de independência espacial, assim, não houve mudanças nas interpretações apresentadas na Tabela 5. Os semivariogramas experimentais, comprovando os entendimentos expostos, podem ser consultados no Apêndice.

No método PP, três amostras apresentam independência espacial, a saber: dp1, dp2 e dp5. Nesses casos, o teste de normalidade *Kolmogorov-Smirnov* (KS), ao nível de confiança de 95%, foi aplicado. Contudo, o emprego do teste KS revelou que as amostras não seguem normalidade (valor-p ≈ 0). Diante disso, num primeiro instante, recorreu-se a aplicação do Teorema Central do Limite, conforme a metodologia proposta na seção 2. O tamanho amostral dos agrupamentos foi definido com o valor mínimo sugerido, em tese, 4, enquanto a dissimilaridade foi quantificada pela distância "euclidiana". A partir dos agrupamentos formados, computou-se a média das discrepâncias para cada *cluster,* obtendo as amostras TCL, para cada base de dados (dp1, dp2 e dp5). Conforme afirma o TCL, a distribuição das médias tende a uma

distribuição normal, com a mesma média da amostra original e com a variância dividida pelo tamanho amostral dos agrupamentos, ou seja, $\sigma^2_{TCL} = {\sigma^2}/{4}$.

Visando comprovar as premissas do Teorema Central do Limite, aplicou-se o teste KS sobre as amostras TCL obtidas na etapa anterior. Constatando-se que as amostras geradas com base nos conjuntos de dados dp1 e dp2 apresentaram, como previsto, normalidade ao nível de confiança de 95%. Entretanto, a amostra dp5, mesmo após a transformação, não se mostrou normal. Em vista disso, foi aplicado a abordagem robusta com os intervalos de confiança construídos através da técnica *Bootstrap.*

Todas as demais amostras, incluindo aquelas obtidas através dos métodos SS e SP, apresentaram dependência espacial. Nessas circunstâncias, calculam-se a incerteza amostral através dos estimadores apresentados no Capítulo 2, ao mesmo tempo que os intervalos de confiança são gerados por meio da técnica *Block Bootstrap* (Seção 2.3, Capítulo 2). Os resultados obtidos encontram-se na Tabela 8.

Através do exame analítico e minucioso dos resultados apresentados na Tabela 8, pode-se concluir que o método PP se mostra mais coerente e efetivo que os demais métodos (SS e SP). Comparando, por exemplo, os valores de $\Phi_{Robusta}$ médios, computados a partir das bases de dados eivadas de *outliers* (Tabela 7) e das bases de dados sem *outliers* (Tabela 8), constata-se que no método PP essa diferença, como esperado, foi nula, enquanto que nos métodos SS e SP, foram, respectivamente, 0,008 metros e 0,024 metros. Ressalta-se que, teoricamente, as diferenças deveriam ser nulas, visto que a estatística $\Phi_{Robusta}$ foi concebida para tratar dados independentemente da presença, ou não, de dados discrepantes.

Tabela 8 – Intervalo da incerteza vertical amostral gerada com base na metodologia proposta.

Método	**Nome do Arquivo**	**Análise de Independência**	**Análise de Normalidade**	**Método para Estimativa do $IC_{95\%}$**	**$RMSE\ (m)$**		**$\Phi_{Robusta}\ (m)$**	
					Pontual	**$IC_{95\%}$**	**Pontual**	**$IC_{95\%}$**
PP	dp1	Independente	Não-Normal	TCL	0,061	[0,057; 0,067]	0,046	[0,042; 0,051]
	dp2	Independente	Não-Normal	TCL	0,068	[0,064; 0,075]	0,044	[0,039; 0,045]
	dp3	Dependente	Não aplicável	*Block Bootstrap*	0,038	[0,037; 0,039]	0,031	[0,029; 0,046]
	dp4	Dependente			0,034	[0,032; 0,035]	0,031	[0,029; 0,039]
	dp5	Independente	Não-Normal	*Bootstrap*	-	-	0,036	[0,029; 0,046]
	dp6	Dependente	Não aplicável	*Block Bootstrap*	0,044	[0,040; 0,047]	0,049	[0,047; 0,050]
SS	dp1_ss	Dependente	Não aplicável	*Block Bootstrap*	0,111	[0,100; 0,112]	0,076	[0,073; 0,077]
	dp2_ss	Dependente			0,311	[0,241; 0,461]	0,060	[0,058; 0,071]
	dp3_ss	Dependente			1,559	[1,468; 1,713]	0,081	[0,077; 0,104]
	dp4_ss	Dependente			0,943	[0,906; 0,963]	0,120	[0,112; 0,122]
SP	dp1_sp	Dependente	Não aplicável	*Block Bootstrap*	0,101	[0,092; 0,119]	0,077	[0,069; 0,091]
	dp2_sp	Dependente			0,078	[0,059; 0,080]	0,057	[0,044; 0,065]
	dp3_sp	Dependente			0,765	[0,743; 0,795]	0,061	[0,055; 0,076]
	dp4_sp	Dependente			0,891	[0,880; 0,900]	0,100	[0,092; 0,122]

Em termos de $RMSE$, os métodos PP, SS e SP apresentaram diferenças médias de, respectivamente, -0,007 metros, 0,088 metros e 0,447 metros. Nesse âmbito, esperava-se diferenças, no geral, mais significativas, uma vez que o estimador $RMSE$ é altamente influenciado por *outliers*. Atribui-se o ocorrido a alta qualidade dos dados coletados. O valor discrepante de -0,007 metros, obtido para o método PP, sugere que as estimativas obtidas através dos dados eivados de *outliers* estão menores que aquelas computadas com base na metodologia proposta. Fato confirmado e devido as estimativas alcançadas pelo emprego do TCL nas amostras dp1 e dp2, conforme será melhor discutido adiante.

As amostras dp1 e dp2 refletiram valores de $RMSE$ (Tabela 8), matematicamente, maiores que aqueles computados pelo método tradicional (Tabela 7), mesmo após a eliminação dos *outliers*. Isso deve-se a essência da metodologia aplicada, isto é, o Teorema Central do Limite. A partir da aplicação deste Teorema, percebe-se que a estimativa pontual da incerteza amostral é sempre, mesmo que de forma branda, superestimada. Em contrapartida, as estimativas dos intervalos de confiança são bastante coerentes e confiáveis. Resultados similares foram obtidos no Capítulo 2. Por outro lado, dada a complexidade analítica e computacional da aplicação do TCL, principalmente, para grandes conjuntos de dados, aliado aos valores de $\Phi_{Robusta}$ obtidos (Tabela 7 e Tabela 8), recomenda-se nesses casos a adoção da abordagem robusta.

Quanto as amostras dp3 e dp4, os resultados obtidos pelo método proposto foram equivalentes aqueles computados pelo método tradicional, com exceção, dos intervalos de confiança, extremamente importantes num contexto estatístico, porém gerados apenas na metodologia proposta. Enfatiza-se que, o fato de os valores de $RMSE$ apresentados na Tabela 7 e Tabela 8 serem iguais ou praticamente iguais, pode indicar falhas na etapa de detecção de *outliers*. Todavia, analisando os valores de $\Phi_{Robusta}$ é nítido que tais semelhanças advêm da alta qualidade dos dados aliados a robustez do método PP. Quanto a classificação perante as normativas, o levantamento hidrográfico, analisado com base nas metodologias propostas e desenvolvidas nesta tese, enquadra-se na Ordem especial/categoria A. Mesmos resultados alcançados através da análise tradicional.

No método SS, as estimativas de $RMSE$ referentes aos dados com e sem *outliers*, Tabelas 7 e 8, apresentaram diferenças significativas, isto é, centimétricas, apenas para as amostras dp2_ss e dp3_ss. No que concerne o estimador $\Phi_{Robusta}$, com

exceção da base de dados dp3_ss, todas as diferenças foram milimétricas ou, no caso da amostra dp1_ss, nula. Os intervalos de confiança construídos com recurso a método *Block Bootstrap* mostraram-se ótimos. Evidencia-se que, durante a geração dos intervalos de confiança para os métodos SS e, principalmente, SP, um alto esforço computacional foi requerido. Em todos os casos, adotou-se, *a priori,* 1.000 replicações *Bootstrap* e o tamanho da diagonal do bloco foi configurado com o valor aproximado do alcance, obtido por meio do semivariograma. Como previsto, em determinadas ocasiões, os intervalos com 95% de confiança mostraram-se inconsistentes. Nesses cenários, solucionou-se tal problemática apenas executando novamente o algoritmo *Block Bootstrap* implementado ou, em casos mais extremos, aumentando o número de replicações *Bootstrap.* No que se refere o enquadramento da sondagem na ordem pretendida, obteve-se os mesmos resultados da análise tradicional, isto é, apenas a abordagem robusta, na média e isoladamente, gerou estimativas intervalares, com 95% de confiança, capazes de classificar o levantamento na Ordem Especial e Categoria A.

Resultados menos promissores foram gerados por meio do método SP. Ao confrontar as estimativas de $RMSE$ (Tabela 7 e Tabela 8), obteve-se uma diferença média de quase meio metro, indicando, que o método SP produziu demasiados *outliers* que estavam mascarando a avaliação da qualidade vertical do levantamento hidrográfico. Isoladamente, todas as amostras apresentaram diferenças com magnitudes significativas. Já o emprego do estimador robusto revelou resultados mais promissores. Entretanto, quando comparado aos resultados obtidos através dos métodos SS e, especialmente, PP, conclui-se que o método SP pode não ser a melhor escolha, pois apresenta baixa eficiência.

Os intervalos de confiança, gerados através do método *Block Bootstrap*, mostram-se coerentes e confiáveis. Todavia, salienta-se a ocorrência dos mesmos problemas relatados durante o emprego do método SS. Por fim, no que tange a classificação junto a S-44, apenas o $\Phi_{Robusta}$ gerou resultados condizentes com a Ordem Especial. Em termos de estimativas do $RMSE$, apesar das bases dp1_sp e dp2_sp apresentarem intervalos, ao nível de confiança de 95%, dentro das tolerâncias estipuladas pela S-44, na média, isso não ocorreu, concluindo que o levantamento não pode ser, num todo, classificado na ordem e classe desejadas.

Diante dos fatos apresentados, primeiramente, conclui-se que a MAIB deve-se ser utilizada em substituição às análises tradicionais, embora, quando o método PP é empregado, a abordagem tradicional (Tabela 7) torna-se apropriada. Todavia, como os

intervalos de confiança não são construídos, estatisticamente, esses exames mostram-se pouco eficientes e, portanto, não são recomendados. Em contrapartida, na hipótese do emprego dos métodos SS ou SP, técnicas comumente empregadas na prática devido, principalmente, as facilidades computacionais e de diagnóstico, sugere-se recorrer a metodologia proposta associada à abordagem robusta para estimativa da qualidade vertical da sondagem batimétrica. Por fim, as análises realizadas neste capítulo, apontam que os métodos desenvolvidos são apropriados, acurados e consistentes.

Quanto a avaliação visando validar a estimativa do intervalo, ao nível de confiança de 95%, da incerteza amostral por meio das sobreposições de faixas regulares de sondagem, foram geradas, conforme já descrito, as amostras dp5 e dp6. A base dp5 apresentou independência amostral, porém, o tratamento através do TCL não gerou os resultados esperados. Nesse âmbito, para a estimativa da incerteza amostral, bem como os intervalos de confiança, recorreu-se a abordagem robusta. Tal fato reduziu a confiabilidade da validação do emprego do método proposto aliado a abordagem convencional. Todavia, analisando as estimativas de $RMSE$ (Tabela 7 e Tabela 8), nota-se que a amostra dp6 apresentou uma diferença de apenas 1 milímetro.

Em termos de $\Phi_{Robusta}$, essa diferença, para ambas as amostras dp5 e dp6, foi nula. A avaliação da qualidade da sondagem batimétrica através das linhas de verificação, exibiu uma incerteza pontual, calculada pelo estimador robusto, de 0,038 metros, enquanto a estimativa através das sucessivas varreduras regulares, foi de 0,043 metros, uma diferença de 5 milímetros. No que tange os intervalos de confiança, todos apresentaram amplitude estatisticamente equivalentes. Assim sendo, a avaliação através da sobreposição adjacente das varreduras regulares se mostrou uma alternativa viável, acurada e consistente, podendo, assim, ser empregada na prática.

Nos processamentos e análises realizada neste trabalho foram utilizadas 3 máquinas, com as seguintes configurações:

- Máquina 1: sistema operacional Windows 10, memória RAM de 8GB (parcialmente dedicada) e processador Intel® Core™ i7-4500U CPU @ 1,80GHz 2,40 GHZ;
- Máquina 2: sistema operacional Windows 10 Pro, memória RAM de 4GB (3,89GB utilizáveis) e processador Intel® Core™ i5-7400 CPU @ 3,00GHz 3,00GHz;

- Máquina 3: sistema operacional Windows 10 Home Single, memória RAM de 4GB (3,80GB utilizáveis) e processador Intel® Core™ i5-3337U CPU @ 1,80 GHz 1,80 GHz.

Enfatiza-se que, por diversas vezes, a mesma máquina foi requisitada na realização de análises simultâneas, o que influenciou nos tempos de processamento apresentados neste estudo.

4. CONCLUSÕES

Este capítulo teve como principal objetivo propor uma nova técnica para extração de profundidades homólogos coletadas através de sistemas de sondagem por varrimento, chamado neste trabalho de método PP. Este foi comparado aos métodos comumente utilizado entre a comunidade hidrográfica. Através de uma investigação minuciosa pode-se constatar, em todos os casos analisados, uma maior acurácia e consistência do método PP. Outra vantagem desse método consiste no fato do mesmo prover um menor esforço computacional, o que o torna a melhor opção para geração de amostra de discrepâncias.

Ao avaliar os dados de um levantamento hidrográfico, sabidamente, classificado na Categoria A/Ordem Especial, verificou-se que apenas as discrepâncias advindas do método PP, gerou resultados capazes de classificar o levantamento, pela análise tradicional, na categoria/ordem pretendidas. Os demais métodos, SS e SP, apenas refletiram a real qualidade dos dados batimétricos quando associados ao estimador robusto, proposto no Capítulo 2. No entanto, estas afirmações possuem apenas cunho exemplificativo, visto que as análises e, posterior, comparação com as tolerâncias previstas em norma, somente possuem sentido estatístico quando os intervalos de confiança são construídos.

Os resultados gerados por meio da aplicação do método proposto para avaliação da incerteza vertical amostral, permitiram concluir, novamente, a robustez do método PP frente às demais técnicas utilizadas neste estudo. Apenas a análise das discrepâncias, produzidas pela aplicação do método PP, possibilitou a classificação da batimetria na ordem e classe pretendidas. Para os métodos SS e SP, somente o exame das discrepâncias via abordagem robusta gerou estimativas intervalares, com 95% de confiança, capazes de classificar o levantamento na Ordem Especial.

Diante disso, obtêm-se conclusões importantes. Primeiramente, conclui-se que a MAIB, deve ser utilizada em substituição as análises tradicionais. Para geração das discrepâncias, sugere-se o emprego, preferencialmente, do método PP, dado que os demais métodos utilizados (SS e SP) mostraram-se ineficazes. Outra problemática desses métodos, reside no fato de eles produzirem uma maior quantidade de pontos homólogos, o que torna as análises posteriores morosas, principalmente, no que concerne os tempos de processamento de máquina. Soma-se a isso, ainda, o fato das técnicas SS e SP serem baseadas em modelos batimétricos, o que acrescenta incertezas ao processo. Nas ocasiões em que se empregou os métodos SS e SP, técnicas comumente utilizadas na prática, constatou-se que as análises tradicionais foram falhas, assim sendo, nesses casos, sugere-se que a MAIB, associada a abordagem robusta, deve ser utilizada nas estimativas intervalares da incerteza vertical amostral.

Outra observação importante diz respeito ao emprego do Teorema Central do Limite para geração de intervalos de confiança, na presença de independência e não-normalidade. A aplicação desse teorema, para grandes amostras, tal como aquelas tratadas neste capítulo, mostrou-se computacionalmente ineficiente e bastante complexa. Nesse sentido, permanece como sugestão o uso da abordagem robusta.

Outro objetivo importante e atingido por este trabalho, consiste na validação da análise da qualidade vertical do levantamento hidrográfico, através de discrepâncias provenientes das sobreposições de sucessivas varreduras regulares de sondagem. Tal conclusão é extremamente importante, pois demonstra que a realização de linhas de verificação mostra-se, de um modo geral, dispensável. Enfatiza-se que essa constatação ocorreu tanto na análise tradicional (Tabela 7), quanto no exame realizado por meio da metodologia de avaliação intervalar da incerteza proposta nesta tese (Tabela 8). Em termos gerais, conclui-se que os objetivos deste capítulo foram atingidos, visto que todos as metodologias desenvolvidas e propostas apresentaram-se apropriadas, acuradas e consistentes.

Por se tratarem de proposições inéditas, aperfeiçoamentos ainda são requeridos. Para trabalhos futuros, sugere-se a realização de melhorias nos algoritmos desenvolvidos com vista a reduzir o alto tempo de processamento de máquina. Maiores estudos sobre a utilização das sobreposições de sucessivas varreduras regulares para avaliação da qualidade vertical da sondagem também são recomendados.

REFERÊNCIAS BIBLIOGRÁFICAS

BJØRKE, J. T. & NILSEN, S. Fast trend extraction and identification of spikes in bathymetric data. **Computers & Geosciences**, v. 35, n. 6, p. 1061-1071, 2009.

CALDER, B. R. & MAYER, L. A. Automatic processing of high-rate, high-density multibeam echosounder data. **Geochemistry, Geophysics, Geosystems**, v. 4, n. 6, 2003.

CHS – Canadian Hydrographic Service. **Hydrographic survey management guidelines**. Canadian Hydrographic Service, Fisheries and Oceans Canada. 2013.

CLARKE, J. E. H. **Imaging and Mapping II: Submarine Acoustic Imaging Methods**. Notes of classes. Ocean Mapping Group. University of New Brunswick. 2014.

CRUZ, J.; VICENTE, J.; MIRANDA, M.; MARQUES, C.; MONTEIRO, C.; ALVES, A. Benefícios da utilização de sondadores interferométricos. **3as Jornadas de Engenharia Hidrográfica**. Instituto Hidrográfico Português, Lisboa, Portugal, 2014.

DHN – Diretoria de Hidrografia e Navegação. **NORMAM 25: Normas da Autoridade Marítima para Levantamentos Hidrográficos**. Marinha do Brasil, Brasil, 52p., 2014.

EEG, J. Multibeam Crosscheck Analysis: A Case Study. **The International Hydrographic Review**, n. 4, p. 25-33, 2010.

ESRI. **ArcGIS Desktop: Release 10.3**. Redlands, CA: Environmental Systems Research Institute. 2014.

FERREIRA, Í. O. ; RODRIGUES, D. D. ; SANTOS, A. P. Levantamento batimétrico automatizado aplicado à gestão de recursos hídricos. Estudo de caso: represamento do ribeirão São Bartolomeu, Viçosa-MG. **In: IV Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação**, 2012, Recife. Geotecnologias para o Planejamento e a Gestão Eficiente do território, 2012.

FERREIRA, Í. O.; RODRIGUES, D. D.; NETO, A. A.; MONTEIRO, C. S. Modelo de incerteza para sondadores de feixe simples**. Revista Brasileira de Cartografia**, v. 68, n. 5, p. 863-881, 2016.

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.; **Coleta, processamento e análise de dados batimétricos**. 1ª ed. Saarbrucken: Novas Edições Acadêmicas, v. 1, 100p., 2015.

FERREIRA, Í. O.; RODRIGUES, D. D.; SANTOS, G. R.; ROSA, L. M. F. In bathymetric surfaces: IDW or Kriging? **Boletim de Ciências Geodésicas**, v. 23, n. 3, p. 493-508, 2017.

FERREIRA, Í. O.; SANTOS, G. R.; RODRIGUES, D. D. **Estudo sobre a utilização adequada da krigagem na representação computacional de superfícies batimétricas**. Revista Brasileira de Cartografia, Rio de Janeiro, v. 65, n. 5, p. 831-842, 2013.

FGDC – Federal Geographic Data Committee. **National Standard for Spatial Data Accuracy**, **Part 3: National Standard for Spatial Data Accuracy.** Federal Geographic Data Committee: Reston, USA, 25p., 1998.

GREENWALT, C. R. & SCHULTZ, M.E. Principles of Error Theory and Cartographic Applications. **Aeronautical Chart and Information Center**: St. Louis, MO, USA, 98p., 1962.

HARE, R. Depth and position error budgets for multibeam echosounding. **The International Hydrographic Review**, v. 72, n. 2, p. 37-69, 1995.

HARE, R.; EAKINS, B.; AMANTE, C. Modelling bathymetric uncertainty. **The International Hydrographic Review,** n. 6, p. 31-42, 2011.

HÖHLE, J. & HÖHLE, M. Accuracy assessment of digital elevation models by means of robust statistical methods. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 64, n. 4, p. 398-406, 2009.

HYPACK, Inc. **Hypack – Hydrographic Survey Software User Manual**. Middletown, USA, 1784p., 2012.

IHO – International Hydrographic Organization. **C-13: IHO Manual on Hydrography**. Mônaco: International Hydrographic Bureau, 540p., 2005.

IHO – International Hydrographic Organization. **S-44: IHO Standards for Hydrographic Surveys**. Special Publication n. 44–5th. Mônaco: International Hydrographic Bureau, 36p., 2008.

LI, Z.; ZHU, Q.; GOLD, C. M. **Digital terrain modelling. Principles and methodology**. New York: CRC Press, 319p., 2005.

LINZ – Land Information New Zealand. **Contract Specifications for Hydrographic Surveys**. New Zealand Hydrographic Authority, V. 1.2, 111p., 2010.

MALEIKA, W. The influence of the grid resolution on the accuracy of the digital terrain model used in seabed modeling. **Marine Geophysical Research**, v. 36, n. 1, p. 35-44, 2015.

MAUNE, D. F. Digital Elevation Model Technologies and Applications: The DEM Users Manual. **American Society for Photogrammetry and Remote Sensing**, 2007.

MIGUENS, A P**. Navegação: a Ciência e a Arte.** Volume I - Navegação costeira, estimada e em águas restritas. Rio de Janeiro: DHN, Brasil, 538p., 1996.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the Theory of Statistics**. McGraw-Hill International, 577p.1974.

MORETTIN, P. A. & BUSSAB, W. O. **Estatística básica**. 5ª ed. São Paulo: Editora Saraiva, 526p., 2004.

OLIVA, J. A. B. Cenário Atual do Transporte Hidroviário Brasileiro. **5º seminário internacional em logística agroindustrial: O Transporte Hidroviário (Fluvial e Cabotagem) de Granéis Agrícolas**. ANTAQ – Agência Nacional de Transportes Aquaviários. 2008.

POMPERMAYER, F. M.; NETO, C. A. S. C.; PAULA, J. M. P. **Hidrovias no brasil: perspectiva histórica, custos e institucionalidade**. Texto para Discussão. IPEA – Instituto de Pesquisa Econômica Aplicada. Rio de Janeiro, 2014.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2017.

RIBEIRO JÚNIOR, P.J. & DIGGLE, P.J. **GeoR: a package for geostatistical analysis**. R-News. v.1, p. 15-18, 2001.

SANTOS, A. M. R. T.; SANTOS, G. R.; EMILIANO, P. C.; MEDEIROS, N. G.; KALEITA, A. L.; PRUSKI, L. O. S. Detection of inconsistencies in geospatial data with geostatistics. **Boletim de Ciências Geodésicas,** v. 23, n. 2, p. 296-308, 2017.

SANTOS, A. P. **Controle de qualidade cartográfica: metodologias para avaliação da acurácia posicional em dados espaciais**. Tese (Doutorado). Programa de Pós-Graduação em Engenharia Civil, Departamento de Engenharia Civil, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 172p., 2015.

SEKELLICK, A. J., BANKS, W. S. L. **Water Volume and Sediment Accumulation in Lake Linganore, Frederick County, Maryland**. Scientific Investigations Report 2010–5174. USGS - U.S. Geological Survey. 2010.

SOUZA, A. V. & KRUEGER, C. P. **Avaliação da qualidade das profundidades coletadas por meio de ecobatímetro multifeixe.** Anais Hidrográficos, Rio de Janeiro, n. 66, p. 90-97, 2009.

SUSAN, S. & WELLS, D. Analysis of Multibeam Crosschecks Using Automated Methods. **In: US Hydro 2000 Conference paper**, Biloxi, Mississippi. 2000.

VICENTE, J. P. D. **Modelação de dados batimétricos com estimação de incerteza**. Dissertação (Mestrado). Programa de Pós-Graduação em Sistemas de Informação Geográfica Tecnologias e Aplicações, Departamento de Engenharia Geográfica, Geofísica e Energia, Universidade de Lisboa, Portugal, 158p., 2011.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. In. NOVAES, R. F.; ALVAREZ V., V. H.; SCHAEFER, C. E G. R. **Tópicos em ciências do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, v.1. p. 2-54, 2000.

WARRICK, A.W. & NIELSEN, D.R. Spatial variability of soil physical properties in the field. In: HILLEL, D. **Applications of soil physics**. New York: Academic Press, p.319-344, 1980.

CONCLUSÕES GERAIS

A partir dos resultados obtidos nesta pesquisa, percebe-se que a aplicação dos métodos propostos contribuirá de sobremaneira com o processo de controle de qualidade no tratamento de dados de batimetria, uma vez que ambas as metodologias desenvolvidas se mostraram robustas e bastante versáteis para os fins a que se propõem.

No Capítulo 1 foi proposta uma metodologia, fundamentada em teoremas da estatística clássica e Geoestatística, para localização de *spikes* em dados batimétricos coletados a partir de sistemas de sondagem por faixa, denominada AEDO (Algoritmo Espacial para Detecção de *Outliers*). Toda a metodologia, incluindo a parte inovadora, foi implementado no *software* livre R. No desenvolvimento do AEDO, a Geoestatística apresentou-se importante, pois através dessa técnica foi possível incorporar a localização geográfica ao processo analítico dos dados. Para detecção das profundidades espúrias, o AEDO utiliza três limiares, dentre os quais, destaca-se o *Método* δ, também desenvolvido e apresentado por meio das pesquisas realizadas neste trabalho de tese. Esse limiar é apoiado por estatísticas robustas e leva em consideração a variabilidade espacial dos dados.

Através da aplicação da metodologia AEDO, principalmente aliada ao *Método* δ, pôde-se constatar a sua eficiência tanto na detecção de profundidades sabidamente caracterizadas como *spikes* quanto na não localização de dados referentes a estruturas submersas que, em análises tradicionais, poderiam ser erroneamente assinaladas como profundidades espúrias, provocando a exclusão de pontos que causam perigos ao navegante. Melhorias computacionais, no que tange o tempo de processamento de máquina, ainda são requeridas. Embora o foco tenha sido dado às sondagens multifeixe, o método pode ser aplicado em uma infinidade de áreas. Em trabalhos futuros, espera-se realizar testes e experimentos com objetivo de definir o P_{limiar}.

No segundo capítulo, foi proposto um método para avaliação intervalar da qualidade das profundidades coletadas por ecobatímetros monofeixe, abordando a normalidade e independência espacial da base de dados, características que afetam as análises, porém quase sempre são negligenciadas. A metodologia desenvolvida, intitulada MAIB (Metodologia para Avaliação da Incerteza de dados Batimétricos), foi baseada no trabalho de Santos (2015) e é fundamentada em teoremas básicos da

estatística clássica e Geoestatística. Visando facilitar as análises e estimativas, toda a parte inovadora foi implementada em ambiente R. Os exames realizados também permitiram o desenvolvimento de um novo estimador para calcular a incerteza vertical amostral dos dados. Esse estimador mostrou-se bastante eficiente, principalmente quando aplicado a conjuntos de dados sabidamente eivados de *outliers.*

A partir dos resultados alcançados, pôde-se perceber a importância de uma análise estatística coerente, verificando a presença de discrepâncias anômalas e analisando a normalidade e autocorrelação da amostra. Outra relevante conclusão deste capítulo, diz respeito a importância da apresentação das estimativas de incerteza sempre em conjunto com os respectivos intervalos de confiança, uma vez que, o simples fornecimento de uma medida pontual, não é capaz de descrever com clareza a qualidade dos dados. São desejáveis melhorias pontuais nos algoritmos implementados, principalmente, no que concerne o tratamento de grandes massas de dados.

No último capítulo, em alternativa as técnicas geralmente utilizadas pela comunidade hidrográfica, foi apresentada uma metodologia para extração de pontos homólogos de levantamento batimétricos realizados por meio de sistemas de sondagem por varrimento. O método é chamado de PP *(Point to Point)* e apresenta inúmeras vantagens frente as demais metodologias. As discrepâncias obtidas pelas técnicas usuais e pelo método PP foram, então, submetidas às análises tradicionais e à metodologia desenvolvida para avaliação da qualidade das profundidades coletadas num levantamento hidrográfico. Vislumbrando eliminar a necessidade da realização de varreduras de verificação, foi proposto o emprego do método PP sobre as áreas de sobreposição das varreduras regulares, objetivando gerar discrepâncias para posterior estimativa intervalar da incerteza vertical. Ao final, inúmeros exames comparativos foram realizados.

Diversas conclusões foram obtidas a partir dos experimentos realizados neste capítulo. A principal diz respeito a robustez do método PP para geração de discrepâncias que visam a avaliação da qualidade vertical do levantamento hidrográfico. Mesmo nas ocasiões em que a análise tradicional é utilizada, o método PP mostrou-se eficaz. Todavia, a análises tradicionais, isto é, baseadas na utilização de estimadores de acurácia teórica e/ou porcentagem de discrepâncias iguais ou inferiores aos valores máximos estabelecidos pela S-44, não são recomendadas, visto que a não verificação da presença de *outliers*, bem como a negligência quanto a

construção de intervalos de confiança, tornam estas análises, estatisticamente, equivocadas e incoerentes. De um modo geral, pode-se dizer que essas técnicas de avaliação são, até mesmo, obsoletas. Outra vantagem do método PP que merece destaque, é o baixo esforço computacional, o que o torna a melhor opção para obtenção das amostras de discrepâncias.

Durante a aplicação do demais métodos para extração de pontos homólogos, pôde-se constatar que a análise tradicional dessas discrepâncias é ineficaz, não somente pela indisponibilidade de estimativas intervalares, mas também pelos resultados inconsistentes que foram gerados. Por outro lado, quando se empregou a MAIB, associado ao estimador robusto, ambos desenvolvidos e apresentados no Capítulo 2, a análise das discrepâncias obtidas pelas técnicas usuais, mostrou-se coerente. Por outro lado, uma problemática inerente a esses métodos, geralmente utilizados na prática hidrográfica, consiste na grande quantidade de discrepâncias produzidas, o que aumenta o tempo de processamento e análise. O inevitável uso de modelos batimétricos por essas técnicas também introduz incertezas que afetam os resultados finais.

Outra conclusão importante refere-se à aplicação do Teorema Central do Limite, naqueles casos em que a base de dados apresentar independência espacial e não normalidade. Apesar do TCL ser teoricamente válido, a sua aplicação em grandes amostras, tal como aquelas tratadas no Capítulo 3, apresenta-se demasiadamente complexa e morosa, principalmente, no tempo de processamento. Assim sendo, recomenda-se nesses casos o uso da abordagem robusta.

No que tange a avaliação da possibilidade de análise vertical do levantamento hidrográfico, via sobreposição de sucessivas varreduras regulares de sondagem, os resultados alcançados mostraram-se muito promissores. Em todos os casos, as avaliações por meio de varreduras de verificação apresentaram-se estatisticamente iguais as estimativas obtidas pelas análises através das sobreposições de faixas adjacentes. Essa conclusão é extremamente importante, pois demonstra que linhas de verificação podem ser dispensáveis. Em vista disso, haverá uma redução da duração da campanha batimétrica e, consequentemente, dos tempos de processamento. Contudo, sugere-se que pesquisas mais detalhadas e em outras áreas de estudo sejam realizadas. Para trabalhos futuros, sugere-se também a execução de melhorias nos algoritmos desenvolvidos com vistas a reduzir o alto tempo de processamento de máquina.

Baseado nas exposições supra apresentadas, pode-se concluir que o trabalho atingiu os objetivos propostos, gerando resultados bastante significativos para o controle de qualidade dos levantamentos hidrográficos. Em suma, este estudo resultou no desenvolvimento da metodologia AEDO; de um novo limiar para detecção de *outliers* em conjuntos de dados georreferenciados; de uma nova metodologia para avaliação intervalar da incerteza vertical de sondagens batimétricas, denominada MAIB; de um estimador robusto e inovador para o cálculo da incerteza amostral e de um novo método para extração de pontos homólogos de levantamentos hidrográficos realizados por meio de sistema de sondagem por varrimento, intitulado PP. Foi apresentado também um estudo que validou o emprego das discrepâncias, resultantes das áreas de sobreposição de faixas de sondagem regular, na avaliação da qualidade vertical do levantamento hidrográfico.

No futuro, vislumbra-se a implementação das metodologias propostas nesta tese em um *software livre*. Por hora, enfatiza-se que todos os códigos desenvolvidos se encontram disponíveis no Apêndice.

APÊNDICES

Capítulo 1

a) Algoritmo da metodologia AEDO

```
#####################################################################
#Controle de qualidade em levantamentos hidrográficos
#Prof. ITALO O FERREIRA / UFV / DEC / EAM
#Prof. JÚLIO CÉSAR DE OLIVEIRA / UFV / DEC / EAM
#Artigo 01 – Tese
#Metodologia AEDO - Algoritmo Espacial para Detecção de Outliers
#####################################################################
#Obs: Para a aplicação correta deste algoritmo, favor consultar o texto da tese.

 #Caminho dos dados
 setwd("C:/Users/Ítalo/Documents")

 getwd()

 #Lista de pacotes a serem usados no Script
 pkg <- c("geoR","moments","rgeos","tcltk2",
        "raster", "rgdal", "ggplot2","plyr", "robustbase")

 sapply(pkg, require, character.only=TRUE)

#####################################################################
#Leitura da base de dados (arquivo shp ou txt)
#####################################################################

#Leitura dos dados no formato shp
 options(digits = 12)
 library(rgdal)
 data.shape<-readOGR(dsn=".",layer="DADOS")
 summary(data.shape)

 #Leitura dos dados no formato txt
 txt <- read.table("DADOS.txt", sep=";", dec=".", header = TRUE)
 coordinates(txt) <- ~X+Y
 #Definido sistema de projeção do arquivo txt
 proj4string(txt) = CRS("+proj=utm +zone=23 +south +ellps=GRS80 +units=m
+no_defs")

#####################################################################
#Atenção!
#####################################################################
#Análise realizada no arquivo shp
 aux<-data.shape@data
```

```
#Análise realizada no arquivo de texto:
  aux<-txt
#########################################################################

  #Gerar arquivo para análise de independência
  write.table(aux, "dados_para_semivariogram.txt", dec=",")

  #Leitura dos dados para análise de independência
  dados <- read.geodata("dados_para_semivariogram.txt", header=T,
dec=",",coords=1:2, data.col=3)
  names(dados)
  dados

#########################################################################
#Análise exploratória
#########################################################################

  (res=summary(dados))

  attach(res)

  # Principais medidas
  (med = round(mean(aux$Z),3))
  (min = round(min(aux$Z),3))
  (max = round(max(aux$Z),3))
  (des = round(sd(aux$Z),4))
  (var = round(var(aux$Z),4))
  (curt = round(kurtosis(aux$Z),2))
  (assim = round(skewness(aux$Z),2))
  (n=length(aux$Z))
  (dist.min= round((distances.summary[1]),3))
  (dist.max= round((distances.summary[2]),3))

  #Exportando informações:
  sink("Resultados.txt", type="output", append=T)
  cat("##### Tese de doutorado #####\n Prof. Ítalo O. Ferreira\n
italo.ferreira@ufv.br\n\n Metodologia AEDO - Algoritmo Espacial para Detecção de
Outliers
      \n Análise Exploratória dos Dados:","\n",
      "-----------------------------------------------------","\n",
      n, "observações"    ,"\n",
      "Média:"      ,med,'metros'  ,"\n",
      "Mínimo:"  ,min,"metros"  ,"\n",
      "Máximo:"     ,max,"metros"  ,"\n",
      "Variância:"  ,var,"metros²"  ,"\n",
      "Desvio Padrão:"     ,des,"metros"  ,"\n",
      "Coef. de Curtose: "         ,curt    ,"\n",
      "Coef. de Assimetria:"      ,assim  ,"\n",
      "dist.min:"  ,dist.min,"metros"  ,"\n",
```

```
    "dist.max:" ,dist.max,"metros" ,"\n",
    "-----------------------------------------------------","\n",
    fill=F)
  sink()
  shell.exec("Resultados.txt")

  #Gráficos para análise exploratória

  windows(8,6,title="Gráficos para análise exploratória")
  par(mfrow=c(2,2), family="serif")
  hist(aux$Z, xlab="Profundidades (m)", ylab= "Frequência", main=" Histograma")
  plot(density(aux$Z), xlab="Profundidades (m)", ylab= "Frequência", main="
Densidade")
  boxplot(aux$Z, ylab= "Profundidades (m)", main="Boxplot (Tukey)")
  qqnorm(aux$Z, xlab="Quantis Teóricos", ylab= "Quantis Amostrados", main="
Normal Q-Q Plot")
  qqline(aux$Z,lty=2, col='red')
  par(mfrow=c(1,1), family="serif")

  windows(8,6,canvas="snow2",title="Profundidades (m)")
  ggplot(aux, aes(x = X, y = Y, colour = Z)) + geom_point()+
    xlab("E (m)") + ylab("N (m)") + ggtitle("Área de Estudo") +
    theme_bw()+theme(plot.title = element_text(hjust = 0.5))

 ######################################################################
  #Análise de Independência (semivariograma)
 ######################################################################

  #Semivariograma empírico

  #Construção de 3 semivariogramas:
  #1º com alcance igual a 75% da distância máxima
  #2º com alcance igual a 50% da distância máxima
  #3º com alcance igual a 25% da distância máxima


  windows(8,6,title="Semivariograma Omnidirecional")
  escala.y=2*var

  par(mfrow=c(2,2), family="serif")
  vario.emp.1 <- variog(dados,max.dist=(dist.max), direction="omnidirectional")
  plot(vario.emp.1,ylim=c(0,escala.y),xlab="Distâncias (m)",ylab="Semivariâncias
(m²)", main=("75% da Distância Máxima"))
  abline(var(dados$data),0, col="gray60", lty=2, lwd=2)
  legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')

  vario.emp.1 <- variog(dados,max.dist=(0.75*dist.max),
direction="omnidirectional")
  plot(vario.emp.1,ylim=c(0,escala.y),xlab="Distâncias (m)",ylab="Semivariâncias
(m²)", main=("50% da Distância Máxima"))
```

```
  abline(var(dados$data),0, col="gray60", lty=2, lwd=2)
  legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')

  vario.emp.1 <- variog(dados,max.dist=(0.50*dist.max),
direction="omnidirectional")
  plot(vario.emp.1,ylim=c(0,escala.y),xlab="Distâncias (m)",ylab="Semivariâncias
(m²)", main=("25% da Distância Máxima"))
  abline(var(dados$data),0, col="gray60", lty=2, lwd=2)
  legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')


  dist.max
  0.75*dist.max
  0.50*dist.max
  0.25*dist.max

#########################################################################
 #Semivariograma omnidirecional das discrepâncias/Envelope de Monte Carlo
#########################################################################

  M<-(0.60*dist.max) #Semivariograma das discrepâncias para distância de M metros.

  windows(8,6,title="Semivariograma Omnidirecional")
  par(mfrow=c(1,1), family="serif")
  vario.emp.1 <- variog(dados,max.dist= M, direction="omnidirectional")
  plot(vario.emp.1,  ylim=c(0,escala.y),xlab="Distâncias  (m)",ylab="Semivariâncias
(m²)", main=("Semivariograma das Profundidades"))
  abline(var(dados$data),0, col="gray60", lty=2, lwd=2)
  legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')

  vario.env <- variog.mc.env(dados, obj.v=vario.emp.1)
  plot(vario.emp.1, env=vario.env,ylim=c(0,escala.y),xlab="Distâncias
(m)",ylab="Semivariâncias (m²)", main=("Semivariograma das Profundidades"))
  abline(var(dados$data),0, col="gray60", lty=2, lwd=2)
  legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')


  #Exportando informações:
  sink("Resultados.txt", type="output", append=T)
  cat(" Resultados do cálculo do Semivariograma:","\n",
     "-----------------------------------------------------","\n",
     vario.emp.1$n.data, "observações"    ,"\n",
     "Distâncias:"      , vario.emp.1$u  ,"\n",
     "Semivariâncias:"  ,vario.emp.1$v  ,"\n",
     "Número de pares em cada lote:"     ,vario.emp.1$n  ,"\n",
     "Desvio padrão de cada lote:"  ,vario.emp.1$sd,"\n",
     "Distância máxima:"     ,vario.emp.1$max.dist,"\n",
     "Direção:"         ,vario.emp.1$direction  ,"\n",
     "-----------------------------------------------------","\n",
     fill=F)
```

```
 sink()
 shell.exec("Resultados.txt")

#####################################################################
#Amostra independente, continuar a análise...
#Amostra dependente, efetuar análise geoestatística e trabalhar com RPs.
#####################################################################

#####################################################################
#Determinação do Raio de Busca: 3x a distância mínima
#####################################################################

#####################################################################
#Atenção!
#####################################################################
#Análise realizada no arquivo shp
 aux1<-data.shape
#Análise realizada no arquivo de texto
 aux1<-txt
#####################################################################

 #Raio de cada buffer
 dist <- gDistance(aux1, byid=T) # aplica para todos os pontos
 dist[which(dist==0)] <- NA
 raio <- 3* mean (apply(dist,1, min, na.rm=T))
 #raio <- 3* dist.min #análise geoestatística
 raio

 #Raio inserido pelo usuário
 raio  <- as.numeric(readline("Informe o valor do raio que será utilizado para analisar
os dados :"))
 raio

#####################################################################
#Constante do método Delta:
#1 para relevos irregulares ou canais artificiais (variabilidade alta).
#2 para relevos ondulados (variabilidade média)
#3 para relevos planos (variabilidade baixa)
#####################################################################

 c <- 1

 #Aplicar técnicas de detecção de outliers
 out_box_ALL <- out_MAD_ALL<- out_sd_ALL<- NULL
 count_ptsInBuffer <- NULL
 aa <- Sys.time()

 for (i in 1:dim(aux1)[1])
 {
  ptsInBuffer <- NULL
```

```
   out_box <- out_MAD <- out_sd<-NULL
   bufferedPoints <- gBuffer(aux1[i,],width=raio,byid=TRUE)
   bufPolygons <- bufferedPoints@polygons
   bufSpPolygons <- SpatialPolygons((bufPolygons))
    plot(bufSpPolygons)
    points(aux1)
   bufSpPolygonDf                                                                    <-
SpatialPolygonsDataFrame(bufSpPolygons,bufferedPoints@data)
   crs(bufSpPolygonDf) <- crs(aux1)
   ptsInBuffer <- which(gIntersects(aux1, bufSpPolygonDf, byid=TRUE)==TRUE)
   count_ptsInBuffer <- c(ptsInBuffer, count_ptsInBuffer)
   #Número mínimo para efetuar a análise de outlier
   if (length(ptsInBuffer)>7)
   {

     print(i)

    # ======= BOXPLOT AJUSTADO =========
    #identificação de outlier para cada buffer pelo BOXPLOT AJUSTADO


    out_box <- ptsInBuffer[which(aux1@data$Z[ptsInBuffer]==
                          unique(adjbox(aux1@data$Z[ptsInBuffer],
plot=FALSE)$out))]
    if (sum(!is.na(out_box)) != 0) {out_box_ALL <-  c(out_box_ALL, out_box)}

    #Fim análise pelo boxplot ajustado

    # ======= Z-SCORE MODIFICADO =========

    #identificação de outlier para cada buffer pelo Z-SCORE MODIFICADO
    dz <- aux1@data$Z[ptsInBuffer]

    ZSM = abs((0.6745*(dz-median(dz)))/(mad(dz,constant = 1)))

    treshold <- 3.5

    out_MAD <- ptsInBuffer[(ZSM > treshold)]

    if (sum(!is.na(out_MAD)) != 0)
    {out_MAD_ALL <-  c(out_MAD_ALL, out_MAD)}
    #Fim análise pelo z-score modificado

    # ======= Método Delta =========

    if (mad(aux1@data$Z) > mad(aux1@data$Z[ptsInBuffer]))
    {
     lim <- 0.5*(mad(aux1@data$Z)+mad(aux1@data$Z[ptsInBuffer]))
    }
    if (mad(aux1@data$Z) <= mad(aux1@data$Z[ptsInBuffer]))
```

```
    {
     lim <- mad(aux1@data$Z)
    }

    treshold_low <-   median(aux1@data$Z[ptsInBuffer])- c * lim          # limiar
inferior
    treshold_high <-   median(aux1@data$Z[ptsInBuffer])+c * lim          # limiar
superior

    out_sd <- ptsInBuffer[(aux1@data$Z[ptsInBuffer] < treshold_low)
                    | (aux1@data$Z[ptsInBuffer] > treshold_high)]

    if (sum(!is.na(out_sd)) != 0)
    {out_sd_ALL <-  c(out_sd_ALL, out_sd)}

   }
  }

  Sys.time() - aa

  #Analisar e localizar os outliers

  # ======= BOXPLOT AJUSTADO =========
  if (sum(!is.na(out_box_ALL)) != 0)
  {
    n_ocorrencia_BOX                                                    <-
merge.data.frame(count(count_ptsInBuffer),count(out_box_ALL), by = c("x","x"))
    n_ocorrencia_BOX                <-                cbind(n_ocorrencia_BOX,
round(n_ocorrencia_BOX[,3]/n_ocorrencia_BOX[,2],2))
    colnames(n_ocorrencia_BOX) <- c("ponto","freq_lido", "freq_out", "percentual")
# aplicando nome nos atributos da tabela

    #Probabilidade do dado ser um spike
    p <- 0.5

    pts_eliminados                                                      <-
n_ocorrencia_BOX$ponto[which(n_ocorrencia_BOX$percentual>=p)]
    aux1_BOX <- aux1[-pts_eliminados,]
    length(n_ocorrencia_BOX$ponto[which(n_ocorrencia_BOX$percentual>=p)])

  }

  if (sum(!is.na(out_box_ALL)) == 0)
  {aux1_BOX <- aux1}

  windows(8,6,title="Spikes detectados pelo Boxplot Ajustado")
  par(mfrow=c(1,1), family="serif")
  #Plotar base de dados sem outlier
  plot(aux1_BOX, pch=3, col=4,
       xlab="E (m)", ylab= "N (m)", main=" Spikes Detectados pelo Boxplot Ajustado")
```

```
 #Plotar os outliers
 points(aux1[pts_eliminados,], pch=19, col=2)
 legend('bottomleft',legend=c('Pontos Batimétricos','Spikes'),col=c(4,
2),pch=c(3,19))

 # ======= Z-SCORE MODIFICADO =========

 if (sum(!is.na(out_MAD_ALL)) != 0)
 {
    n_ocorrencia_MAD <-
merge.data.frame(count(count_ptsInBuffer),count(out_MAD_ALL), by = c("x","x"))
    n_ocorrencia_MAD <- cbind(n_ocorrencia_MAD,
round(n_ocorrencia_MAD[,3]/n_ocorrencia_MAD[,2],2))
    colnames(n_ocorrencia_MAD) <- c("ponto","freq_lido", "freq_out",
"percentual")  # aplicando nome nos atributos da tabela

    #Probabilidade do dado ser um spike
    p <- 0.8

    pts_eliminados <-
n_ocorrencia_MAD$ponto[which(n_ocorrencia_MAD$percentual>=p)]
    aux1_MAD <- aux1[-pts_eliminados,]
    length(n_ocorrencia_MAD$ponto[which(n_ocorrencia_MAD$percentual>=p)])

 }

 if (sum(!is.na(out_MAD_ALL)) == 0)
 {aux1_MAD <- aux1}

 windows(8,6,title="Spikes Detectados pelo Z-Score Modificado")
 par(mfrow=c(1,1), family="serif")
 #Plotar base de dados sem outlier
 plot(aux1_MAD, pch=3, col=4,
    xlab="E (m)", ylab= "N (m)", main=" Spikes Detectados pelo Z-Score
Modificado")
 #Plotar os outliers
 points(aux1[pts_eliminados,], pch=19, col=2)
 legend('bottomleft',legend=c('Pontos Batimétricos','Spikes'),col=c(4,
2),pch=c(3,19))

 # ======= MÉTODO DELTA =========

 if (sum(!is.na(out_sd_ALL)) != 0)
 {
    n_ocorrencia_sd <-
merge.data.frame(count(count_ptsInBuffer),count(out_sd_ALL), by = c("x","x"))
    n_ocorrencia_sd <- cbind(n_ocorrencia_sd,
round(n_ocorrencia_sd[,3]/n_ocorrencia_sd[,2],3))
    colnames(n_ocorrencia_sd) <- c("ponto","freq_lido", "freq_out", "percentual")  #
aplicando nome nos atributos da tabela
```

```
    #Probabilidade do dado ser um spike
    p <- 0.5

    pts_eliminados <-
n_ocorrencia_sd$ponto[which(n_ocorrencia_sd$percentual>=p)]
    aux1_sd <- aux1[-pts_eliminados,]
    length(n_ocorrencia_sd$ponto[which(n_ocorrencia_sd$percentual>=p)])

  }

  if (sum(!is.na(out_sd_ALL)) == 0)
  {aux1_sd <- aux1}

  windows(8,6,title="Spikes Detectados pelO Método Delta")
  par(mfrow=c(1,1), family="serif")
  #Plotar base de dados sem outlier
  plot(aux1_sd, pch=3, col=4,
      xlab="E (m)", ylab= "N (m)", main=" Spikes Detectados  pelo Método Delta")
  #Plotar os outliers
  points(aux1[pts_eliminados,], pch=19, col=2)
  legend('bottomleft',legend=c('Pontos Batimétricos','Spikes'),col=c(4,
2),pch=c(3,19))


  #Salvar tabelas para análise
  write.table(n_ocorrencia_BOX, "Tab_boxplot_Ajustado.txt", dec=",")
  write.table(n_ocorrencia_MAD, "Tab_Z_Scores.txt", dec=",")
  write.table(n_ocorrencia_sd, "Tab_3_Delta.txt", dec=",")

  #Comparação da concordância entre os spikes detectados por cada limiar
  Repetidos <- function(z,b)
  {BZ <- unique( c(z, b))
  return(((abs(length (BZ) - length(c(z, b))))/ (min (length(z), length(b))))*100)
  }

  #Gerar arquivo de dps sem outliers no formato txt (X, Y, Z, dz)
  write.table(aux1_BOX, "dados_semout_boxplot_ajustado.txt", dec=",")

  write.table(aux1_MAD, "dados_semout_ZSM.txt", dec=",")

  write.table(aux1_sd, "dados_semout_delta.txt", dec=",")

  #Gerar arquivo de dps sem outliers no formato shp (X, Y, Z, dz)
  writeOGR(aux1_BOX, dsn=".", layer="dados_semout_boxplot_ajustado",
driver="ESRI Shapefile")

  writeOGR(aux1_MAD, dsn=".", layer="dados_semout_ZSM", driver="ESRI
Shapefile")
```

```
writeOGR(aux1_sd, dsn=".", layer="dados_semout_delta", driver="ESRI Shapefile")
```

Capítulo 2

a) Algoritmo da MAIB

```
#####################################################################
#Controle de qualidade em levantamentos hidrográficos
#Prof. ITALO O FERREIRA / UFV / DEC / EAM
#Prof. JÚLIO CÉSAR DE OLIVEIRA / UFV / DEC / EAM
#Artigo 02 e 03 - Tese
#MAIB – Metodologia para Avaliação da Incerteza de dados Batimétricos
#####################################################################
#Obs: Para a aplicação correta deste algoritmo, favor consultar o texto da tese.

 #Caminho dos dados
 setwd("C:/Users/Ítalo/Documents")

 getwd()

 #Lista de pacotes a serem usados no Script
 pkg <- c("geoR","moments","scatterplot3d","tcltk2",
        "sp", "rgdal", "ggplot2" , "cluster" ,
        "bootstrap", "plyr", "robustbase", "MBESS", "rgeos", "gstat")

 sapply(pkg, require, character.only=TRUE)

 #Leitura dos dados
 dados <- read.table("discrepâncias.txt", header=T, dec=",")
 names(dados)
 dados
 length(dados$dz)

 #Leitura dos dados para análise Geoestatística
 dados1 <- read.geodata("discrepâncias.txt", header=T, dec=",",coords=1:2,
data.col=4)
 names(dados1)
 dados1
 length(dados1$data)
 dup.coords(dados1)

#####################################################################
#Análise NORMAM-25 / S-44
#Estimativa da Tolerância NORMAM-25 / S-44 --> Profundidade Média
#####################################################################

 #Cálculo da tolerância
 esp <- sqrt((0.25^2)+((0.0075*dados$Z)^2)) #Ordem Especial
 A <- sqrt((0.50^2)+((0.013*dados$Z)^2)) #Ordem 1A e Ordem 1B
 B <- sqrt((1.0^2)+((0.023*dados$Z)^2)) #Ordem 2
```

```
#Ordem Especial
aux <- round(((sum (abs(dados$dz)<=esp)/length(dados$dz))*100),2)
#Ordem 1A ou 1B
aux1 <- round(((sum (abs(dados$dz)<=A)/length(dados$dz))*100),2)
#Ordem 2
aux2 <- round(((sum (abs(dados$dz)<=B)/length(dados$dz))*100),2)

#cálculo da tolerância usando a profundidade média
esp1<- sqrt((0.25^2)+((0.0075*mean(dados$Z))^2)) #Ordem Especial
A1 <- sqrt((0.50^2)+((0.013*mean(dados$Z))^2)) #Ordem 1A e Ordem 1B
B1 <- sqrt((1.0^2)+((0.023*mean(dados$Z))^2)) #Ordem 2

#Exportando informações:
sink("Resultados1.txt", type="output", append=T)

cat("##### Tese de doutorado #####\n Prof. Ítalo O. Ferreira\n
italo.ferreira@ufv.br\n\n MAIB - Metodologia para Avaliação da Incerteza de dados
Batimétricos
    \n Avaliação NORMAM-25 / S-44:","\n",
    "-----------------------------------------------------","\n",
    "Ordem Especial: "        ,aux ,"%"   ,"\n",
    "Ordem 1A/1B:"            ,aux1,"%"  ,"\n",
    "Ordem 2:"                ,aux2,"%"  ,"\n",
    "-----------------------------------------------------","\n\n",
    "Tolerância NORMAM-25 / S-44","\n",
    "Intervalo de Confiança de 95%","\n",
    "Profundidade Média (m):", round(mean(dados$Z),3),"\n",
    "-----------------------------------------------------","\n",
    "Ordem Especial (m):" , "[",round(-1*esp1,3),";",round(esp1,3),"]" ,"\n",
    "Ordem 1A/1B (m):"    , "[",round(-1*A1,3),";",round(A1,3),"]" ,"\n",
    "Ordem 2 (m):"        , "[",round(-1*B1,3),";",round(B1,3),"]" ,"\n",
    "-----------------------------------------------------","\n",
    fill=F)
sink()
shell.exec("Resultados1.txt")

#Plotando análise de acordo com a NORMAM-25/S44
Ordem <- NULL
s44 <- dados
for(i in 1:length(dados$dz)) {

  if (s44$dz [i] >= -1*esp1 & s44$dz [i] <= esp1){

    Ordem[i] <- "Especial"

  }

  else if (s44$dz[i] >= -1*A1 & s44$dz[i] < -1*esp1 | s44$dz[i] <= A1 &  s44$dz[i]
> esp1){
```

```
    Ordem[i] <- "1A/1B"
   }

   else if (s44$dz[i] >= -1*B1 & s44$dz[i] < -1*A1 | s44$dz[i] <= B1 &  s44$dz[i] >
A1){

    Ordem[i] <- "2"
   }

   else { Ordem [i] <- "Sem Classificação"}

  }

  s44 <- cbind(dados,Ordem)

  #Plotando via ggplot2
  windows(8,6,title="Classificação: NORMAM-25 / S44")
  par(mfrow=c(1,1), family="serif")
  ggplot(s44, aes(x = X, y = Y, colour = Ordem)) + geom_point(size=2)+
   labs(title= "Classificação: NORMAM-25 / S-44", x= "E (m)", y= "N (m)") +
   theme_bw()+theme(plot.title = element_text(hjust = 0.5, size = 16))+
   theme(legend.title = element_text(size=14, face="bold"))+
   scale_color_manual(values = c( "2"="red","Sem Classificação" = "blue",
                              "Especial" = "forestgreen", "1A/1B" = "black"))

  #Plotando usando plot
  windows(8,6,title="Classificação: NORMAM-25 / S44")
  par(mfrow=c(1,1), cex=1.2, family="serif")
  plot(s44$X, s44$Y, col= s44$Ordem,
      xlab=" E (m)", ylab="N (m)", main="Classificação: NORMAM-25 / S-
44",pch=16)
  legenda <- aggregate(s44, by = list(unique.values = s44$Ordem), FUN=length)
  legend("bottomleft", inset=.05, legend= legenda$unique.values,
        col= legenda$unique.values, pch=16,bty="o", title="Ordem")

#########################################################################
#########################Proposição do Método############################
#########################################################################

#########################################################################
#Análise exploratória dos dados de discrepâncias
#########################################################################

  (res=summary(dados1))

  attach(res)

  #Principais medidas
  (med = round(mean(dados$dz),3))
  (min = round(min(dados$dz),3))
```

```
 (max = round(max(dados$dz),3))
 (des = round(sd(dados$dz),4))
 (var = round(var(dados$dz),4))
 (curt = round(kurtosis(dados$dz),2))
 (assim = round(skewness(dados$dz),2))
 (descritiva = data.frame(med,min,max,var,des, curt,assim))
 (n=length(dados$dz))
 (dist.min= round((distances.summary[1]),3))
 (dist.max= round((distances.summary[2]),3))

 #Exportando informações:
 sink("Resultados.txt", type="output", append=T)
 cat("  #### Proposição do Método ####","\n",
    "-----------------------------------------------------","\n",
    " Análise Exploratória dos Dados:","\n",
    "-----------------------------------------------------","\n",
    n, "observações"    ,"\n",
    "Média:"      ,med,'metros'  ,"\n",
    "Mínimo:"  ,min,"metros"  ,"\n",
    "Máximo:"     ,max,"metros"  ,"\n",
    "Variância:"  ,var,"metros²"  ,"\n",
    "Desvio Padrão:"     ,des,"metros"  ,"\n",
    "Coef. de Curtose: "         ,curt    ,"\n",
    "Coef. de Assimetria:"      ,assim  ,"\n",
    "dist.min.:"  ,dist.min,"metros"  ,"\n",
    "dist.max.:" ,dist.max,"metros" ,"\n",
    "-----------------------------------------------------","\n",
    fill=F)
 sink()
 shell.exec("Resultados.txt")

#########################################################################
#Gráficos para análise exploratória
#########################################################################

 windows(8,6,title="Gráficos para análise exploratória")
 par(mfrow=c(2,2), family="serif")

 hist(dados$dz, xlab="Discrepâncias (m)", ylab= "Frequência", main=" Histograma")
 rug(jitter(dados$dz))
 plot(density(dados$dz), xlab="Discrepâncias (m)", ylab= "Frequência", main="
Densidade")
 boxplot(dados$dz, xlab="dZ", ylab= "Discrepâncias (m)", main="Boxplot (Tukey)")
 qqnorm(dados$dz, xlab="Quantis Teóricos", ylab= "Quantis Amostrados", main="
Normal Q-Q Plot")
 qqline(dados$dz,lty=2, col='red')
 par(mfrow=c(1,1), family="serif")


 #Boxplot Ajustado
```

```
  windows(8,6,title="Gráficos para análise exploratória")
  par(mfrow=c(1,2), family="serif")
  adjbox(dados$dz, xlab="dZ", ylab= "Discrepâncias (m)", main="Boxplot Ajustado")
  boxplot(dados$dz, xlab="dZ", ylab= "Discrepâncias (m)", main="Boxplot (Tukey)")
  par(mfrow=c(1,1) ,family="serif")

#####################################################################
#Análise Exploratória Espacial
#####################################################################

  windows(8,6,canvas="snow2",title="Discrepâncias")
  ggplot(dados, aes(x = X, y = Y, colour = dz)) + geom_point()+
    xlab("E (m)") + ylab("N (m)") + ggtitle("Área de Estudo") +
    theme_bw()+theme(plot.title = element_text(hjust = 0.5))

  windows(8,6,title="Discrepâncias")
  scatterplot3d(dados$X,dados$Y,dados$dz,xlab="    E    (m)",    ylab="N    (m)",
zlab="Discrepância (m)", main="Área de Estudo")

  windows(8,6,title="Gráficos para análise exploratória")
  par(mfrow=c(2,2), family="serif")
  points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="equal")
  points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="data.proportional")
  points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="quartiles")
  points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="deciles")
  par(mfrow=c(1,1), family="serif")

  windows(8,6,title="Gráficos para análise exploratória")
  par(mfrow=c(1,1), family="serif")
  points(dados1,xlab="E (m)",ylab="N (m)",pt.divide="quartiles", main="Gráfico de
Quartis")

#####################################################################
#Análise de tendência
#####################################################################

  windows(8,6,title="Gráficos para análise de Tendência")
  par(mfrow=c(1,1), family="serif")
  plot(dados1,low=T) #Com linha de tendência

#####################################################################
#Detecção de outliers para dados (aproximadamente) simétricos/normais
#####################################################################

  #Boxplot de Tukey
  out.box <- boxplot.stats(dados$dz)

  #Isola os outliers detectados pelo boxplot de Tukey
  result <- dados[-which(dados$dz<out.box$stats[1] | dados$dz>out.box$stats[5]),]
```

```
 #Z-Score Modificado
 ZSM <- abs((0.6745*(dados$dz-median(dados$dz)))/(mad(dados$dz,constant = 1)))

 #Isola os outliers detectados pelo z score modificado
 result1 <- dados[-which(ZSM>3),]

#########################################################################
#Detecção de outliers para dados assimétricos
#########################################################################

 #adjbox(dados$dz, xlab="dZ", ylab= "Discrepâncias (m)", main="Boxplot
Ajustado")
 out.box1 <- adjboxStats(dados$dz)

 #isola os outliers detectados pelo boxplot ajustado
 result.box.ajust <- dados[-which(dados$dz<out.box1$stats[1] |
dados$dz>out.box1$stats[5]),]

 #Exportando informações:
 sink("Resultados.txt", type="output", append=T)

 cat(" Detecção de Outliers","\n",
    "-----------------------------------------------------","\n",
    "Boxplot (Tukey): "        ,out.box$out,"\n\n",
    "Boxplot Ajustado: "        ,out.box1$out,"\n\n",
    "Z-Score Modificado: "        ,dados$dz[which(ZSM>3)]   ,"\n\n",
    "-----------------------------------------------------","\n",
    fill=F)
 sink()
 shell.exec("Resultados.txt")

 #Gera arquivo com dados sem outliers

 write.table(result, "dados_semout_boxplot.txt", dec=",")

 write.table(result1, "dados_semout_ZSM.txt", dec=",")

 write.table(result.box.ajust, "dados_semout_boxplot_ajustado.txt", dec=",")

#########################################################################
#Escolher arquivo sem outliers
#########################################################################

 #Leitura dos dados
 dados <- read.table("dados_semout_boxplot_ajustado.txt", header=T, dec=",")
 names(dados)
 dados
 length(dados$dz)

 #Leitura dos dados para análise Geoestatística
```

```
 dados1 <- read.geodata("dados_semout_boxplot_ajustado.txt", header=T,
dec=",",coords=1:2, data.col=4)
 names(dados1)
 dados1
 length(dados1$data)
 dup.coords(dados1)

#########################################################################
#Análise exploratória dos dados de discrepâncias
#########################################################################

 (res=summary(dados1))

 attach(res)

 #Principais medidas
 (med = round(mean(dados$dz),3))
 (min = round(min(dados$dz),3))
 (max = round(max(dados$dz),3))
 (des = round(sd(dados$dz),4))
 (var = round(var(dados$dz),4))

 (curt = round(kurtosis(dados$dz),2))
 (assim = round(skewness(dados$dz),2))
 (descritiva = data.frame(med,min,max,var,des, curt,assim))
 (n=length(dados$dz))
 (dist.min= round((distances.summary[1]),3))
 (dist.max= round((distances.summary[2]),3))

 #Exportando informações:
 sink("Resultados.txt", type="output", append=T)
 cat("  #### Dados sem Outliers ####","\n",
    "-----------------------------------------------------","\n",
    " Análise Exploratória dos Dados:","\n",
    "-----------------------------------------------------","\n",
    n, "observações"    ,"\n",
    "Média:"      ,med,'metros'  ,"\n",
    "Mínimo:"  ,min,"metros"  ,"\n",
    "Máximo:"     ,max,"metros"  ,"\n",
    "Variância:"  ,var,"metros²"  ,"\n",
    "Desvio Padrão:"     ,des,"metros"  ,"\n",
    "Coef. de Curtose: "         ,curt    ,"\n",
    "Coef. de Assimetria:"      ,assim  ,"\n",
    "dist.min:"  ,dist.min,"metros"  ,"\n",
    "dist.max:" ,dist.max,"metros" ,"\n",
    "-----------------------------------------------------","\n",
    fill=F)
 sink()
 shell.exec("Resultados.txt")
```

```
#########################################################################
#Gráficos para análise exploratória
#########################################################################

 windows(8,4,title="Gráficos para análise exploratória")
 par(mfrow=c(1,3), family="serif")
 hist(dados$dz, xlab="Discrepâncias (m)", ylab= "Frequência", main=" Histograma")
 rug(jitter(dados$dz))
 plot(density(dados$dz), xlab="Discrepâncias (m)", ylab= "Frequência", main="
Densidade")
 qqnorm(dados$dz, xlab="Quantis Teóricos", ylab= "Quantis Amostrados", main="
Normal Q-Q Plot")
 qqline(dados$dz,lty=2, col='red')
 par(mfrow=c(1,1), family="serif")


 windows(8,6,title="Gráficos para análise exploratória")
 par(mfrow=c(1,2), family="serif")
 adjbox(dados$dz, xlab="dZ", ylab= "Discrepâncias (m)", main="Boxplot Ajustado")
 boxplot(dados$dz, xlab="dZ", ylab= "Discrepâncias (m)", main="Boxplot (Tukey)")
 par(mfrow=c(1,1), family="serif")

#########################################################################
#Análise Exploratória Espacial
#########################################################################

 windows(8,6,canvas="snow2",title="Discrepâncias")
 ggplot(dados, aes(x = X, y = Y, colour = dz)) + geom_point()+
  xlab("E (m)") + ylab("N (m)") + ggtitle("Área de Estudo") +
  theme_bw()+theme(plot.title = element_text(hjust = 0.5))

 windows(8,6,title="Discrepâncias")
 scatterplot3d(dados$X,dados$Y,dados$dz,xlab="  E  (m)",  ylab="N  (m)",
zlab="Discrepância (m)", main="Área de Estudo")


 windows(8,6,title="Gráficos para análise exploratória")
 par(mfrow=c(2,2), family="serif")
  #Visualização
 points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="equal")
 points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="data.proportional")
 points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="quartiles")
 points(dados1,xlab="E (m)",ylab="N (m)", pt.divide="deciles")
 par(mfrow=c(1,1), family="serif")

 windows(8,6,title="Gráficos para análise exploratória")
 par(mfrow=c(1,1), family="serif")
 points(dados1,xlab="E (m)",ylab="N (m)",
       pt.divide="quartiles", main="Gráfico de Quartis")
```

```
#########################################################################
#Análise de tendência
#########################################################################

 windows(8,6,title="Gráficos para análise exploratória")
 par(mfrow=c(1,1), family="serif")
 plot(dados1,low=T) #Com linha de tendência

#########################################################################
#Análise de Independência (semivariograna)
#########################################################################

 #Semivariograma empírico

 #Construção de 3 semivariogramas:
 #1º com alcance igual a 75% da distância máxima
 #2º com alcance igual a 50% da distância máxima
 #3º com alcance igual a 25% da distância máxima


 windows(8,6,title="Semivariograma Omnidirecional")
 escala.y=2*var

 par(mfrow=c(2,2), family="serif")
 vario.emp.1 <- variog(dados1,max.dist=(0.75*dist.max),
direction="omnidirectional")
 plot(vario.emp.1,ylim=c(0,escala.y),xlab="Distâncias (m)",ylab="Semivariâncias
(m²)", main=("75% da Distância Máxima"))
 abline(var(dados1$data),0, col="gray60", lty=2, lwd=2)
 legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')

 vario.emp.1 <- variog(dados1,max.dist=(0.50*dist.max),
direction="omnidirectional")
 plot(vario.emp.1,ylim=c(0,escala.y),xlab="Distâncias (m)",ylab="Semivariâncias
(m²)", main=("50% da Distância Máxima"))
 abline(var(dados1$data),0, col="gray60", lty=2, lwd=2)
 legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')

 vario.emp.1 <- variog(dados1,max.dist=(0.25*dist.max),
direction="omnidirectional")
 plot(vario.emp.1,ylim=c(0,escala.y),xlab="Distâncias (m)",ylab="Semivariâncias
(m²)", main=("25% da Distância Máxima"))
 abline(var(dados1$data),0, col="gray60", lty=2, lwd=2)
 legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')
 par(mfrow=c(1,1), family="serif")

 dist.max
 0.75*dist.max
 0.50*dist.max
```

```
  0.25*dist.max

#########################################################################
#Semivariograma omnidirecional das discrepâncias/Envelope de Monte Carlo
#########################################################################

  M <- (0.5*dist.max) #Semivariograma das discrepâncias para distância de M m.

  windows(8,6,title="Semivariograma Omnidirecional")
  par(mfrow=c(1,1), family="serif")
  vario.emp.1 <- variog(dados1,max.dist= M, direction="omnidirectional")
  plot(vario.emp.1,   ylim=c(0,escala.y),xlab="Distâncias   (m)",ylab="Semivariâncias
(m²)", main=("Semivariograma das Discrepâncias"))
  abline(var(dados1$data),0, col="gray60", lty=2, lwd=2)
  legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')

  #Envelope de variograma (Simulação Monte Carlo)
  vario.env <- variog.mc.env(dados1, obj.v=vario.emp.1)
  plot(vario.emp.1,                    env=vario.env,ylim=c(0,escala.y),xlab="Distâncias
(m)",ylab="Semivariâncias (m²)", main=("Semivariograma das Discrepâncias"))
  abline(var(dados1$data),0, col="gray60", lty=2, lwd=2)
  legend("topleft","Variância Amostral", col="gray60",lty=2, lwd=2,bty='n')

  #Exportando informações:
  sink("Resultados.txt", type="output", append=T)
  cat(" Resultados do cálculo do Semivariograma:","\n",
     "-----------------------------------------------------","\n",
     vario.emp.1$n.data, "observações"    ,"\n",
     "Distâncias:"      , vario.emp.1$u  ,"\n",
     "Semivariâncias:"  ,vario.emp.1$v  ,"\n",
     "Número de pares em cada lote:"     ,vario.emp.1$n  ,"\n",
     "Desvio padrão de cada lote:"  ,vario.emp.1$sd,"\n",
     "Distância Máxima:"     ,vario.emp.1$max.dist,"\n",
     "Direção:"         ,vario.emp.1$direction  ,"\n",
     "-----------------------------------------------------","\n",
     fill=F)
  sink()
  shell.exec("Resultados.txt")

#########################################################################
#Carregar Funções
#########################################################################

  #Função para cálcular Incerteza
  theta <- function(x){(sqrt((sd(x)^2)+(mean(x)^2)))}

  #Função para cálcular RMSE (dividido por n)
  theta1 <- function(x){(sqrt((sum(x^2))/length(x)))}

  #Função para cálcular Incerteza TCL
```

```
 theta2 <- function(x,y){(sqrt(((sd(x)^2)*y)+(mean(x)^2)))}

 #Função para calcular Incerteza Robusta
 theta3 <- function(x){(sqrt((mad(x)^2)+(median(x)^2)))}

#########################################################################
######################### Amostra Independente ##########################
#########################################################################

#########################################################################
#Análise da Normalidade
#########################################################################

 par(mfrow=c(1,1), family="serif")
 qqnorm(dados$dz, xlab="Quantis Teóricos", ylab= "Quantis Amostrados", main="
Normal Q-Q Plot")
 qqline(dados$dz,lty=2, col='red')


 shapiro.test(dados$dz)
 shap <- shapiro.test(dados$dz)
 ks.test(dados$dz,"pnorm", mean(dados$dz), sd(dados$dz))
 ks <- ks.test(dados$dz,"pnorm", mean(dados$dz), sd(dados$dz))

 #Exportando informações:
 sink("Resultados.txt", type="output", append=T)

 cat(" Análise da Normalidade da Amostra","\n",
    "Tamanho da Amostra: ",n,"\n",
    "-----------------------------------------------------","\n",
    "Teste Shapiro-Wilk "        ,"\n",
    "p-value: "         ,shap$p.value   ,"\n",
    "Normal: "          ,(shap$p.value>0.05),"\n",
    "\n Teste Kolmogorov-Smirnov "        ,"\n",
    "p-value: "         ,ks$p.value   ,"\n",
    "Normal: "          ,(ks$p.value>0.05),"\n",
    "\np-valeu > 0.05, amostra é normal ao nível de significância de 5%","\n",
    "-----------------------------------------------------","\n",
    fill=F)
 sink()
 shell.exec("Resultados.txt")


#########################################################################
#Amostra normal e sem Outliers
#########################################################################

 #Verificação
 tol <- 0.01 #10 centímetros
 (1.96*sd(dados$dz)-quantile(dados$dz,0.95)) < tol
```

```
  (mean(dados$dz)-median(dados$dz)) < tol
  (sd(dados$dz)- quantile(dados$dz,0.683)) < tol

  #cálculo das estatísticas
  ivt = theta(dados$dz)
  rms = theta1(dados$dz)

  #Número de amostras para estimar o IC por bootstrap
  amostra=5000

  #Boott - Bootstrap-t Confidence Limits
  results.boot <- boott(dados$dz,theta,
nboott=amostra,VS=FALSE,perc=c(0.025,0.975))
  results.boot1 <- boott(dados$dz,theta1,
nboott=amostra,VS=FALSE,perc=c(0.025,0.975))

  #Nonparametric BCa Confidence Limits
  results.bca <- bcanon(dados$dz, amostra, theta,alpha=c(0.025, 0.975))
  results.bca1 <- bcanon(dados$dz, amostra, theta1,alpha=c(0.025, 0.975))

  #Intervalo de confiança baseado no qui-quadrado para RMSE.
  results.rms <- ci.rmsea(rms,length(dados$dz),length(dados$dz),conf.level = 0.95,
alpha.lower = NULL, alpha.upper = NULL)

  #Exportando informações:
  sink("Resultados.txt", type="output", append=T)

  cat(" Incerteza Vertical \n Amostra Independente, Normal e sem Outliers","\n",
     "\n Intervalo de Confiança de 95%","\n",
     "-----------------------------------------------------","\n",
     "Incerteza (m): "      ,round(ivt,3)   ,"\n",
     "IC bootstrap-t (m): "      ,"[",round(results.boot$confpoints[1,1],3),
     ";",round(results.boot$confpoints[1,2],3),"]","\n",
     "IC BCa (95%): "      ,"[",round(results.bca$confpoints[1,2],3),
     ";",round(results.bca$confpoints[2,2],3),"]","\n",

     "\n RMSE (m): "      ,round(rms,3)   ,"\n",
     "IC (qui-quadrado)(m): "
,"[",round(results.rms$Lower.Conf.Limit,3),";",round(results.rms$Upper.Conf.Limit,
3),"]","\n",
     "IC bootstrap-t (m): "      ,"[",round(results.boot1$confpoints[1,1],3),
     ";",round(results.boot1$confpoints[1,2],3),"]","\n",
     "IC BCa (m): "      ,"[",round(results.bca1$confpoints[1,2],3),
     ";",round(results.bca1$confpoints[2,2],3),"]","\n",
     "-----------------------------------------------------","\n",
     fill=F)
  sink()
  shell.exec("Resultados.txt")

#####################################################################
```

```
#Amostra não normal: Aplicação do TCL ou abordagem robusta
#########################################################################

  #Aplicação do TCL - Teorema Central do Limite

  #Particionamento em torno dos medoids - funcao PAM {cluster}

  #Número médio de pontos por cluster
  Tamanho_amostral = 4

  #Número de clusters
  k= round((length(dados$dz))/Tamanho_amostral, 0)

  grupos <- pam(dados[,c(1,2)],k=k,metric = "euclidean", stand = TRUE)

  #Agrupa as amostras por indíce e calcula a média de cada uma
  TCL <- aggregate(dados$dz~grupos$clustering, FUN = mean)

  #PLota os agrupamentos
  windows(10,5)
  par(mfrow=c(1,2), family="serif")
  plot(dados$X,dados$Y,xlab=" E (m)", ylab="N (m)", col=grupos$clustering,
      main="Agrupamento k-medóides")
  points(grupos$medoids, pch=16, col=25)
  legend("bottom", inset=.05, legend= "Centróide",
         col= 25, pch=16,bty="o")
  plot(dados$dz~grupos$clustering, xlim=c(0,k+1),
      ylim=c(min(dados$dz-0.1),max(dados$dz+0.1)), xlab="Grupos",
      ylab="Discrepâncias (m)", xaxt="n", main="Distribuição dos Agrupamentos")
  axis(1,at=seq(1,k, by=1)) #adiciona o eixo X
  points(TCL, col=2,pch=16)
  legend("bottom",inset=.05, legend= "Média do Cluster",
         col= 2, pch=16,bty="o")
  par(mfrow=c(1,1), family="serif")

  #Análise da normalidade da nova amostra
  shap1 <- shapiro.test(TCL$`dados$dz`)
  ks1 <- ks.test(TCL$`dados$dz`,"pnorm", mean(TCL$`dados$dz`),
sd(TCL$`dados$dz`))

  windows(8,6,title="Análise Exploratória")
  par(mfrow=c(2,2), family="serif")
  hist(TCL$`dados$dz`, xlab="Média das Discrepâncias (m)", ylab= "Frequência",
main=" Histograma")
  rug(jitter(TCL$`dados$dz`))
  plot(density(TCL$`dados$dz`), xlab="Média das Discrepâncias (m)", ylab=
"Frequência", main=" Densidade")
  boxplot(TCL$`dados$dz`, xlab="dZ", ylab= "Média das Discrepâncias (m)",
main="Boxplot")
```

```
 qqnorm(TCL$`dados$dz`, xlab="Quantis Teóricos", ylab= "Quantis Amostrados",
main=" Normal Q-Q Plot")
 qqline(TCL$`dados$dz`,lty=2, col='red')
 par(mfrow=c(1,1), family="serif")

 #Exportando informações:
 sink("Resultados.txt", type="output", append=T)

 cat(" Teorema Centra do Limite (TCL)","\n",
    "-----------------------------------------------------","\n",
    "Tamanho da Amostra original: ",length(dados$dz)   ,"\n",
    "Número de Agrupamentos: "        ,k   ,"\n",
    "Tamanho Amostral Definido: "        ,Tamanho_amostral   ,"\n",
    "Tamanho Amostral Médio dos Agrupamentos: "
,round(mean(grupos$clusinfo[,1]),3)  ,"\n\n",
    "Média TCL (m): "        ,round(mean(TCL$`dados$dz`),3)   ,"\n",
    "variância TCL: "        ,round(var(TCL$`dados$dz`),3)   ,"\n\n",
    "Teste Shapiro-Wilk (m²) "        ,"\n",
    "p-value: "        ,shap1$p.value   ,"\n",
    "Normal: "        ,(shap1$p.value>0.05),"\n",
    "\n Teste Kolmogorov-Smirnov "        ,"\n",
    "p-value: "        ,ks1$p.value   ,"\n",
    "Normal: "        ,(ks1$p.value>0.05),"\n",
    "\np-valeu > 0.05, amostra é normal ao nível de significância de 5%","\n",
    "-----------------------------------------------------","\n",
    fill=F)
 sink()
 shell.exec("Resultados.txt")


#####################################################################
#Amostra obtida pelo TCL: normal e sem Outliers
#####################################################################

 #Verificação
 tol <- 0.1 #10 centímetros
 (1.96*sd(TCL$`dados$dz`)-quantile(TCL$`dados$dz`,0.95)) < tol
 (mean(TCL$`dados$dz`)-median(TCL$`dados$dz`)) < tol
 (sd(TCL$`dados$dz`)- quantile(TCL$`dados$dz`,0.683)) < tol

 #Cálculo das estatísticas
 ivt1 = theta2(TCL$`dados$dz`,mean(grupos$clusinfo[,1]))

 #Número de amostras para estimar o IC por bootstrap
 amostra=5000

 #Boott - Bootstrap-t Confidence Limits
 results.boot2 <- boott(TCL$`dados$dz`,theta2, mean(grupos$clusinfo[,1]),
nboott=amostra,VS=FALSE,perc=c(0.025,0.975))
```

```
 #Nonparametric BCa Confidence Limits
 results.bca2 <- bcanon(TCL$`dados$dz`, amostra, theta2
,mean(grupos$clusinfo[,1]), alpha=c(0.025, 0.975))

 #Exportando informações:
 sink("Resultados.txt", type="output", append=T)

 cat(" Incerteza Vertical \n Amostra Independente, Normal e sem Outliers\n
TCL","\n",
    "\n Intervalo de Confiança de 95%","\n",
    "-----------------------------------------------------","\n",
    "Incerteza (m): "       ,round(ivt1,3)   ,"\n",
    "IC bootstrap-t (m): "       ,"[",round(results.boot2$confpoints[1,1],3),
    ";",round(results.boot2$confpoints[1,2],3),"]","\n",
    "IC BCa (95%): "       ,"[",round(results.bca2$confpoints[1,2],3),
    ";",round(results.bca2$confpoints[2,2],3),"]","\n",
    "-----------------------------------------------------","\n",
    fill=F)
 sink()
 shell.exec("Resultados.txt")

#########################################################################
#Abordagem Robusta
#########################################################################

 #Cálculo das estatísticas
 ivt2 = theta3(dados$dz)

 #Número de amostras para estimar o IC por bootstrap
 amostra1=5000

 #Boott - Bootstrap-t Confidence Limits
 results.boot3 <- bootт(dados$dz,theta3,
nboott=amostra1,VS=FALSE,perc=c(0.025,0.975))

 #Nonparametric BCa Confidence Limits
 results.bca3 <- bcanon(dados$dz, amostra1, theta3,alpha=c(0.025, 0.975))

 #Exportando informações:
 sink("Resultados.txt", type="output", append=T)

 cat(" Incerteza Vertical \n Amostra Independente e Não Normal\n Abordagem
Robusta","\n",
    "\n Intervalo de Confiança de 95%","\n",
    "-----------------------------------------------------","\n",
    "Incerteza Robusta (m): "       ,round(ivt2,3)   ,"\n",
    "IC bootstrap-t (m): "       ,"[",round(results.boot3$confpoints[1,1],3),
    ";",round(results.boot3$confpoints[1,2],3),"]","\n",
    "IC BCa (m): "       ,"[",round(results.bca3$confpoints[1,2],3),
    ";",round(results.bca3$confpoints[2,2],3),"]","\n",
```

```
    "\n Mediana (m): "         ,round(median(dados$dz),3)   ,"\n",
    "NMAD (m): "         ,round(mad(dados$dz),3)   ,"\n",
    "Q (0,683) (m): "         ,round(quantile(dados$dz,0.683),3),"\n",
    "Q (0,95) (m): "         ,round(quantile(dados$dz,0.95),3)   ,"\n",
    "-----------------------------------------------------","\n",
    fill=F)
  sink()
  shell.exec("Resultados.txt")

#########################################################################
######################### Amostra dependente ############################
#########################################################################

#########################################################################
#Block Bootstrap: Gerar IC com 95%
#########################################################################

  #Obtendo os dados
  #Carregar dados sem outliers
  dados <- read.table("dados_semout_boxplot_ajustado.txt", header=T, dec=",")
  coordinates(dados) <- c("X", "Y")

  #Cálculo das estatísticas
  ivt3 = theta(dados$dz)
  rms1 = theta1(dados$dz)
  ivt_robs = theta3(dados$dz)

  #Gerar blocos com diagonal de tamanho pre-defindo
  #sugestão: Alcance obtido da análise Geoestatística
  tamanho <- 300


  #Delimitar o numero e localização de cada bloco
  Bloco <- makegrid(bbox(dados), cellsize = (tamanho*sqrt(2)), pretty = FALSE)
  coordinates(Bloco) <- c("x1","x2")
  gridded(Bloco) <- TRUE
  Bloco <- as.SpatialPolygons.GridTopology(Bloco@grid)
  plot(Bloco)     #Plotando os Blocos
  points(dados) #Plotando os dados originais

  #Extrair o numero do Bloco onde cada ponto esta sobrepondo
  ptsInBloco <- as.numeric(gIntersects(dados, Bloco, byid=TRUE,
                                returnDense=FALSE,  checkValidity=TRUE))   #Todos que
tem intersecao para cada buffer

  #Número de replicações Bootstrap
  n_vezes <- 500

  tab_boot <- tab_boot_dz <- NULL
```

```
  for (i in 1:n_vezes)
  {
   #PRIMEIRO - sorteio do Bloco
   Grid <- sample(unique(ptsInBloco),dim(dados)[1], replace = TRUE)
   #SEGUNDO - sorteio de um ponto dentro de cada Bloco selecionado anteriormente
   pontos <- (as.numeric(lapply(Grid,function(x) sample(which(ptsInBloco==x),1))))
#Os pontos repetidos sao contabilizados apenas uma vez
   tab_boot <- rbind(tab_boot,(pontos))
   tab_boot_dz <- rbind(tab_boot_dz,dados@data$dz[pontos])
  }

  #Converter os dados para data.frame
  tab_boot <- as.data.frame(tab_boot)
  tab_boot_dz <- as.data.frame(tab_boot_dz)


  #Gerar um novo conjunto de dados apos o block boostrap
  dados_novos_ivt <- apply(tab_boot_dz,1,theta)
  dados_novos_rms <- apply(tab_boot_dz,1,theta1)
  dados_novos_robs <- apply(tab_boot_dz,1,theta3)

  #Intervalos de confiança block bootstrap
  IC_ivt <- quantile(dados_novos_ivt,c(0.025, 0.975))
  IC_rms <- quantile(dados_novos_rms,c(0.025, 0.975))
  IC_robs <- quantile(dados_novos_robs,c(0.025, 0.975))

  #viés
  vies_ivt <- ivt3 - median(dados_novos_ivt)
  vies_rms <- rms1 - median(dados_novos_rms)
  vies_robs <- ivt_robs - median(dados_novos_robs)

  windows(8,8,title="Gráficos para análise exploratória")
  par(mfrow=c(3,2), family="serif")
  hist(dados_novos_ivt, xlab="Incerteza (m)", ylab= "Frequência", main="
Histograma (bootstrap)")
  qqnorm(dados_novos_ivt, xlab="Quantis Teóricos", ylab= "Quantis Amostrados",
main=" Normal Q-Q Plot (bootstrap)")
  qqline(dados_novos_ivt,lty=2, col='red')
  hist(dados_novos_rms, xlab="RMSE (m)", ylab= "Frequência", main=" Histograma
(bootstrap)")
  qqnorm(dados_novos_rms, xlab="Quantis Teóricos", ylab= "Quantis Amostrados",
main=" Normal Q-Q Plot (bootstrap)")
  qqline(dados_novos_rms,lty=2, col='red')
  hist(dados_novos_robs, xlab="Incerteza Robusta (m)", ylab= "Frequência", main="
Histograma (bootstrap)")
  qqnorm(dados_novos_robs, xlab="Quantis Teóricos", ylab= "Quantis Amostrados",
main=" Normal Q-Q Plot (bootstrap)")
  qqline(dados_novos_robs,lty=2, col='red')
  par(mfrow=c(1,1), family="serif")
```

```
#Exportando informações:
sink("Resultados.txt", type="output", append=T)

cat(" Incerteza Vertical \n Amostra Dependente - Bloco Bootstrap","\n",
   "Número de replicações: " ,n_vezes,"\n",
   "Tamanho do lado do Bloco (m): " ,round (tamanho*sqrt(2),3),"\n",
   "\n Intervalo de Confiança de 95%","\n",
   "-----------------------------------------------------","\n",
   "Incerteza (m): "         ,round(ivt3,3)   ,"\n",
   "IC (m): "         ,"[",round(IC_ivt[1],3),";",round(IC_ivt[2],3),"]","\n",
   "Viés Bootstrap (m): "         ,round(vies_ivt,3)   ,"\n",

   "\n RMSE (m): "         ,round(rms1,3)   ,"\n",
   "IC (m): "         ,"[",round(IC_rms[1],3),";",round(IC_rms[2],3),"]","\n",
   "Viés Bootstrap (m): "         ,round(vies_rms,3)   ,"\n",

   "\n Incerteza Robusta (m): "         ,round(ivt_robs,3)   ,"\n",
   "IC (m): "         ,"[",round(IC_robs[1],3),";",round(IC_robs[2],3),"]","\n",
   "Viés Bootstrap (m): "         ,round(vies_robs,3)   ,"\n",
   "-----------------------------------------------------","\n",
   fill=F)
sink()
shell.exec("Resultados.txt")
```

Capítulo 3

a) Algoritmo do Método PP

```
#####################################################################
#Controle de qualidade em levantamentos hidrográficos
#Prof. ITALO O FERREIRA / UFV / DEC / EAM
#Prof. JÚLIO CÉSAR DE OLIVEIRA / UFV / DEC / EAM
#Artigo 03 - Tese
#Método PP (Point to Point)
#####################################################################
#Obs: Para a aplicação correta deste algoritmo, favor consultar o texto da tese.

 #////////////////////////////////////////////////////////
 # Script para analisar duas faixas de sondagem e encontrar os pontos
 # mais proximos de cada varredura (LRS e LV), gerando as discrepâncias.
 #////////////////////////////////////////////////////////

 #Caminho dos dados

setwd("C:/Users/Ítalo/Documents/2_Pesquisa_Doutorado/1_TESE/2_TESE_AFONSO/Capítulo3/Base de Dados/Dados de Estudo")

 getwd()

 #Lista de pacotes a serem usados no Script
 pkg <- c("rgeos", "raster", "rgdal", "nabor")

 sapply(pkg, require, character.only=TRUE)

 #Função pts_near
 pts_near <- function (LRS,LV,lim_dist=0, b=0)
 {
  #LRS - Faixa Regular de Sondagem (dados do tipo SpatialPointsDataframe)
  #LV - faixa de Verificação (dados do tipo SpatialPointsDataframe)
  #lim_dist - distancia maxima (na mesma unidade do LRS)
  #          em que o ponto mais proximo pode ser considerado valido
  #          o valor inicial sera a diagonal do retangulo envolvente da
  #          sobreposicao entre as duas faixas
  # b - o valor de um limite adicional, ou seja, um buffer na intersecao das faixas
  #     o valor default de b==0, ou seja, nao sera aplicado buffer


  #Extraindo o limite da faixa LRS
  LRS_points_lim <- LRS@coords[chull(LRS@coords[,1:2]),1:2]
  LRS_pol      <-      SpatialPolygons(list(Polygons(list(Polygon(LRS_points_lim)),
ID=1)))
  #Extraindo o limite da faixa LV
  LV_points_lim <- LV@coords[chull(LV@coords[,1:2]),1:2]
```

```
LV_pol <- SpatialPolygons(list(Polygons(list(Polygon(LV_points_lim)), ID=1)))

#Intersecao entre entre as faixas LRS e LV
Int <- intersect(LRS_pol,LV_pol)


#Gerando um buffer na intersecao
Int_B<- gBuffer(Int, width = b)
crs(Int_B)<-crs(LRS)


#Extraindo os pontos LRS e LV que estao dentro da intersecao entre as faixas
LRS_int <- LRS[Int_B,]
LV_int <- LV[Int_B,]

#Se nao for informado o valor da distancia limite para encontrar o ponto mais
proximo
#sera atribuida a diagonal da faixa de interesecao como limite
if (lim_dist==0)
lim_dist <- sqrt(sum((diff(t(as.matrix(extent(LRS_int)))))^2))

#Encontrar os pontos mais proximos
#a partir da faixa de referencia - LV, encontrar o ponto
#mais proximo na faixa a ser analisada - LRS
near <- knn((LV_int@coords[,1:2]), (LRS_int@coords[,1:2]),k=1)

#Qual ponto esta com distancia acima do DISTANCIA LIMITE (lim_dist)
pts_lim <- which(near$nn.dists>lim_dist)

#Atualizar o arquivo NEAR com os valores de
#Z: cota dos pontos
#dp: discrepancia entre a referencia (LV) e o dado (LRS), ou seja dp = LV - LRS
#
 near <- list(near,
          array(LV_int@data$Z[near$nn.idx], dim = dim(near$nn.idx)),  #Cotas dos
pontos proximos obtidas no arquivo LV
          as.numeric(LRS_int@data$Z), #Cotas dos pontos do arquivo  LRS
          coordinates(LRS_int), #Coordenadas dos pontos
          dp              =              as.numeric(LV_int@data$Z[near$nn.idx[,1]])-
as.numeric(LRS_int@data$Z))

names(near) <- c("pontos", "Z_in_ref","Z_in_Dados","coordenadas","dp")

LRS_int@data <- cbind(LRS_int@data,Z2=near$Z_in_ref,dz=near$dp)

#Excluir os pontos que possuem distancias acima do LIMITE_DISTANCIA
LRS_int<- LRS_int[-pts_lim,]

return(LRS_int)
```

```
} #Fim da funcao pts_near


#######################################################################
#Leitura da base de dados (arquivo shp)
#######################################################################

 #Dados de entrada:
 #a) Linha Regular de Sondagem (LRS) no formato shp
 #b) Linhas de verificação (LV) no formato shp

 aa <- Sys.time()

 #Leitura dos dados no formato shp
 Faixa_dados<-readOGR(dsn=".",layer="LRS")
 Faixa_Ver<-readOGR(dsn=".",layer="LV")

 #Executar a funcao pts_near
 limite <- 0.01
 buffer <- 0

 dp <- pts_near(Faixa_dados, Faixa_Ver, limite, buffer)

 #Plotar base de dados
 windows(8,6,title="Base de Dados Analisada")
 par(mfrow=c(1,1), family="serif")
 plot(Faixa_dados, pch=1, col=1,
     xlab="E (m)", ylab= "N (m)", main="Discrepâncias")
 points(Faixa_Ver, pch=1, col=4)
 points(dp, pch=19, col=2)
 legend('bottomleft',legend=c('Faixa Regular de Sondagem','Faixa de Verificação',
                         "Discrepâncias"),col=c(1, 4, 2),pch=c(1, 1, 19))

 #Exportando informações:
 sink("Results.txt", type="output", append=T)
 cat("##### Tese de doutorado #####\n Prof. Ítalo O. Ferreira\n e-mail:
italo.ferreira@ufv.br\n\n Método para Obtenção das Discrepâncias em Sondagens por
Varrimento","\n",
     "-----------------------------------------------------","\n",
     "LRS:"      ,length(Faixa_dados),'Profundidades'  ,"\n",
     "LV:"  ,length(Faixa_Ver),'Profundidades'  ,"\n",
     "Distância Máxima:"     ,limite,"metros"  ,"\n",
     "Buffer:"  ,buffer,"metros","\n",
     "Número de Discrepâncias:",length(dp@data$dz),"\n",
     "-----------------------------------------------------","\n",
     fill=F)
 sink()
 shell.exec("Results.txt")

#verificar colunas para exclusão
```

```
names(dp)

#Gerar arquivo de dps no formato txt (X, Y, Z, dz)
write.table(dp@data[,-c(4,5)], "dp.txt", dec=",")

#Gerar arquivo de dps no formato shp (X, Y, Z, dz)
writeOGR(dp[,-c(4,5)], dsn=".", layer="dp", driver="ESRI Shapefile")

Sys.time() - aa
```

b) Semivariogramas omnidirecionais experimentais e os modelos ajustados

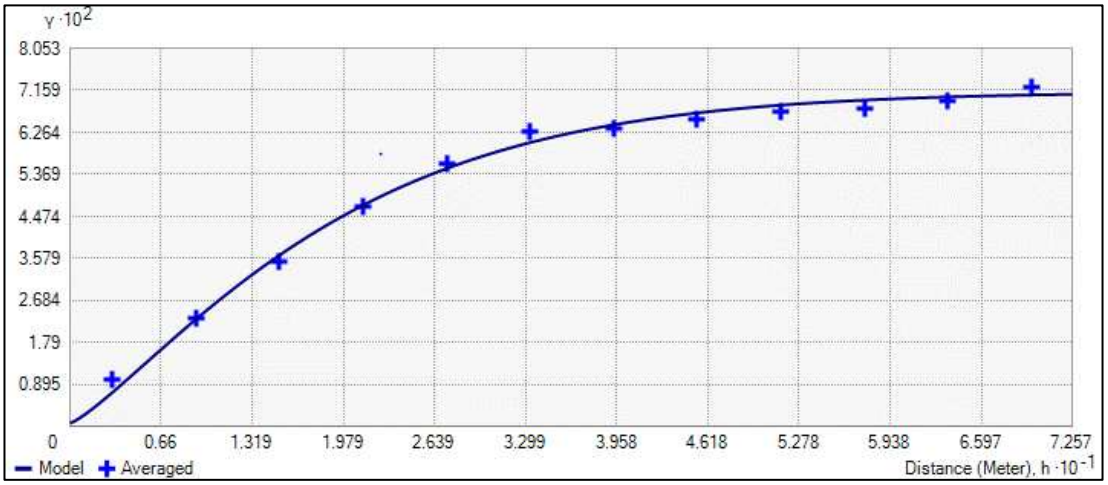

Figura 1 – Semivariograma omnidirecional experimental e modelo ajustado / Modelo Batimétrico referente às LRS.

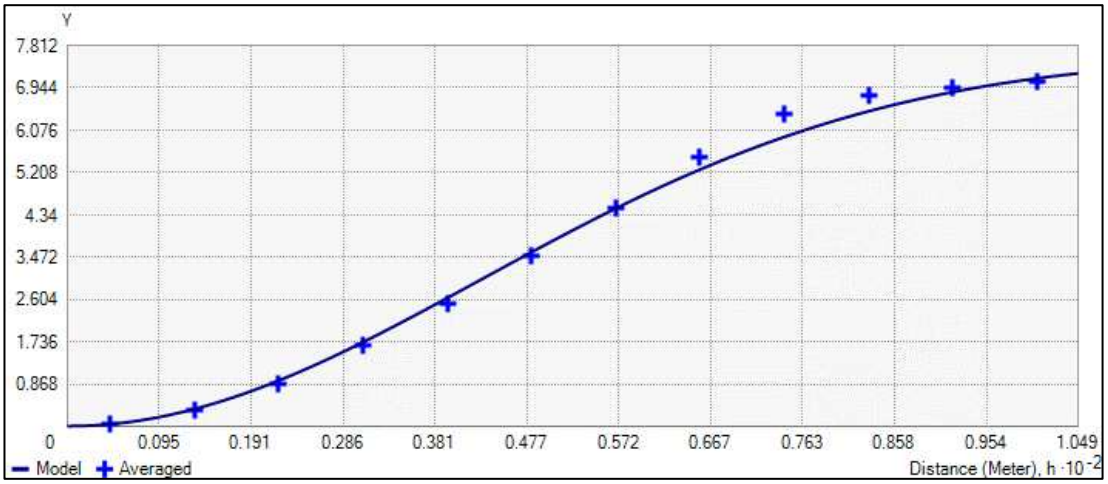

Figura 2 – Semivariograma omnidirecional experimental e modelo ajustado / Modelo Batimétrico referente à LV1.

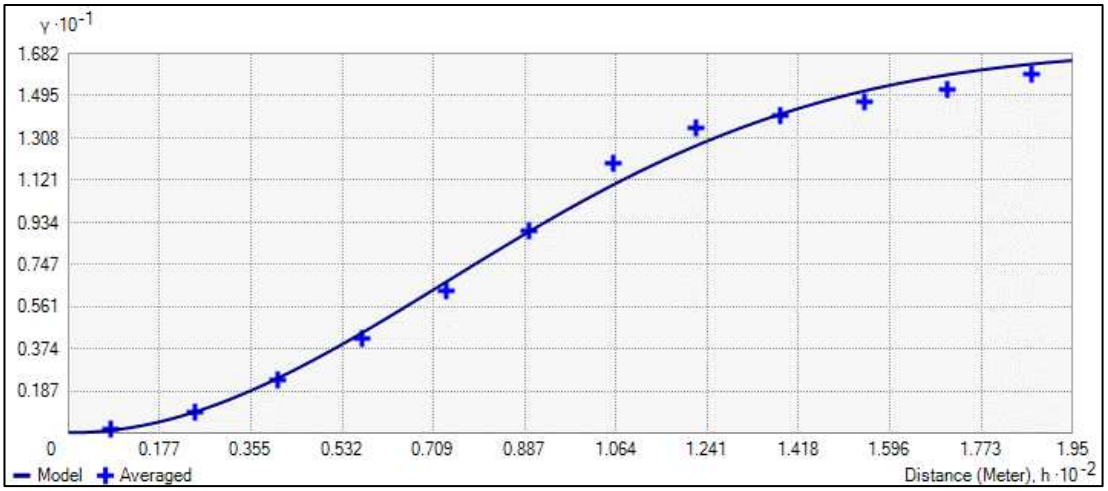

Figura 3 – Semivariograma omnidirecional experimental e modelo ajustado / Modelo Batimétrico referente à LV2.

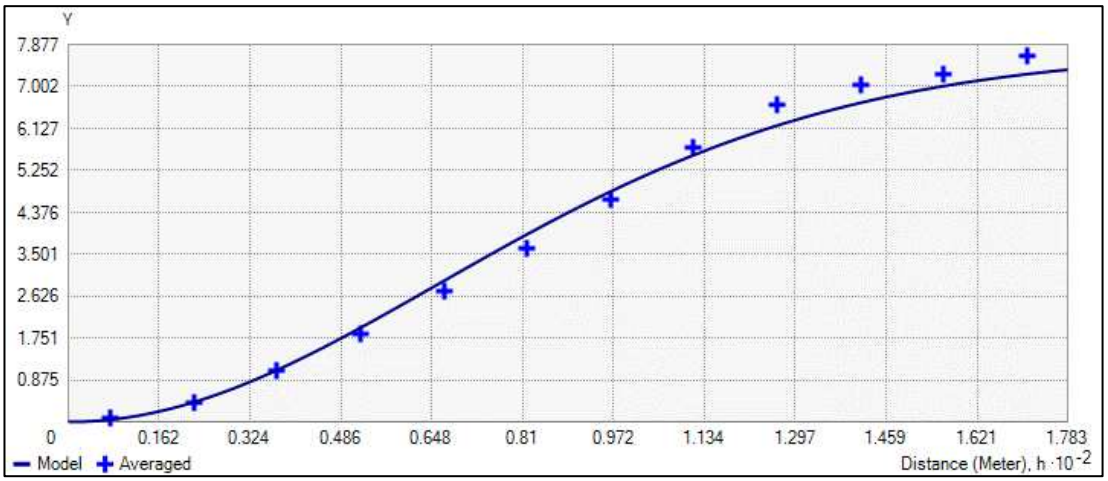

Figura 4 – Semivariograma omnidirecional experimental e modelo ajustado / Modelo Batimétrico referente à LV3.

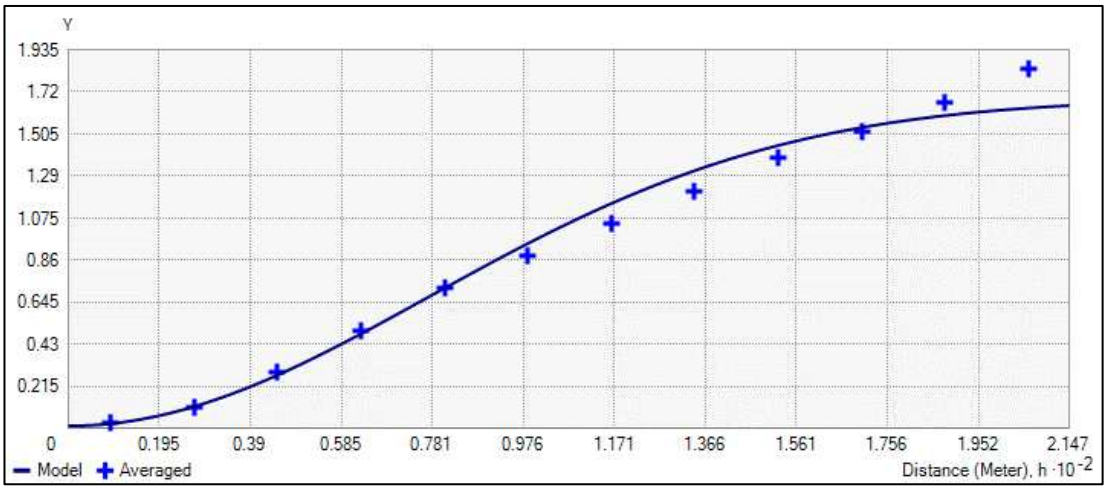

Figura 5 – Semivariograma omnidirecional experimental e modelo ajustado / Modelo Batimétrico referente à LV4.

c) Relatórios das análises geoestatísticas

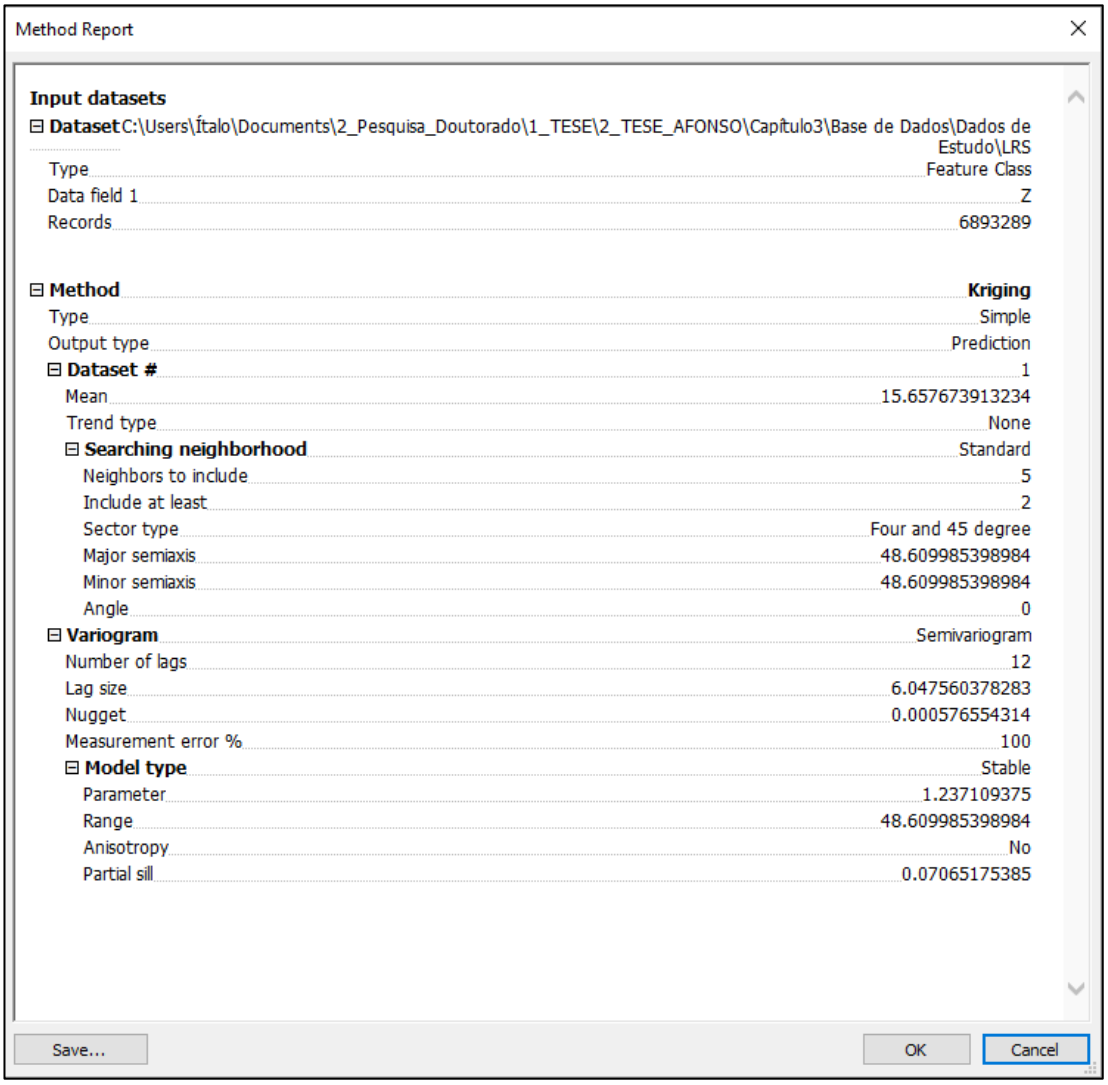

Figura 6 – Relatório da análise geoestatística / Modelo Batimétrico referente às LRS.

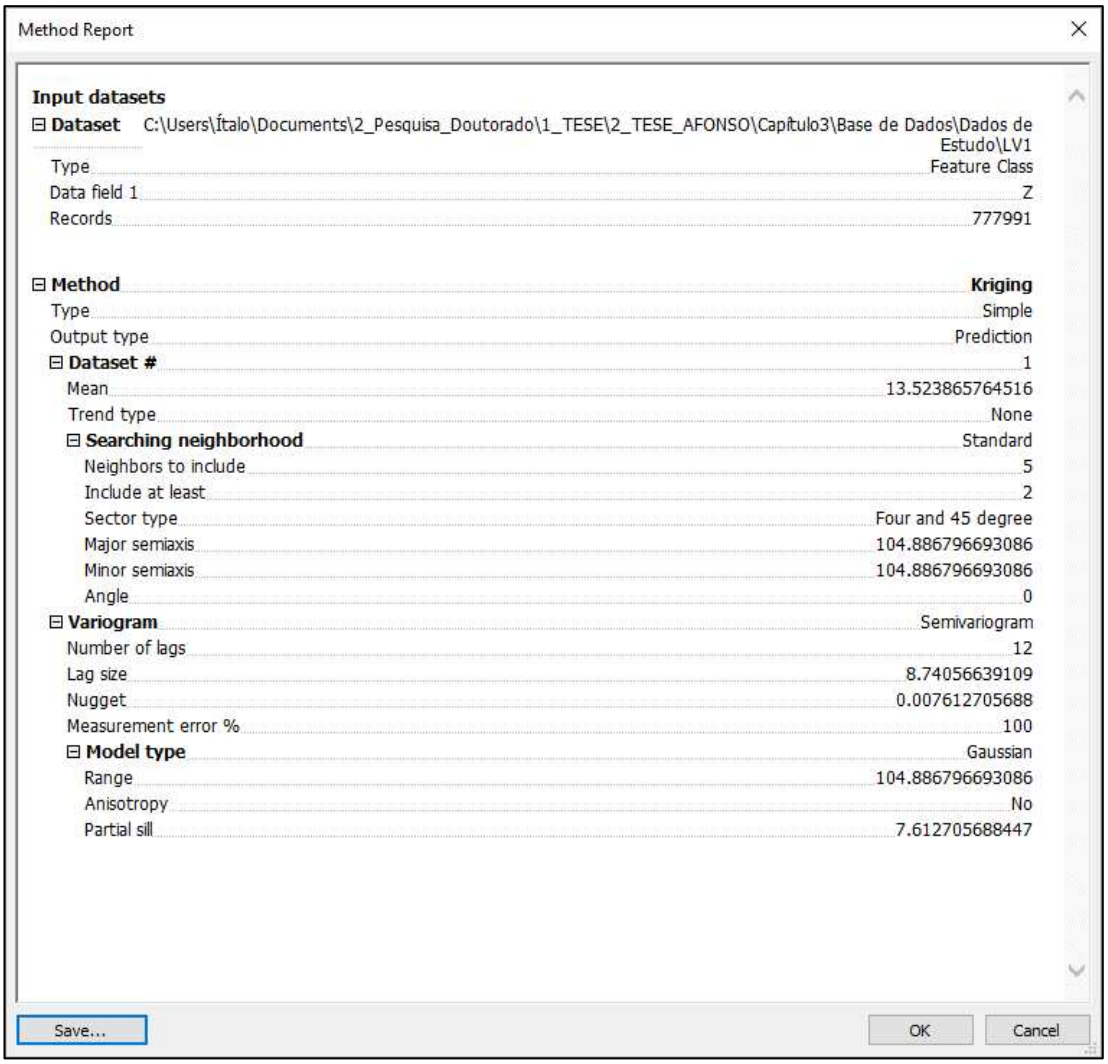

Figura 7 – Relatório da análise geoestatística / Modelo Batimétrico referente à LV1.

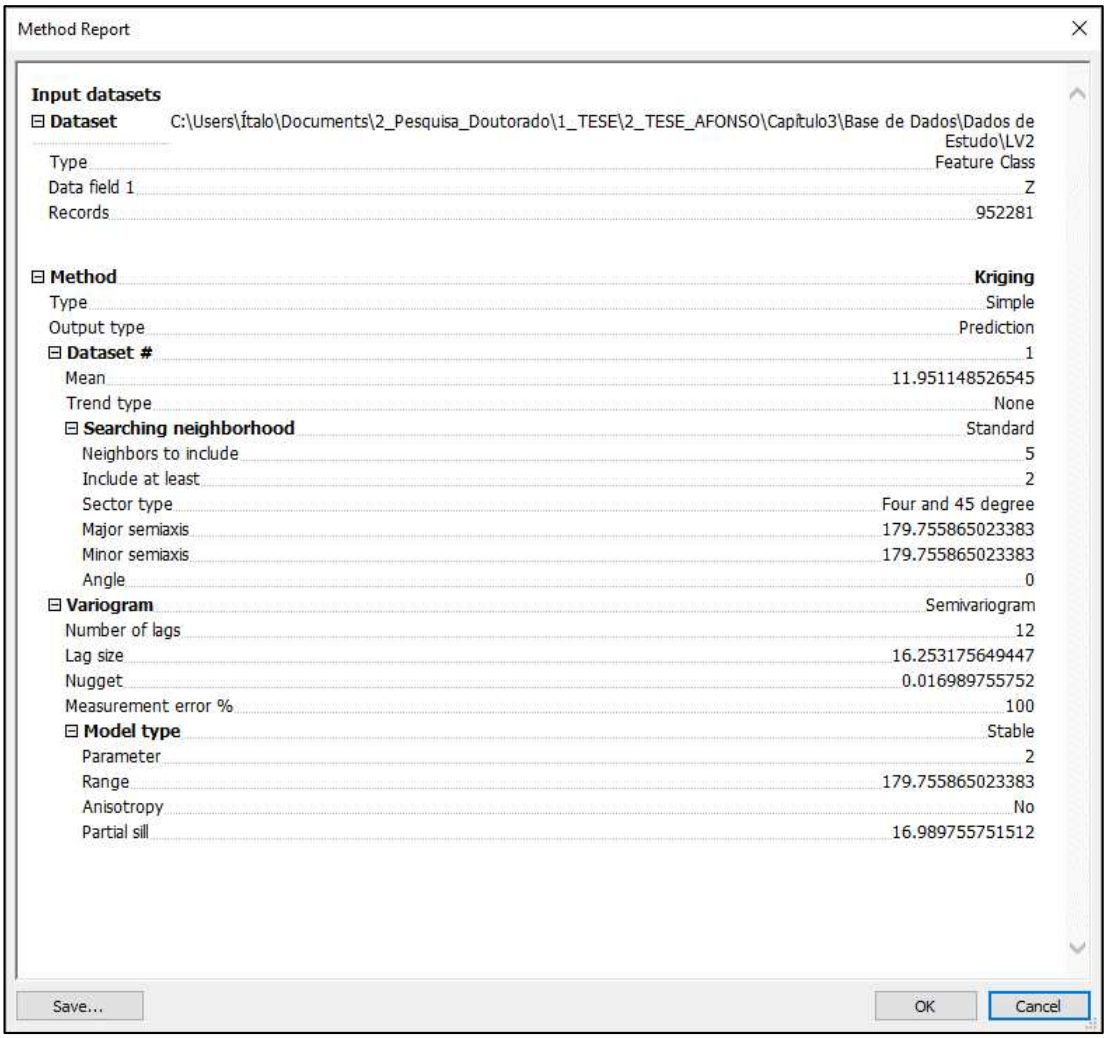

Figura 8 – Relatório da análise geoestatística / Modelo Batimétrico referente à LV2.

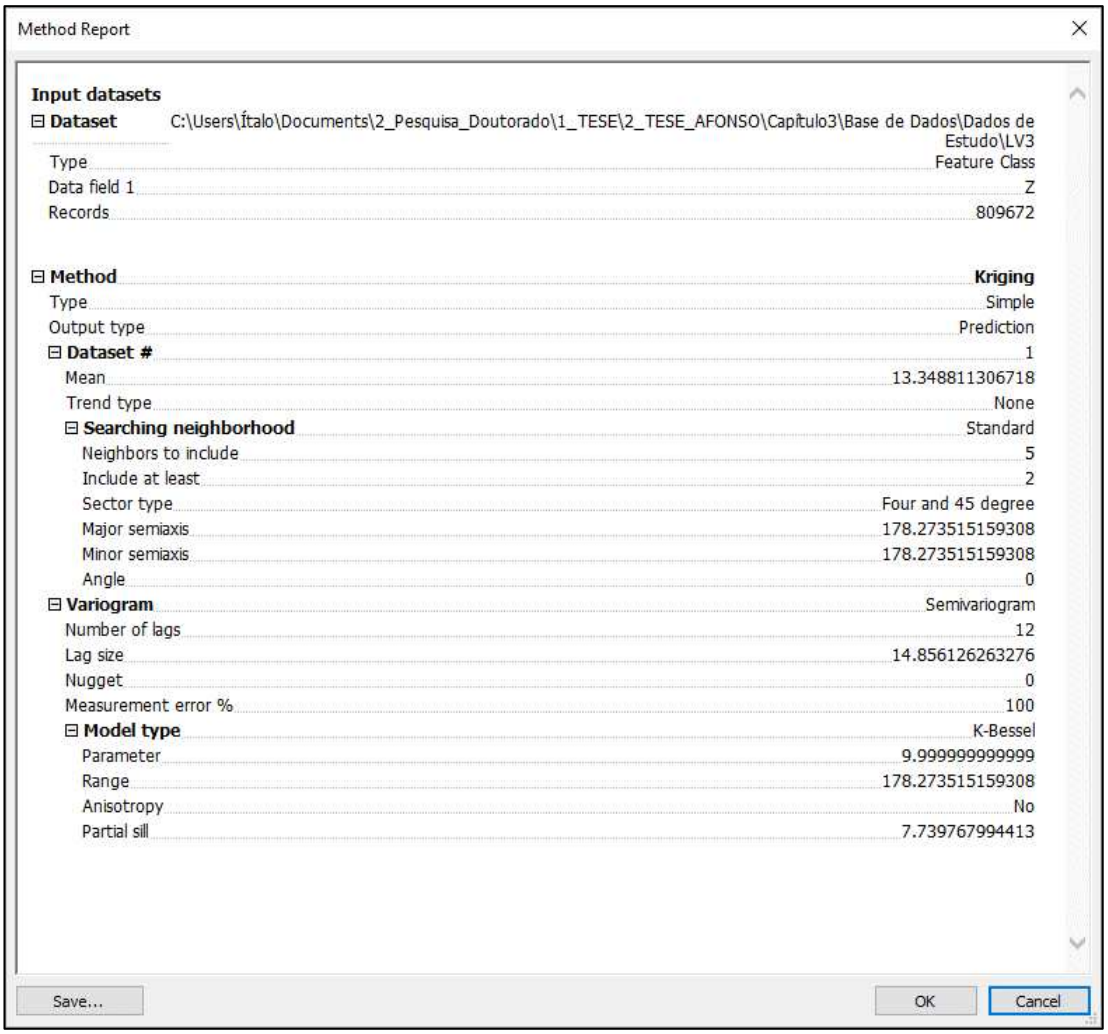

Figura 9 – Relatório da análise geoestatística / Modelo Batimétrico referente à LV3.

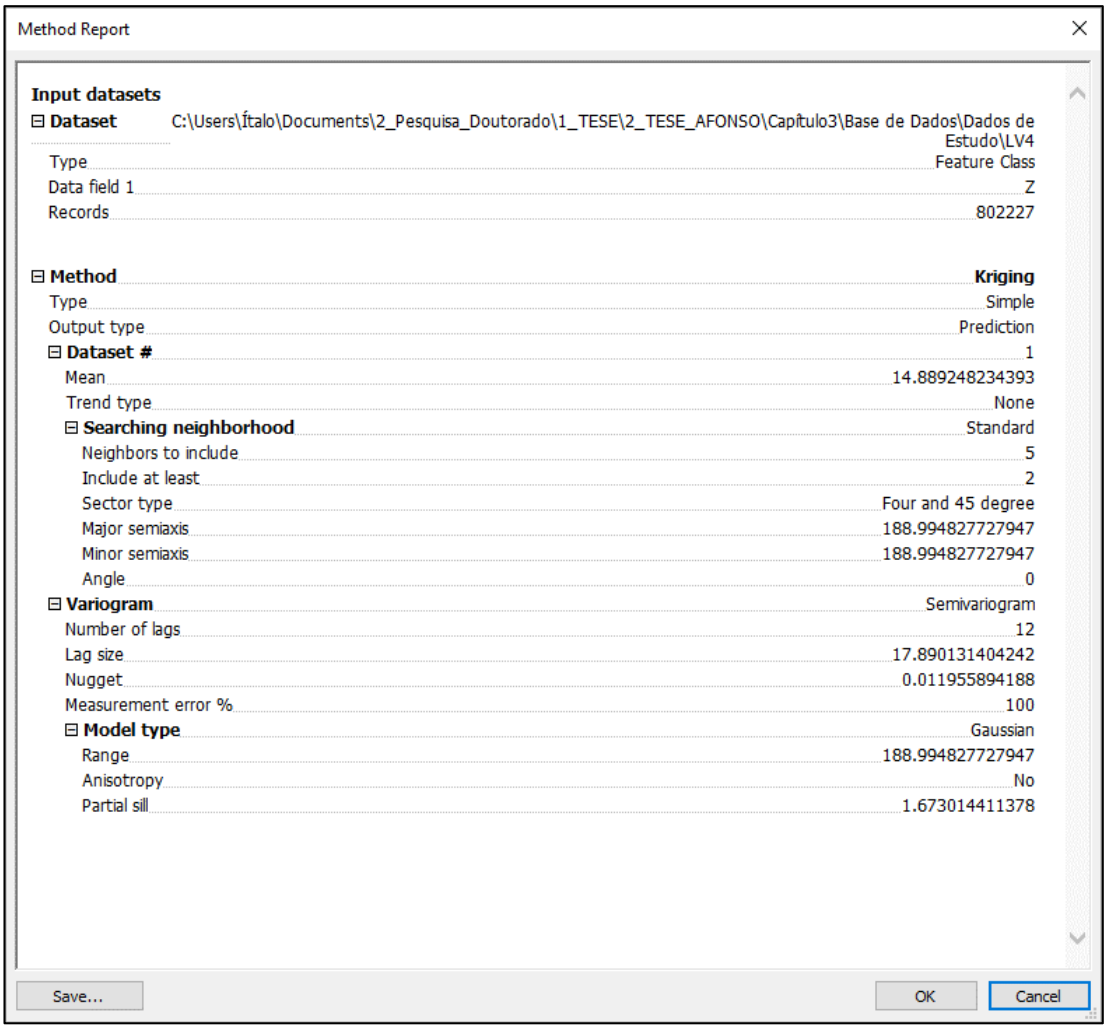

Figura 10 – Relatório da análise geoestatística / Modelo Batimétrico referente à LV4.

d) Apresentação dos *outliers* detectados nas amostras de discrepâncias por meio do emprego do AEDO/Método δ

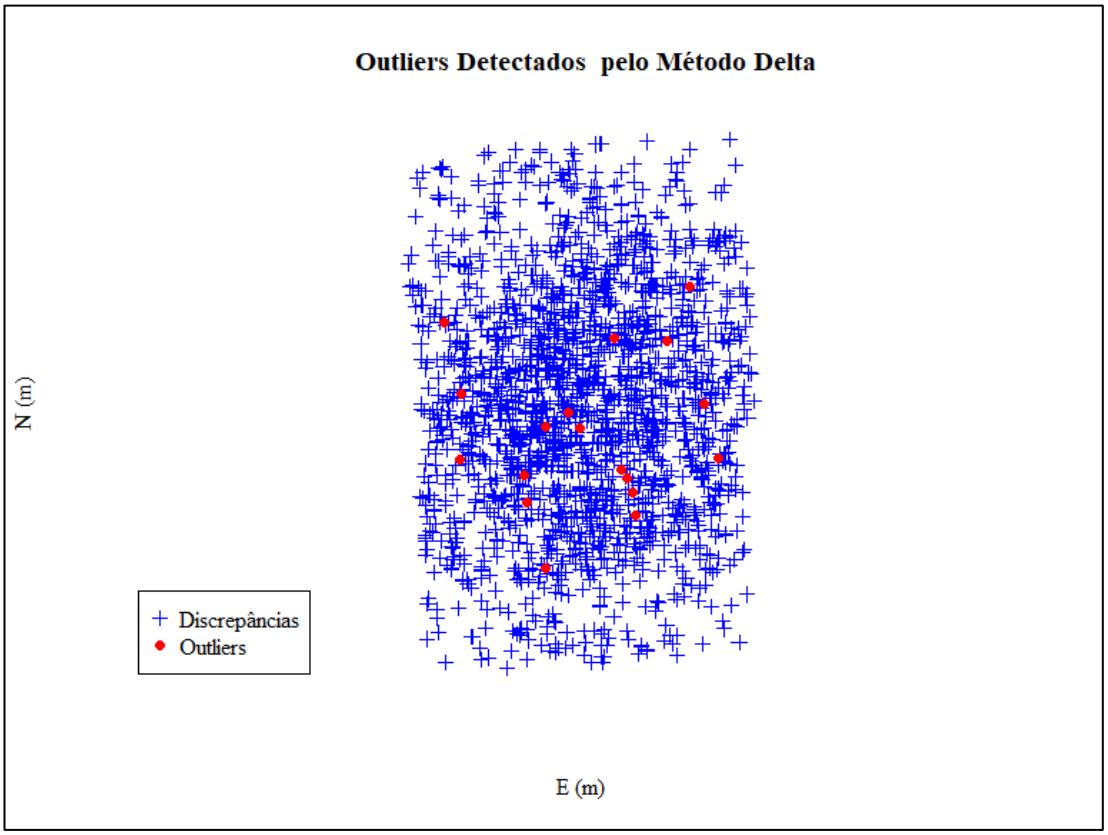

Figura 11 – *Outliers* detectados na amostra de discrepâncias dp1.

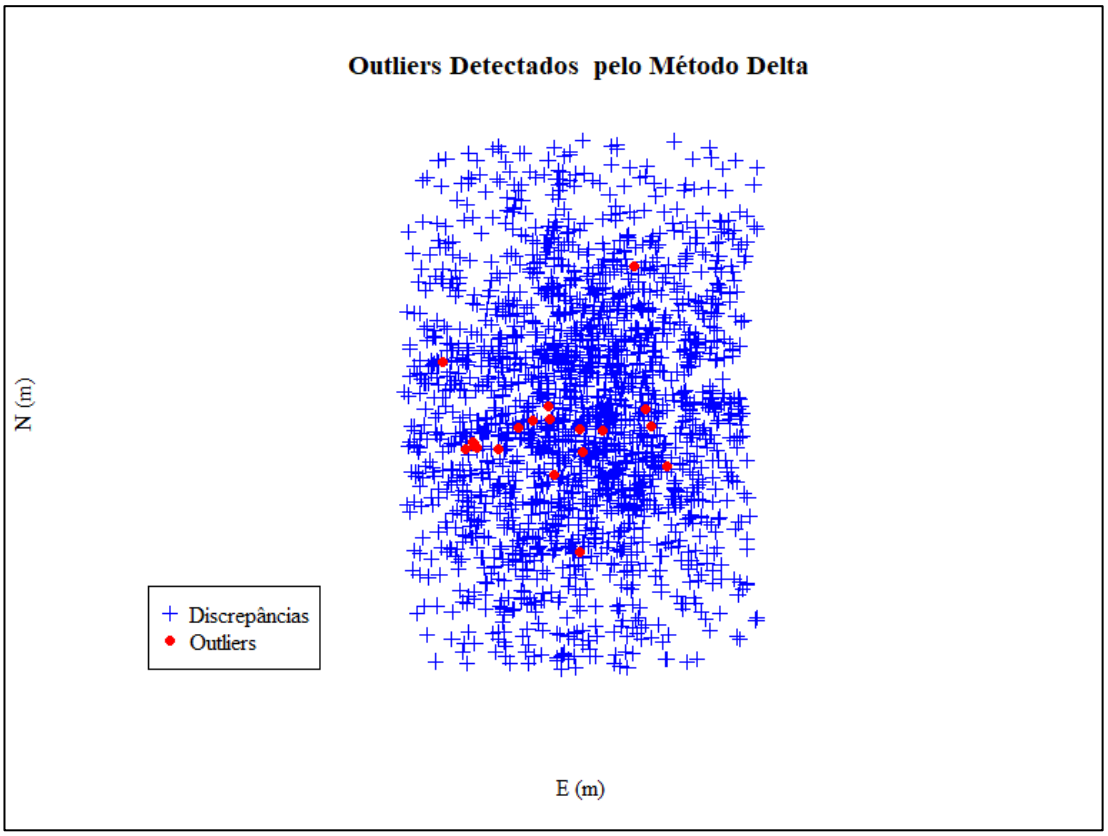

Figura 12 – *Outliers* detectados na amostra de discrepâncias dp2.

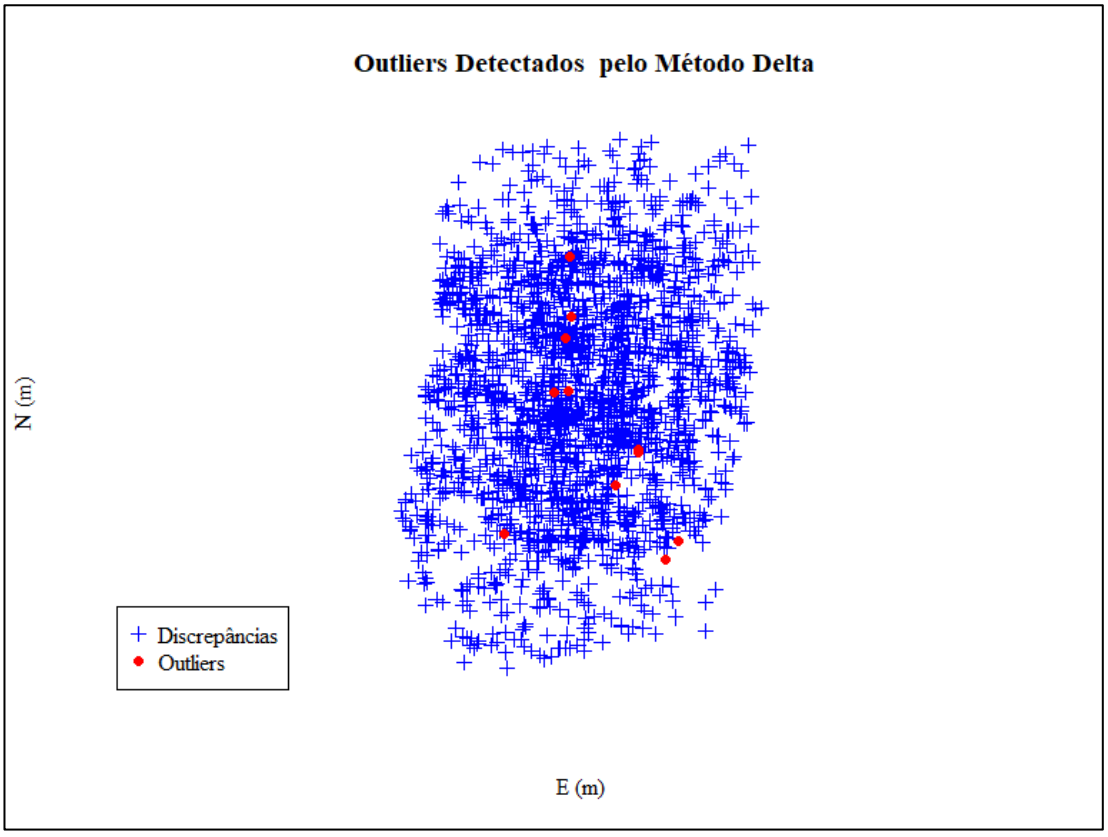

Figura 13 – *Outliers* detectados na amostra de discrepâncias dp3.

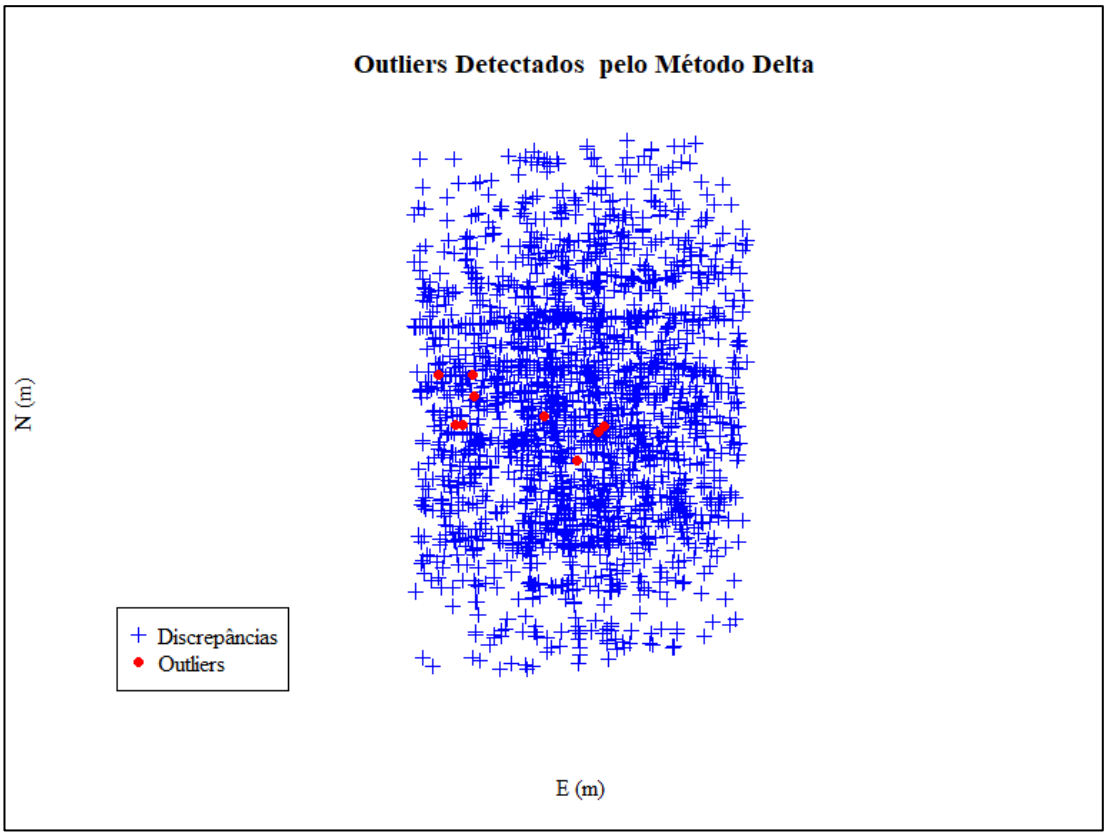

Figura 14 – *Outliers* detectados na amostra de discrepâncias dp4.

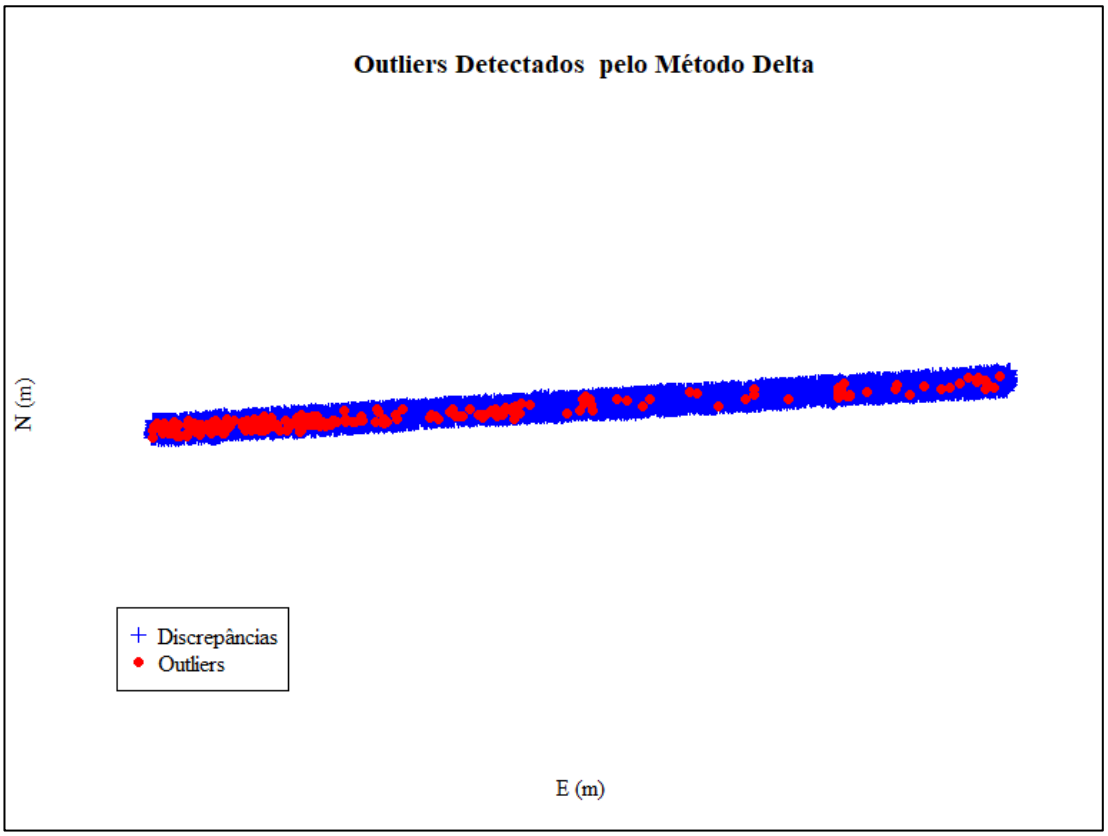

Figura 15 – *Outliers* detectados na amostra de discrepâncias dp5.

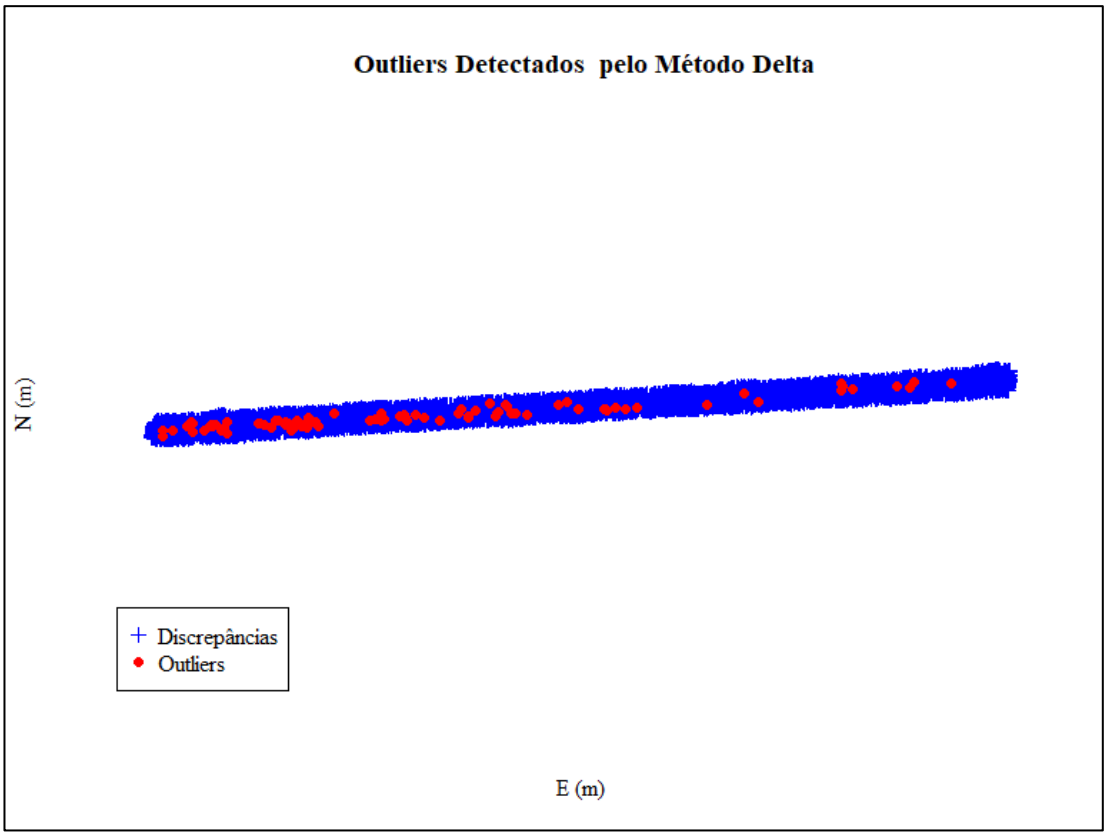

Figura 16 – *Outliers* detectados na amostra de discrepâncias dp6.

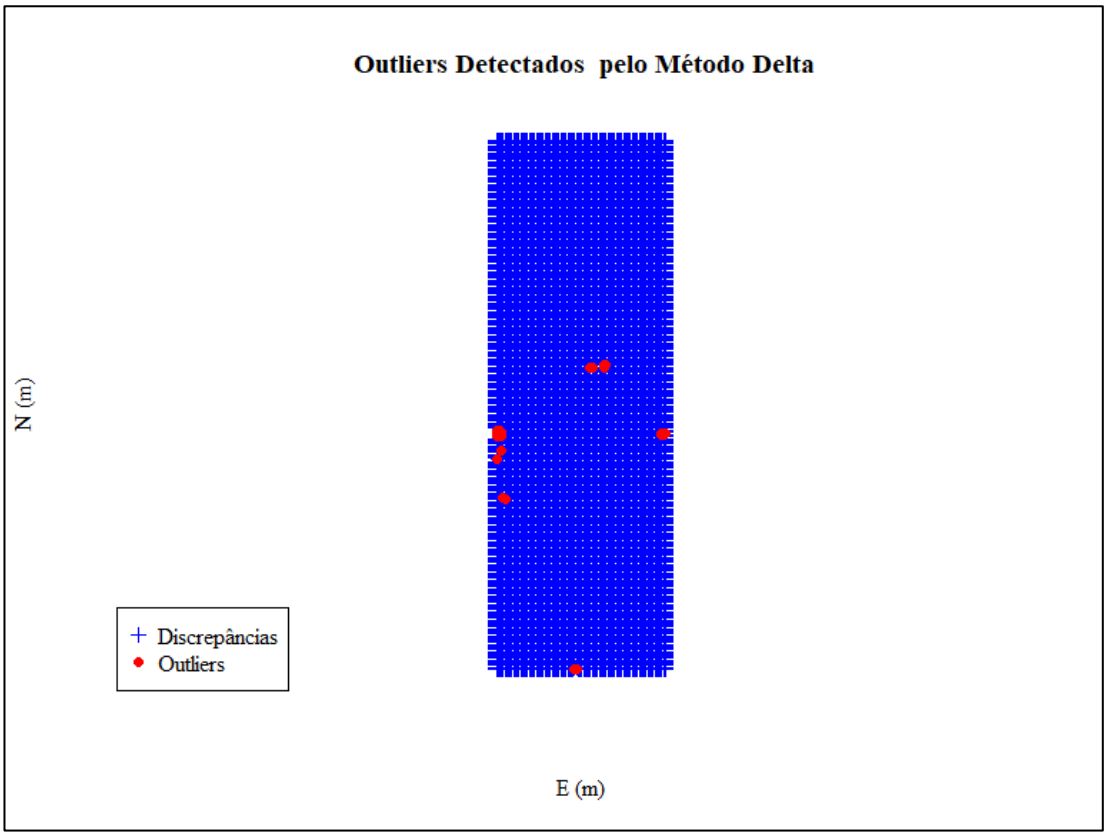

Figura 17 – *Outliers* detectados na amostra de discrepâncias dp1_ss.

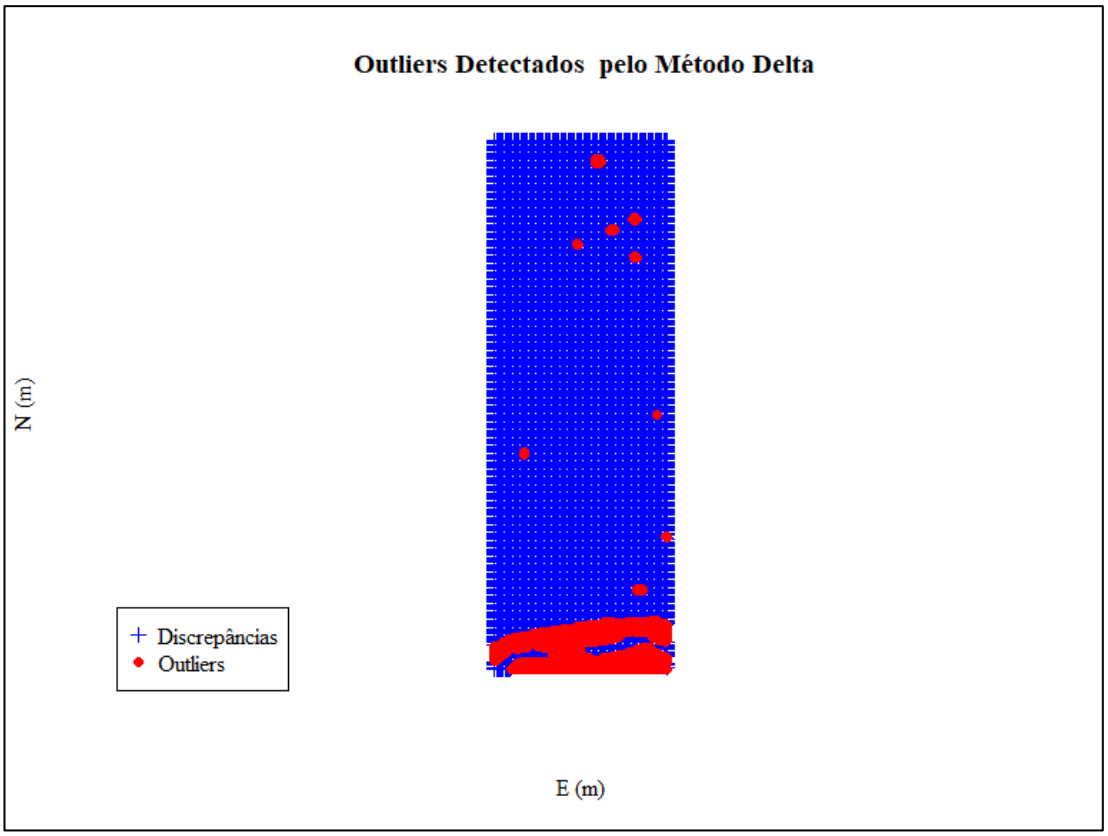

Figura 18 – *Outliers* detectados na amostra de discrepâncias dp2_ss.

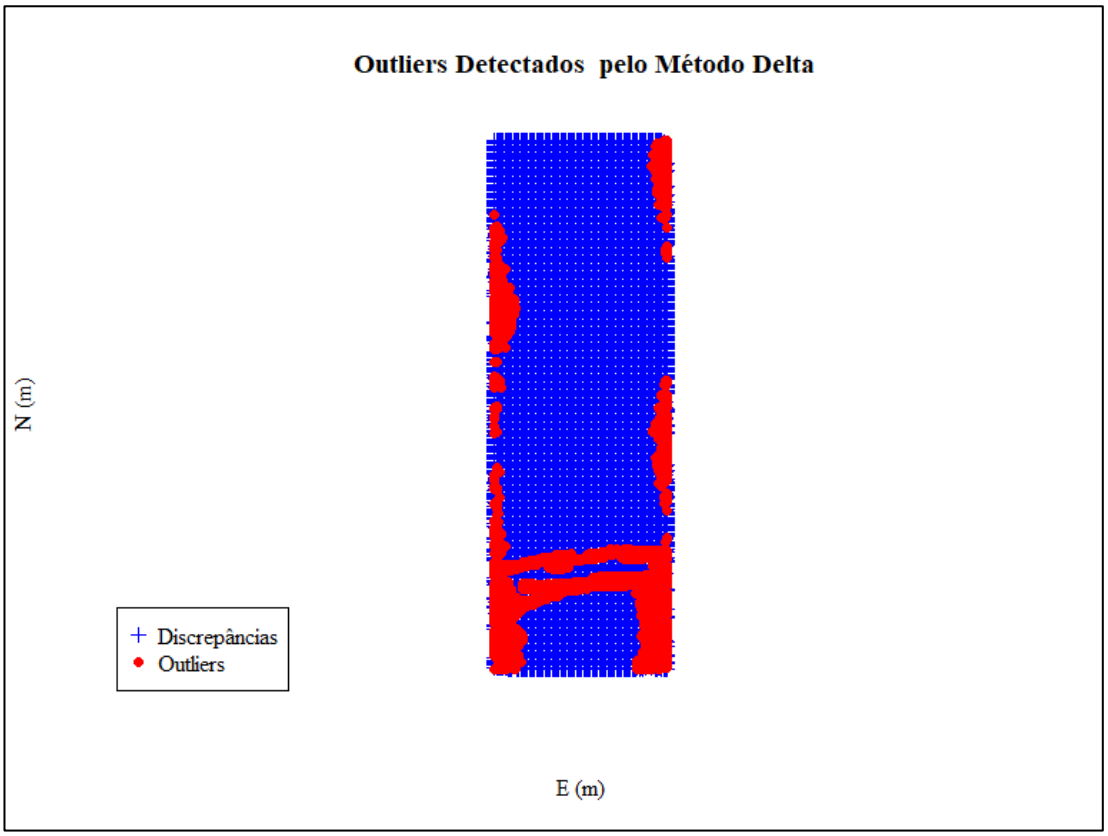

Figura 19 – *Outliers* detectados na amostra de discrepâncias dp3_ss.

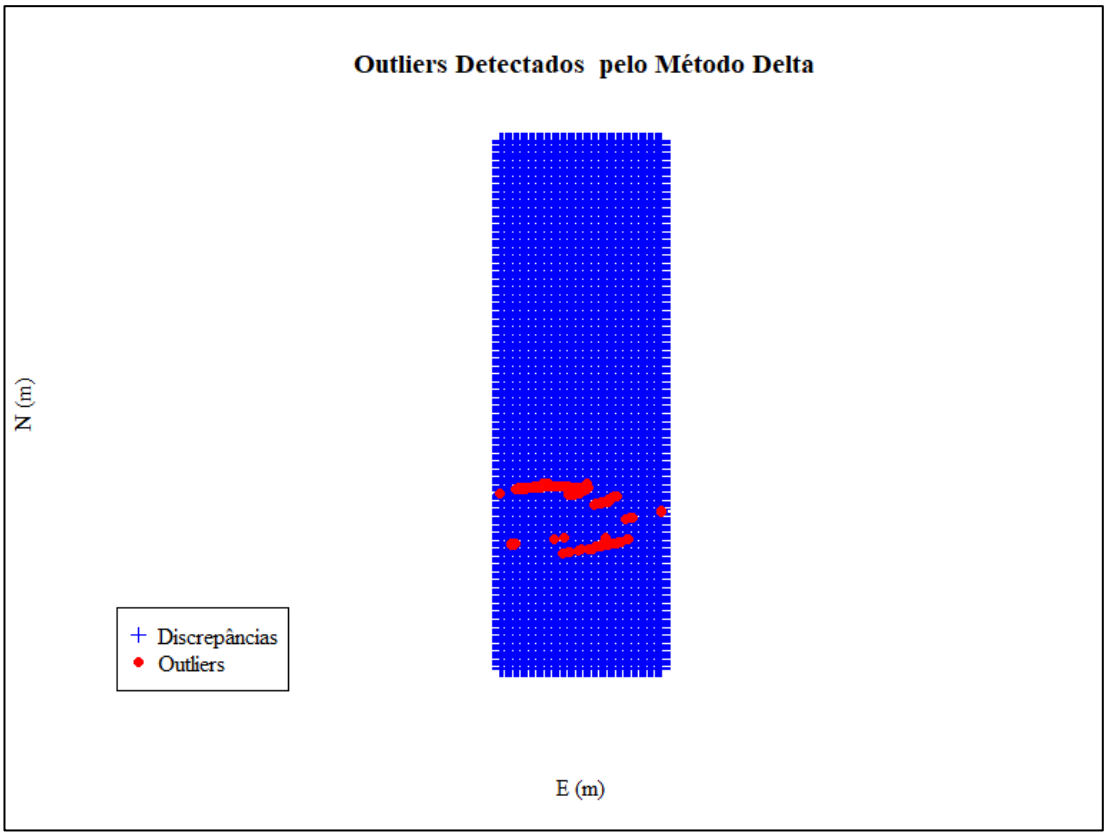

Figura 20 – *Outliers* detectados na amostra de discrepâncias dp4_ss.

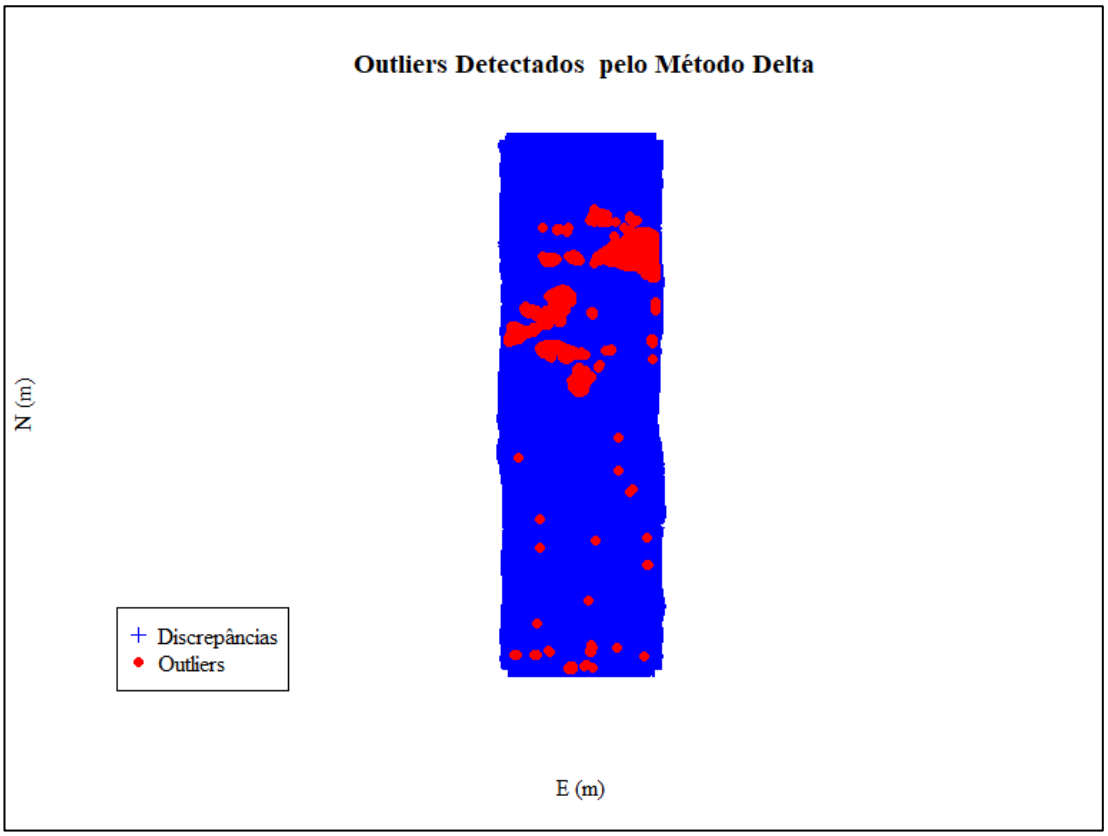

Figura 21 – *Outliers* detectados na amostra de discrepâncias dp1_sp.

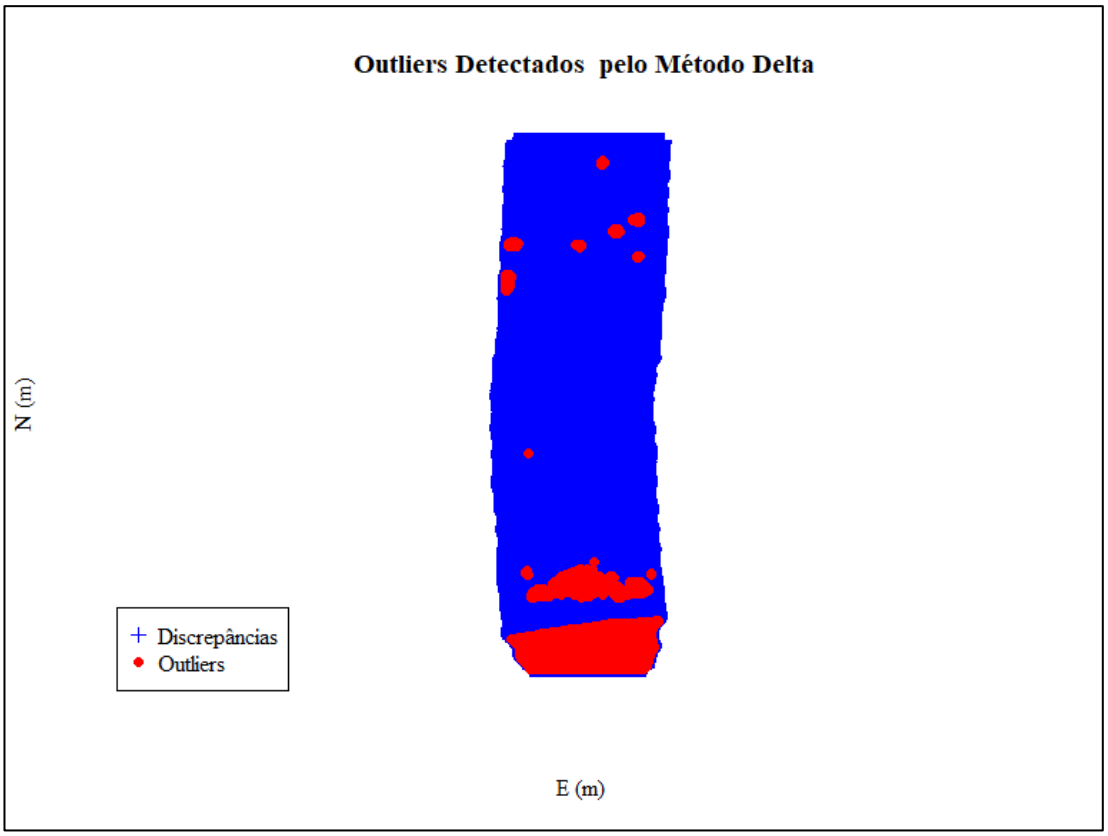

Figura 22 – *Outliers* detectados na amostra de discrepâncias dp2_sp.

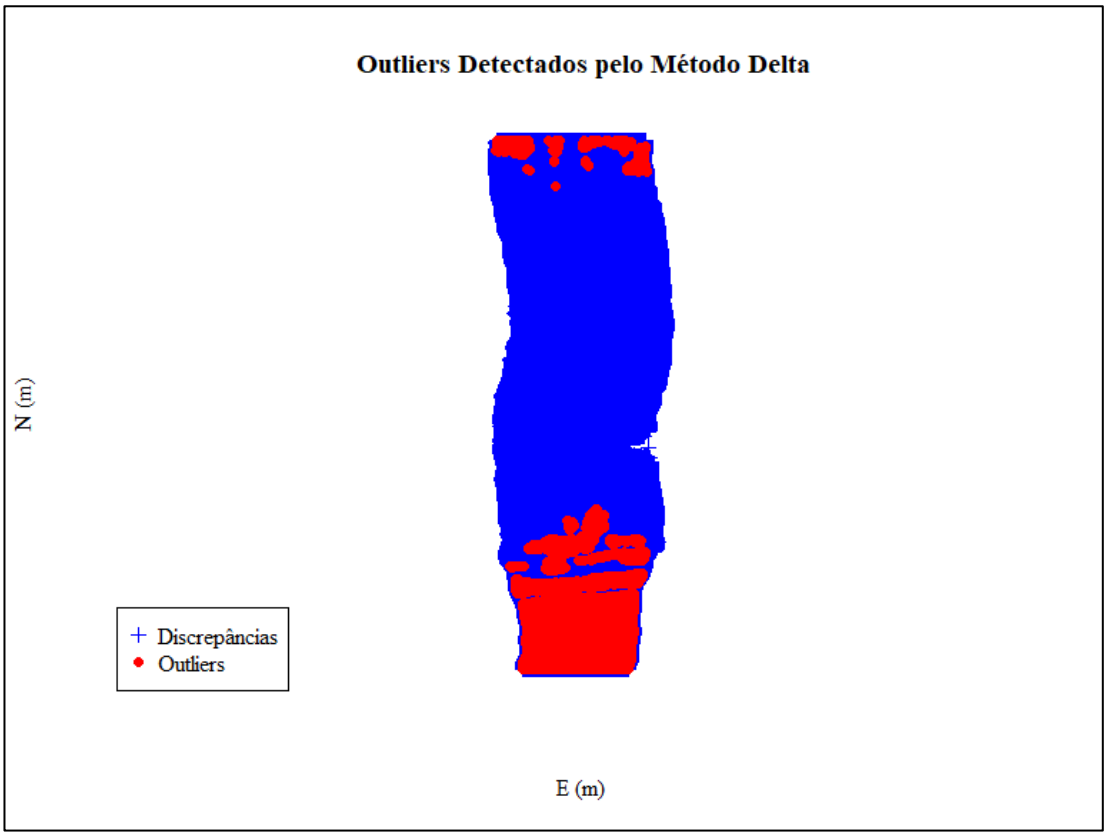

Figura 23 – *Outliers* detectados na amostra de discrepâncias dp3_sp.

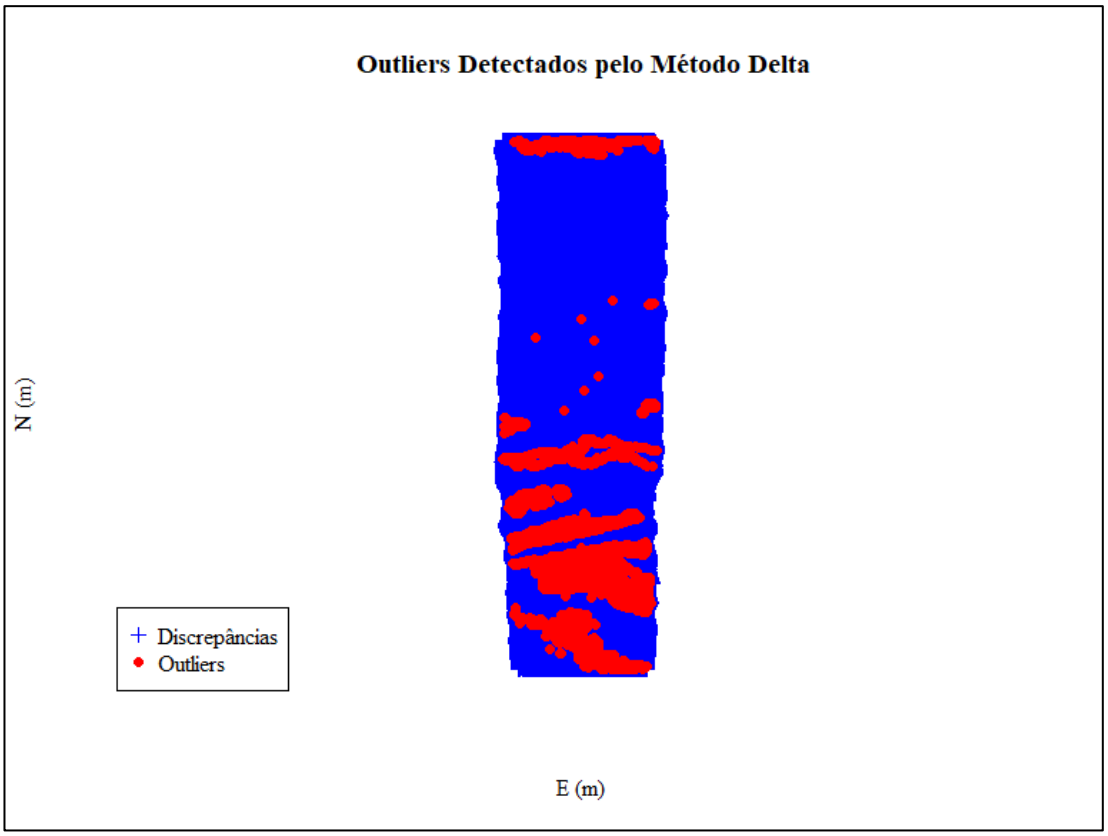

Figura 24 – *Outliers* detectados na amostra de discrepâncias dp4_sp.

e) Análise gráfica exploratória das discrepâncias (dados sem *outliers*)

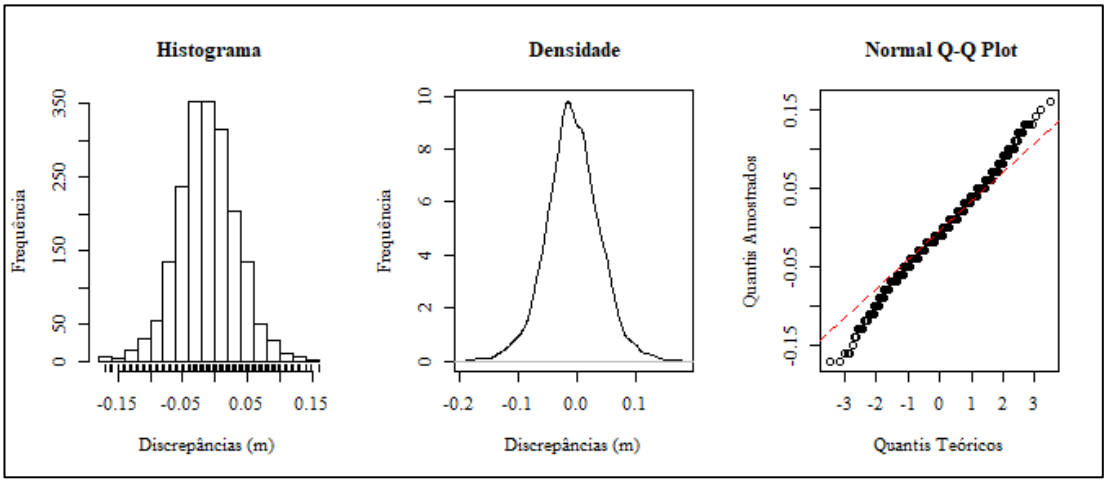

Figura 25 – Análise gráfica exploratória / amostra dp1.

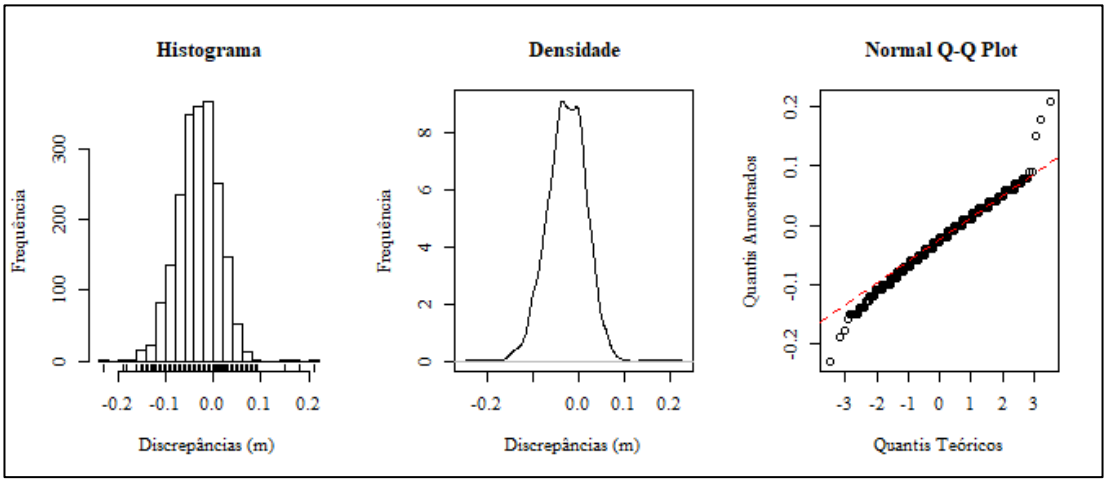

Figura 26 – Análise gráfica exploratória / amostra dp2.

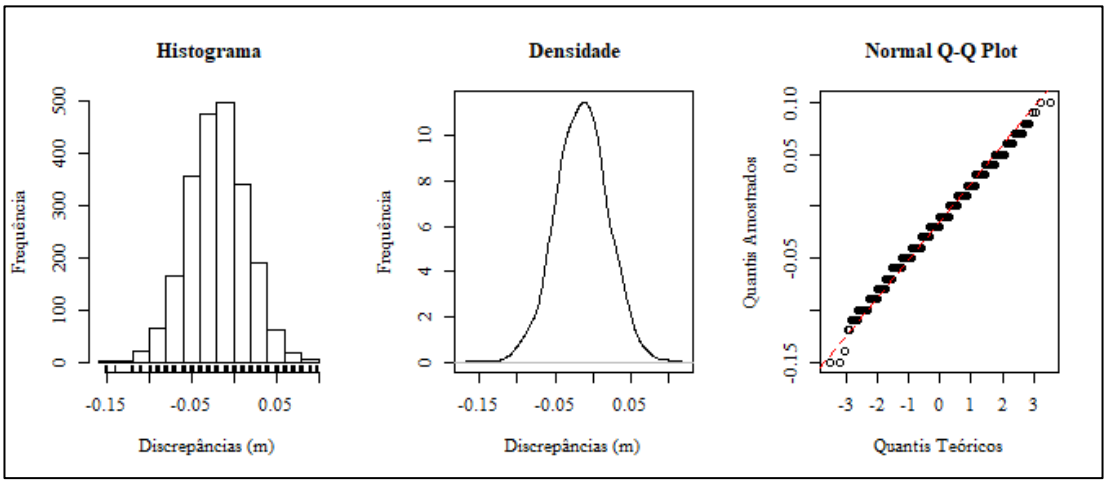

Figura 27 – Análise gráfica exploratória / amostra dp3.

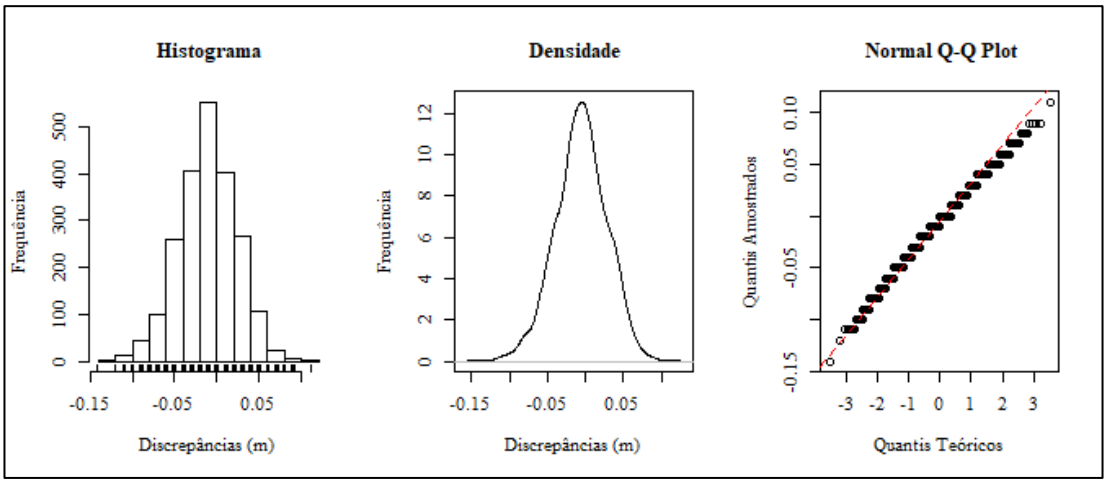

Figura 28 – Análise gráfica exploratória / amostra dp4.

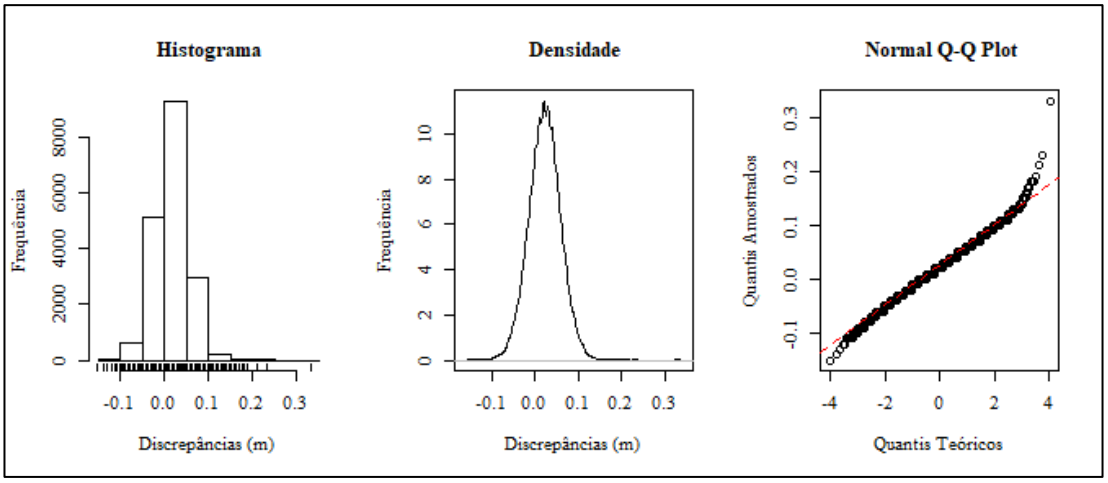

Figura 29 – Análise gráfica exploratória / amostra dp5.

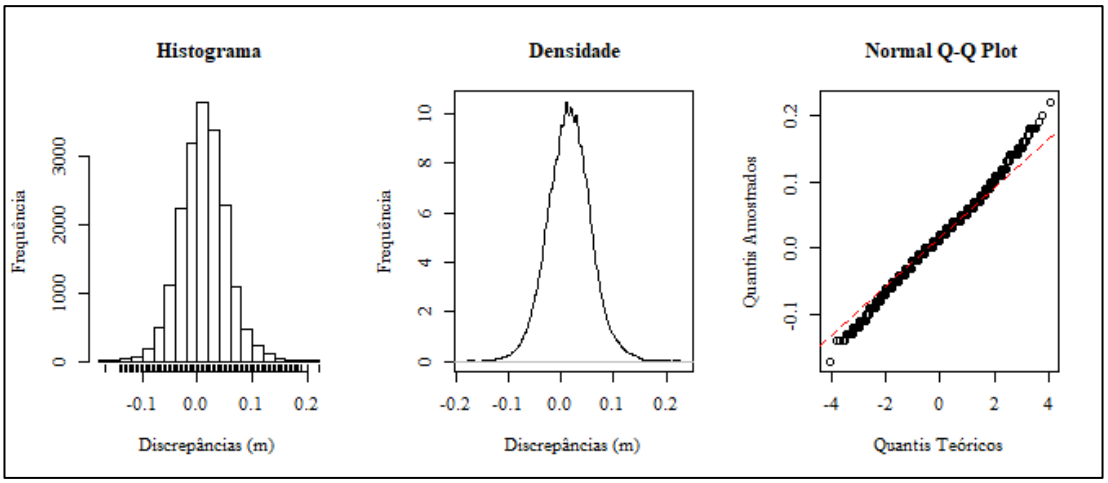

Figura 30 – Análise gráfica exploratória / amostra dp6.

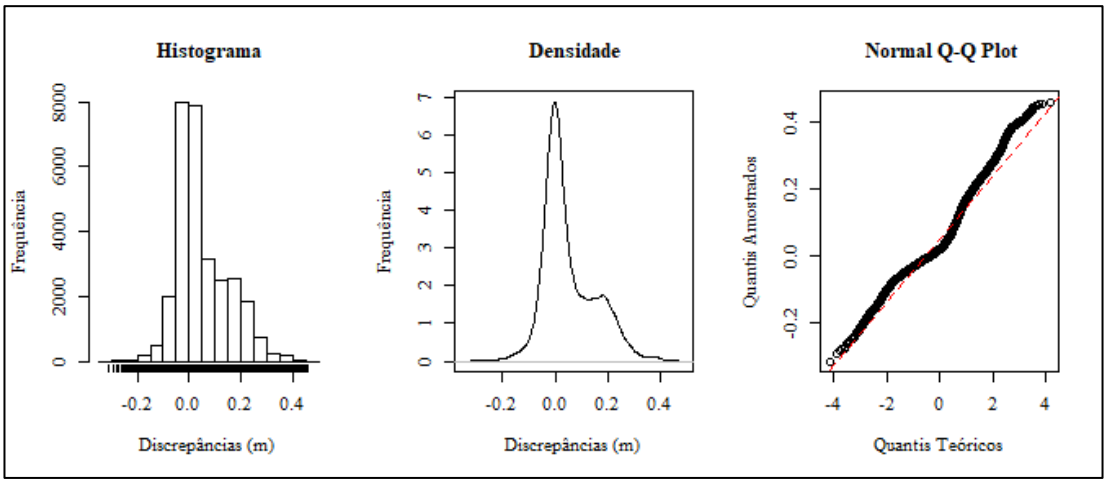

Figura 31 – Análise gráfica exploratória / amostra dp1_ss.

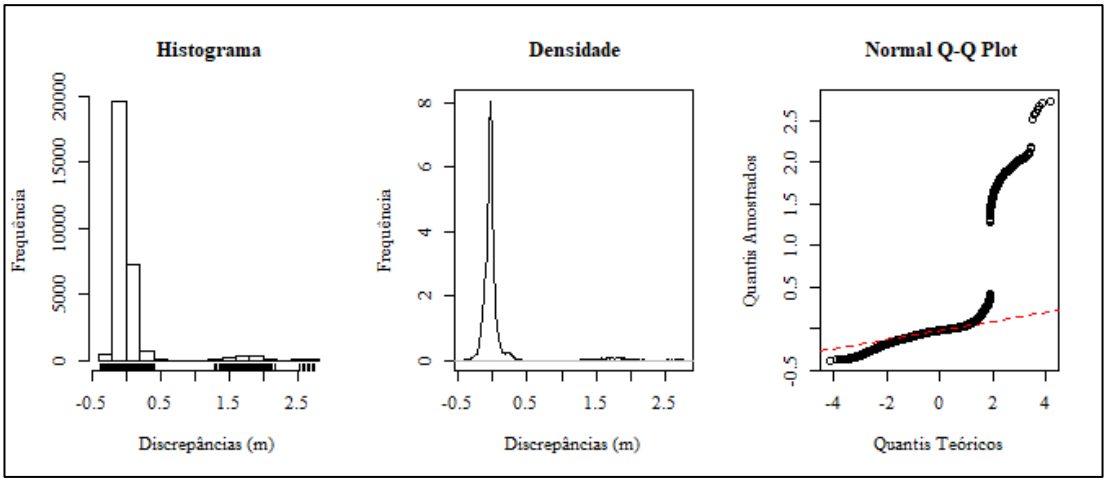

Figura 32 – Análise gráfica exploratória / amostra dp2_ss.

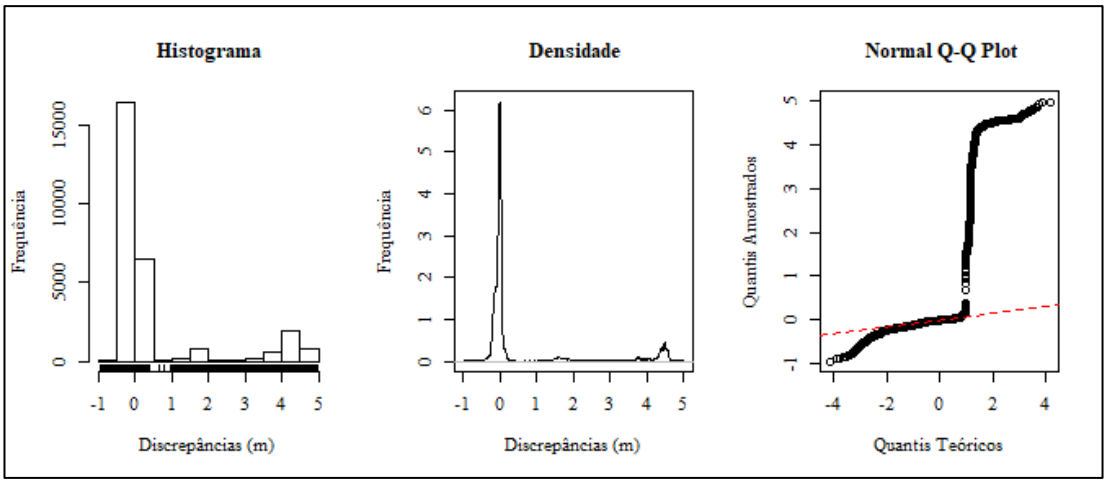

Figura 33 – Análise gráfica exploratória / amostra dp3_ss.

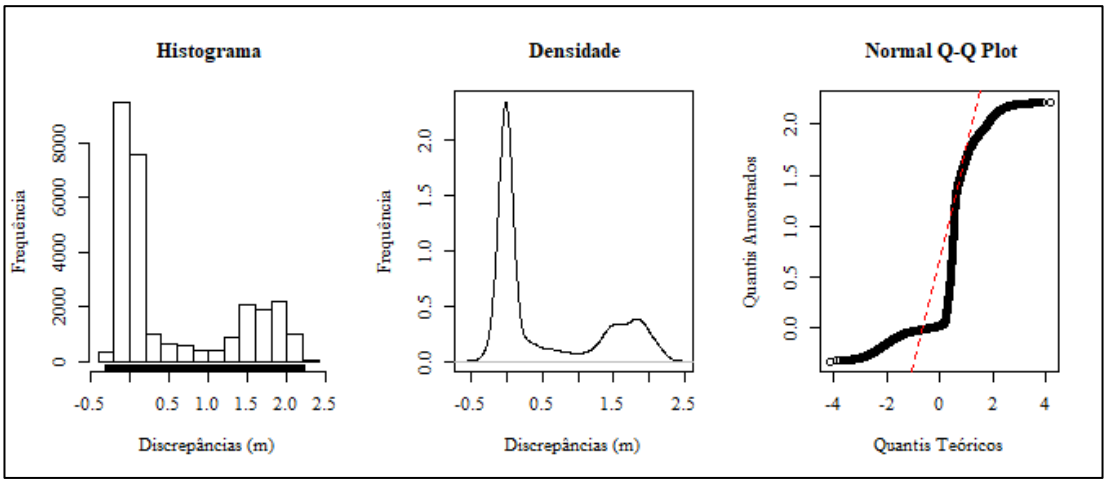

Figura 34 – Análise gráfica exploratória / amostra dp4_ss.

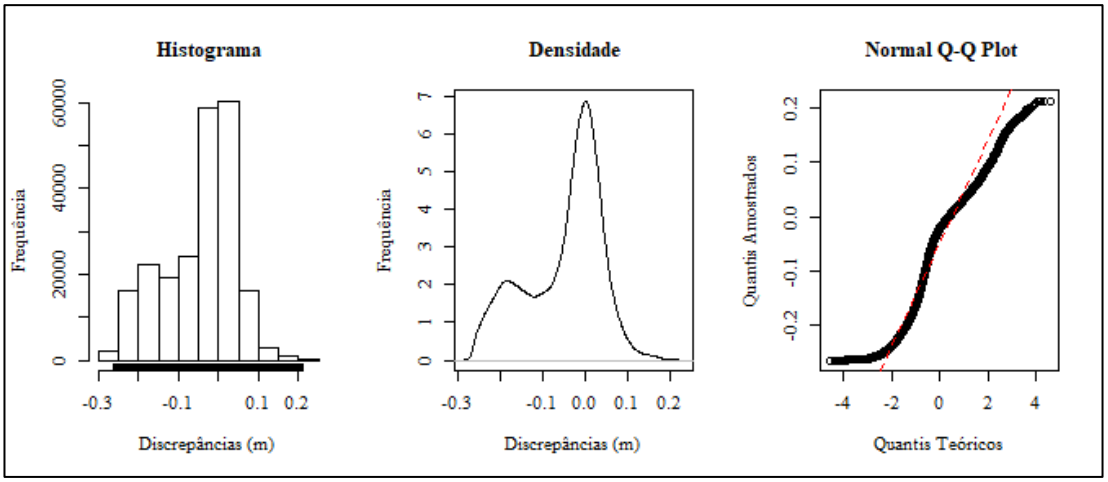

Figura 35 – Análise gráfica exploratória / amostra dp1_sp.

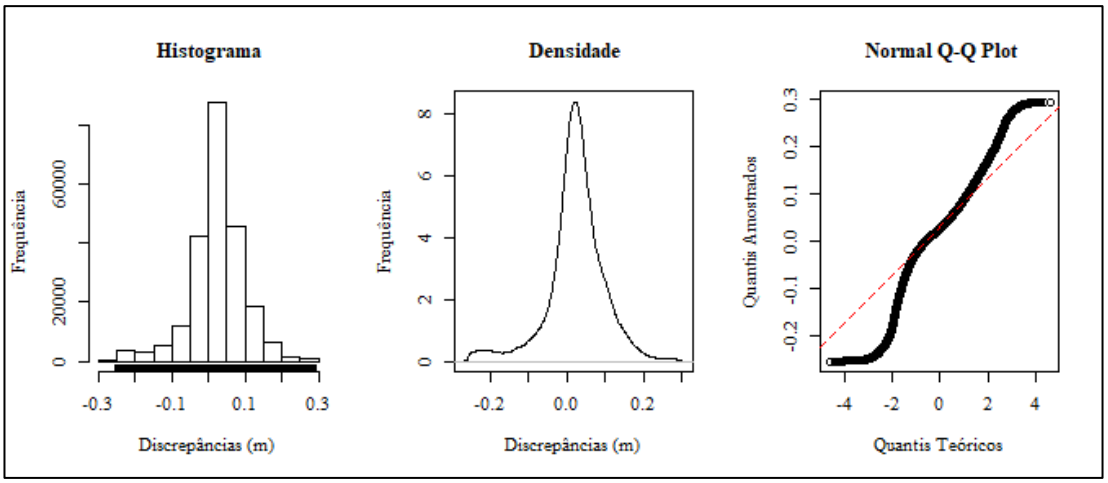

Figura 36 – Análise gráfica exploratória / amostra dp2_sp.

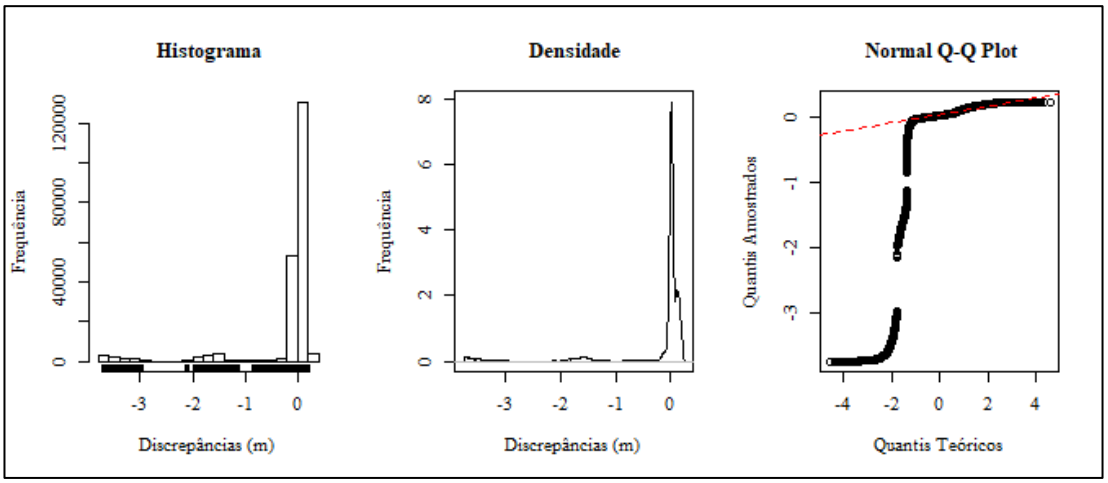

Figura 37 – Análise gráfica exploratória / amostra dp3_sp.

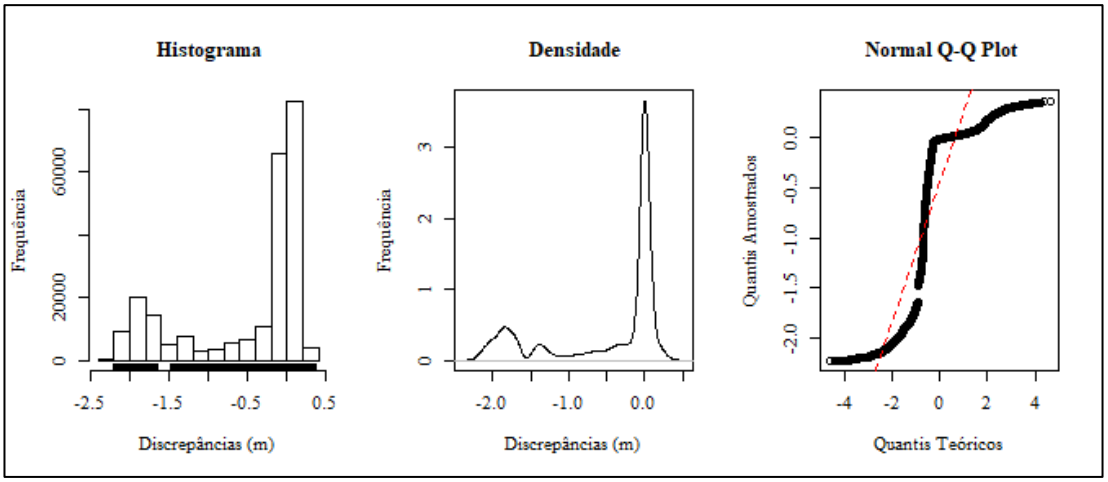

Figura 38 – Análise gráfica exploratória / amostra dp4_sp.

f) Análise de independência – Semivariogramas das discrepâncias (dados sem *outliers*).

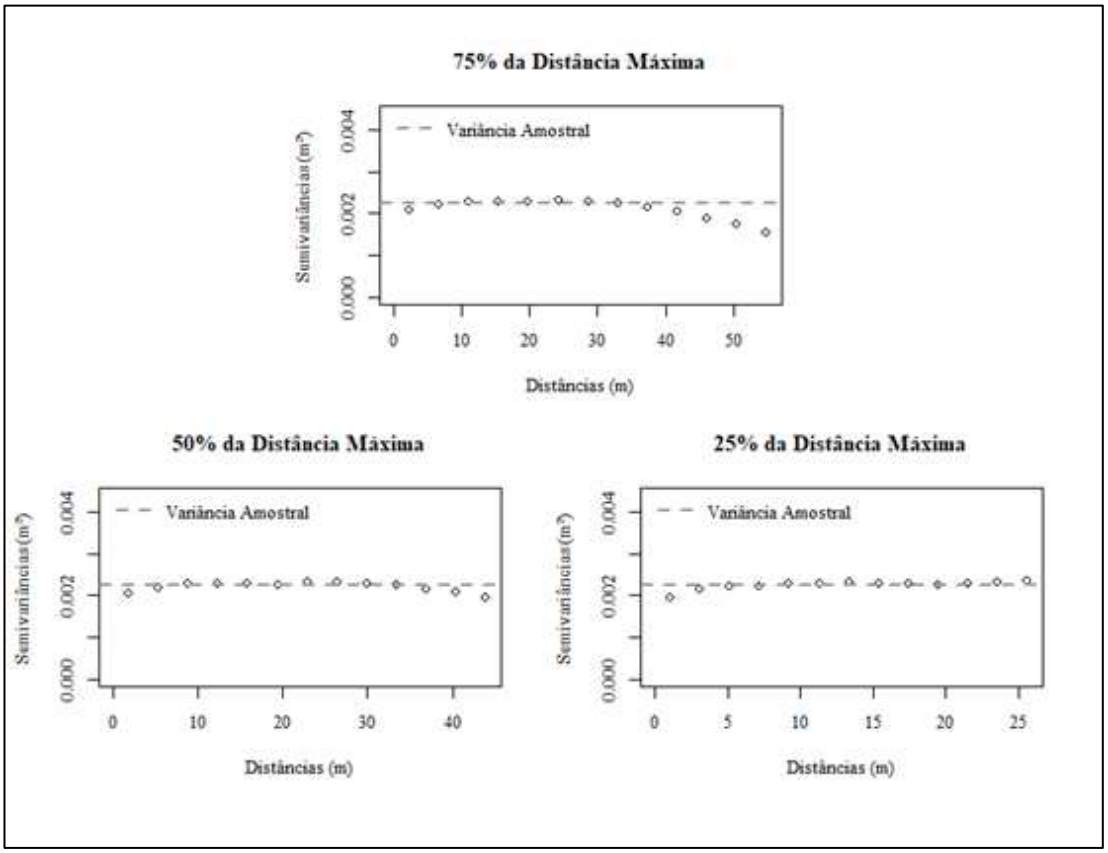

Figura 39 – Semivariograma das discrepâncias / amostra dp1.

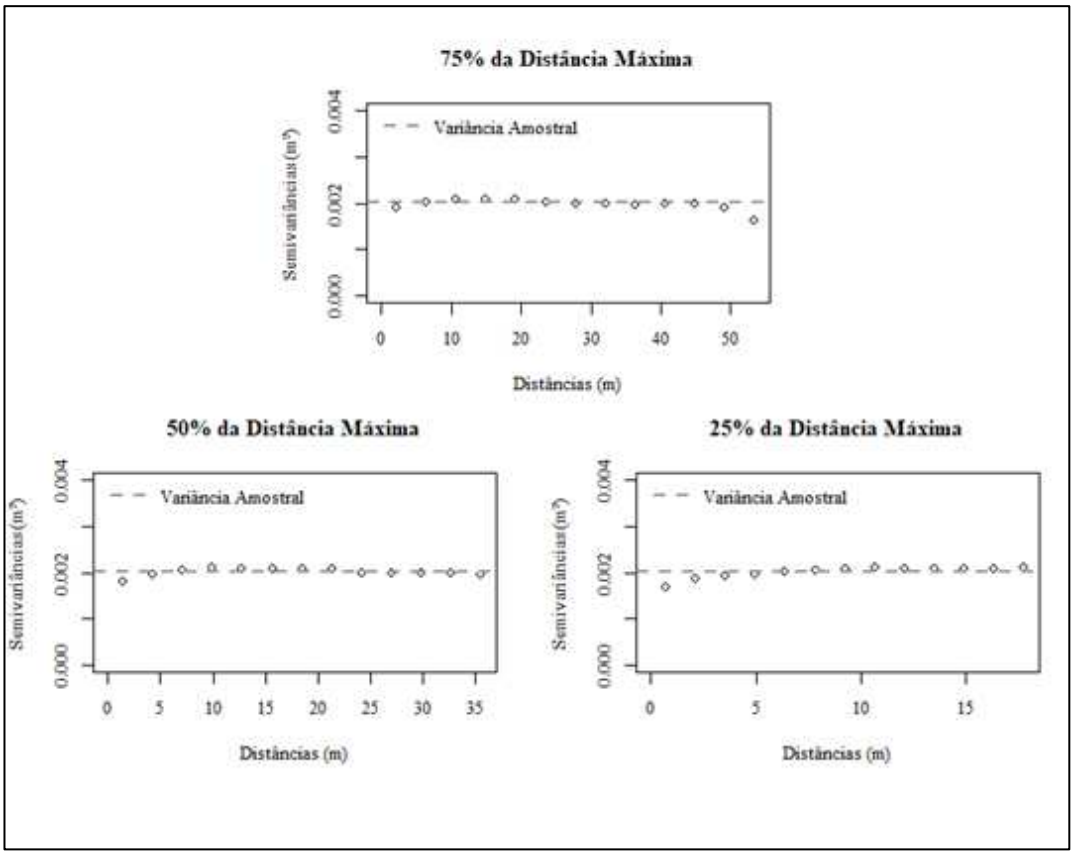

Figura 40 – Semivariograma das discrepâncias / amostra dp2.

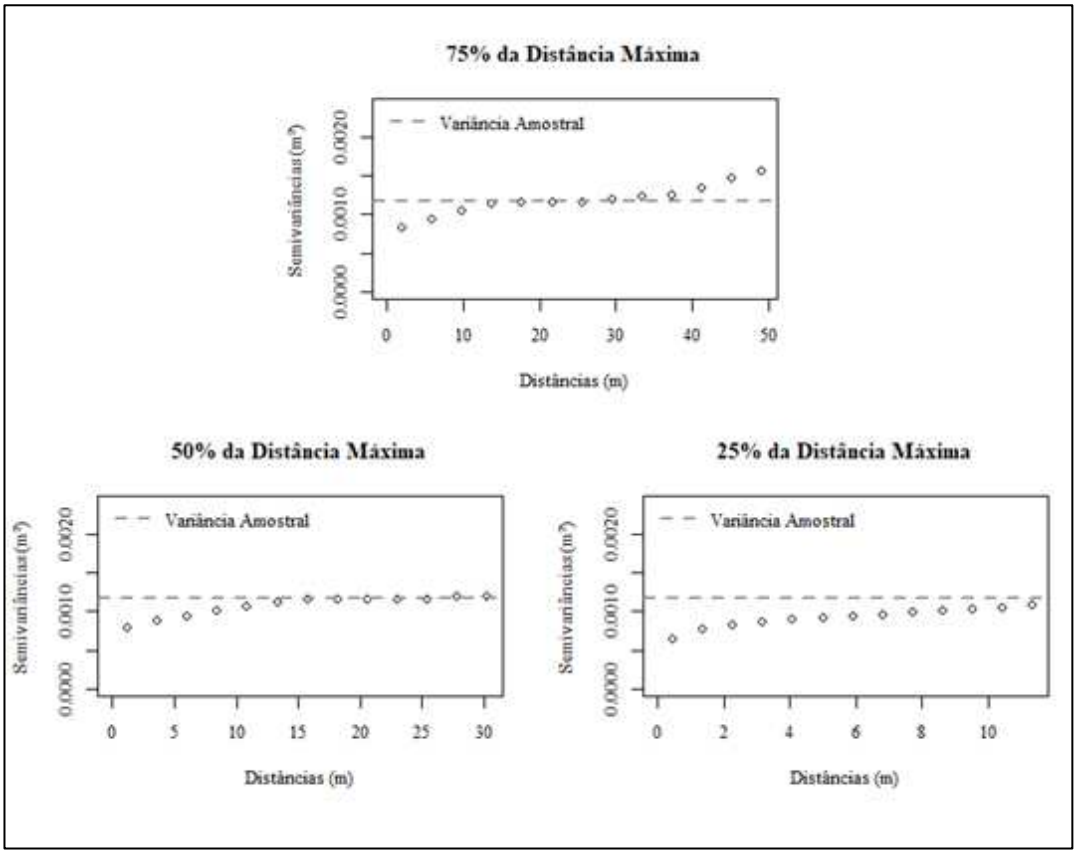

Figura 41 – Semivariograma das discrepâncias / amostra dp3.

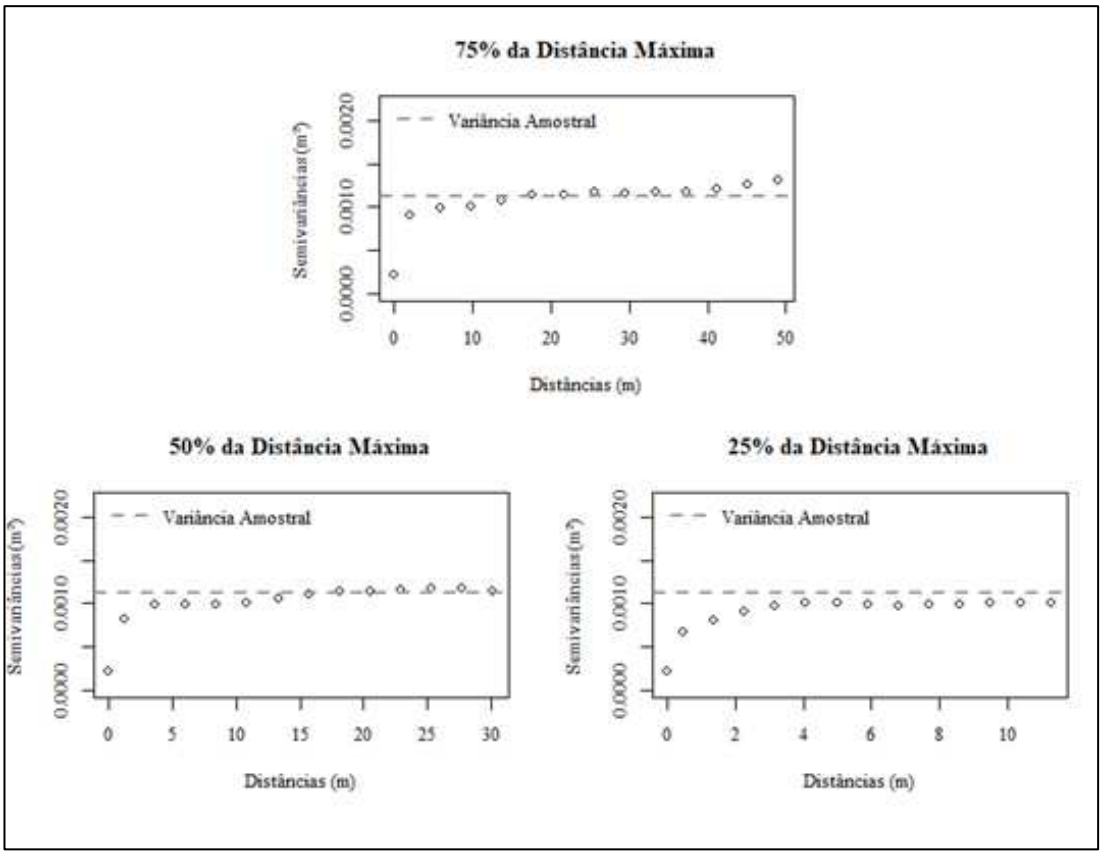

Figura 42 – Semivariograma das discrepâncias / amostra dp4.

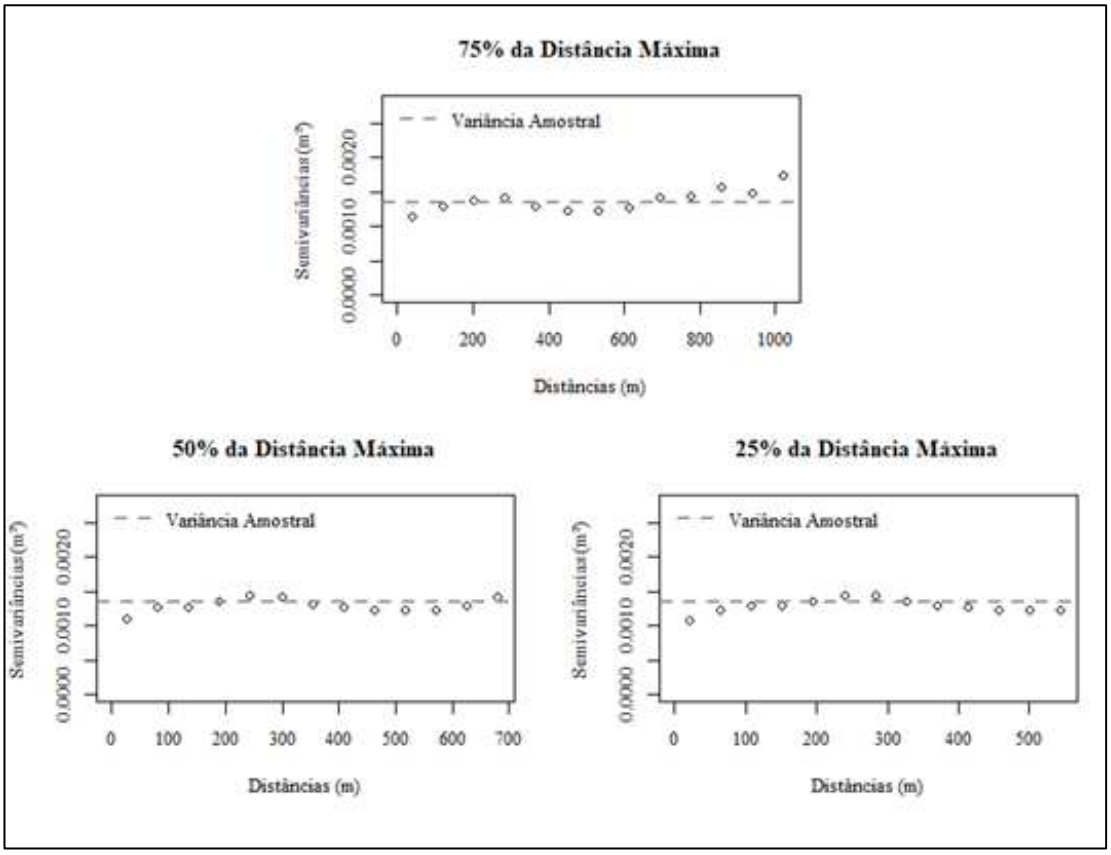

Figura 43 – Semivariograma das discrepâncias / amostra dp5.

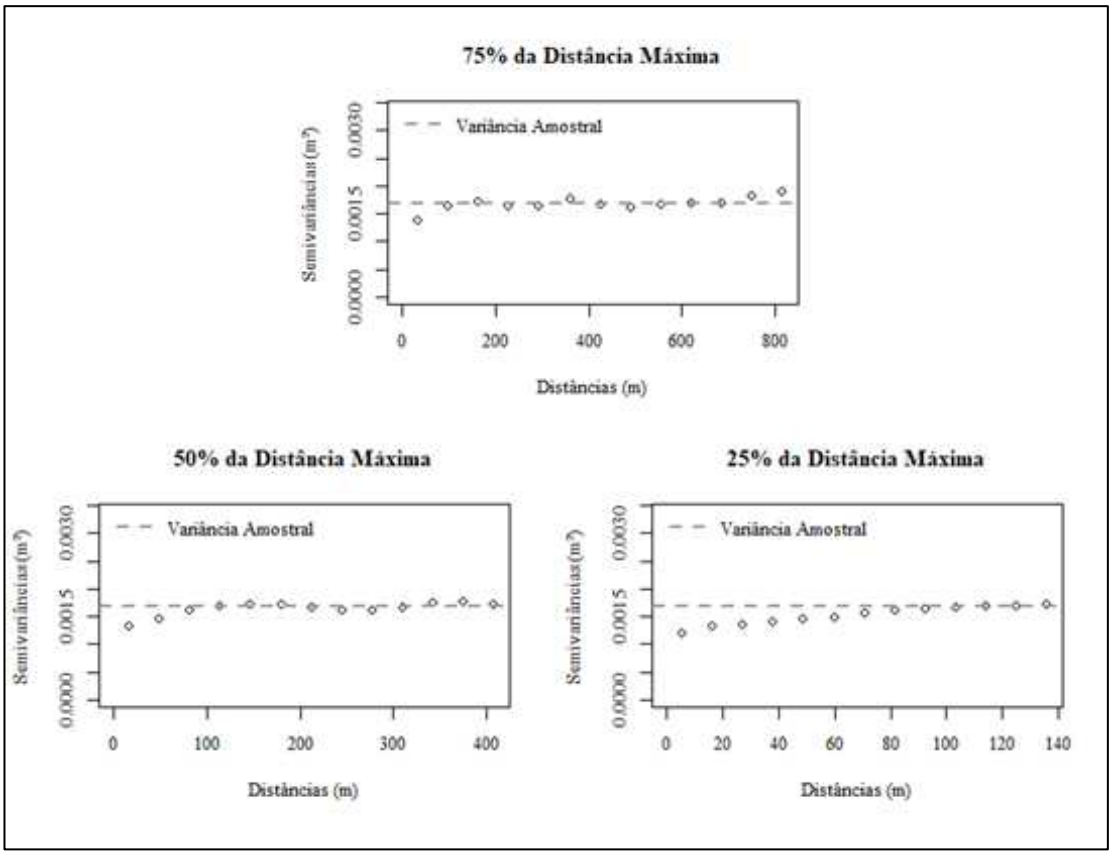

Figura 44 – Semivariograma das discrepâncias / amostra dp6.

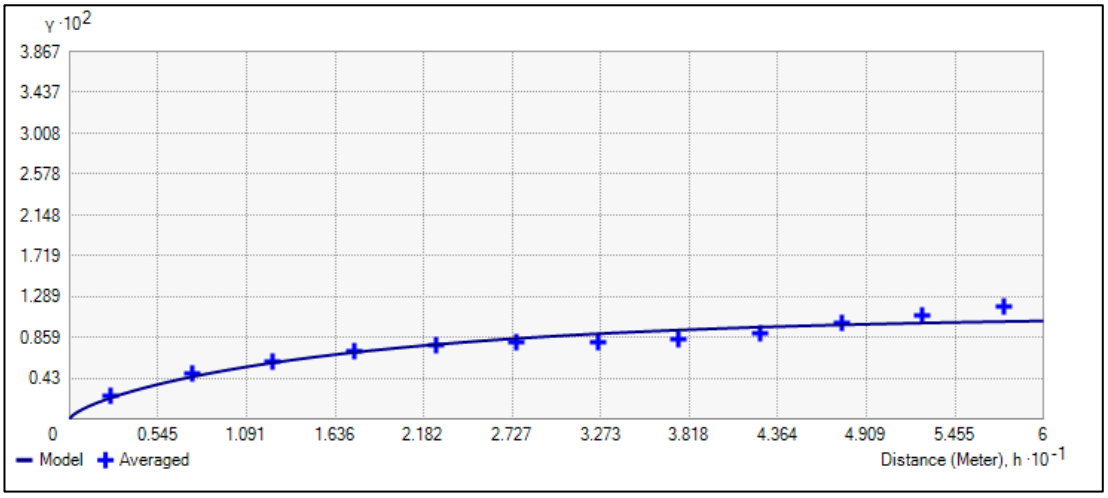

Figura 45 – Semivariograma das discrepâncias / amostra dp1_ss.

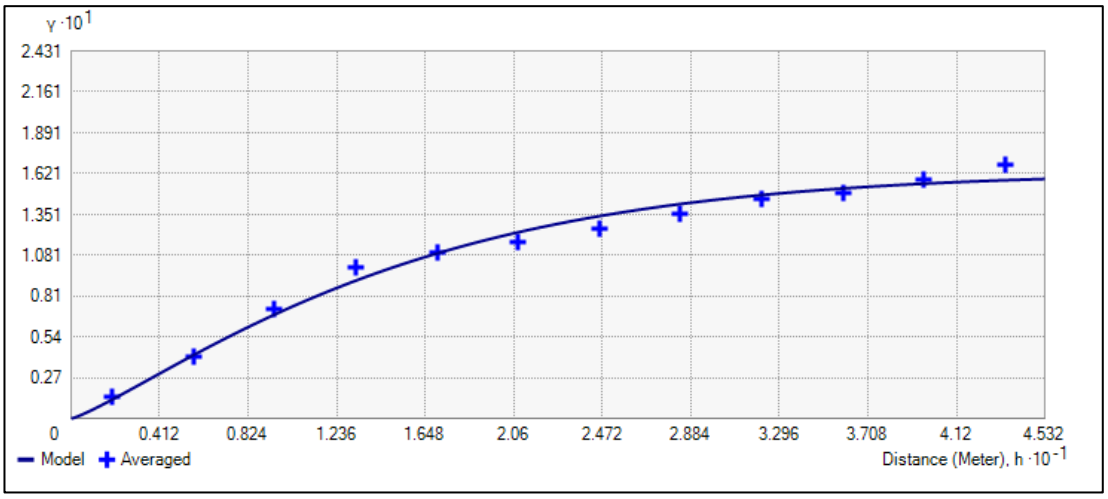

Figura 46 – Semivariograma das discrepâncias / amostra dp2_ss.

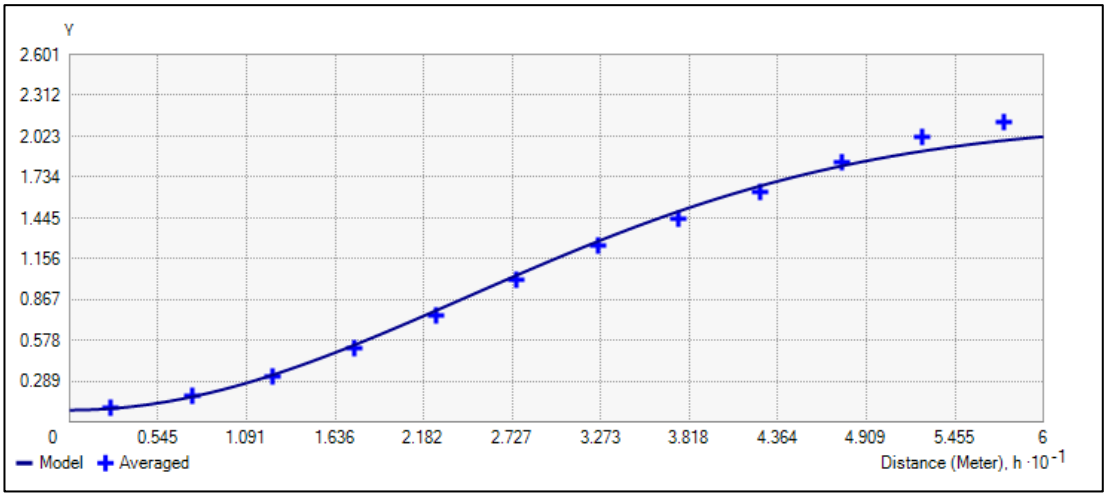

Figura 47 – Semivariograma das discrepâncias / amostra dp3_ss.

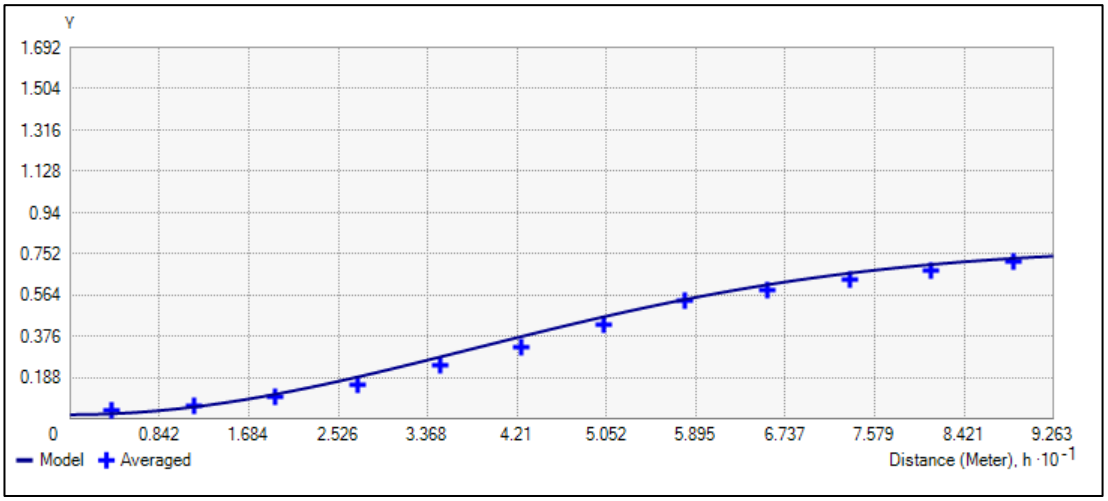

Figura 48 – Semivariograma das discrepâncias / amostra dp4_ss.

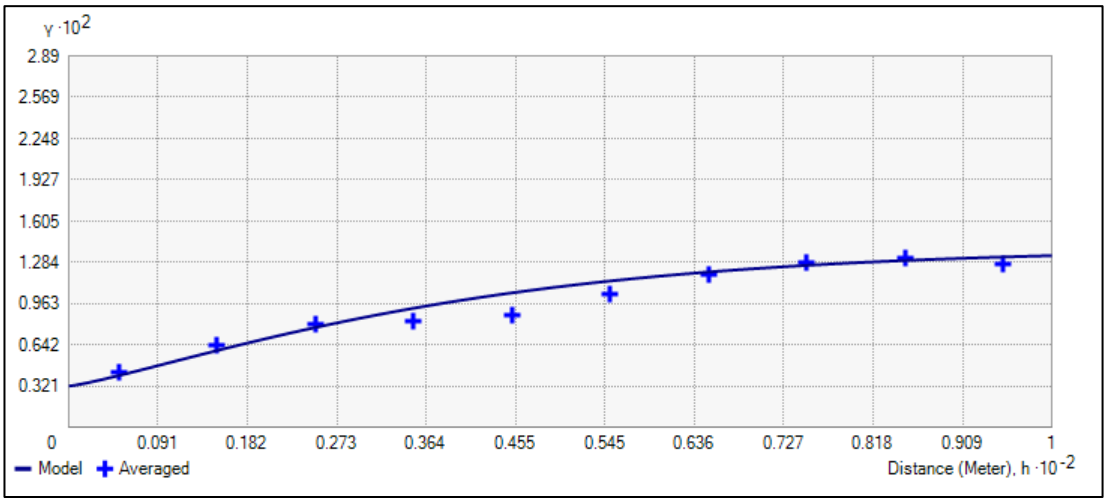

Figura 49 – Semivariograma das discrepâncias / amostra dp1_sp.

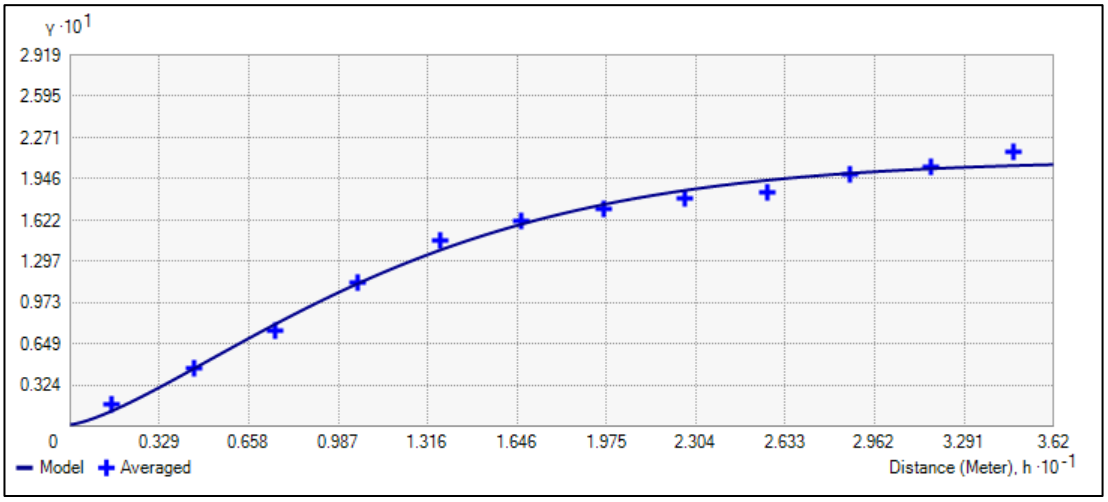

Figura 50 – Semivariograma das discrepâncias / amostra dp2_sp.

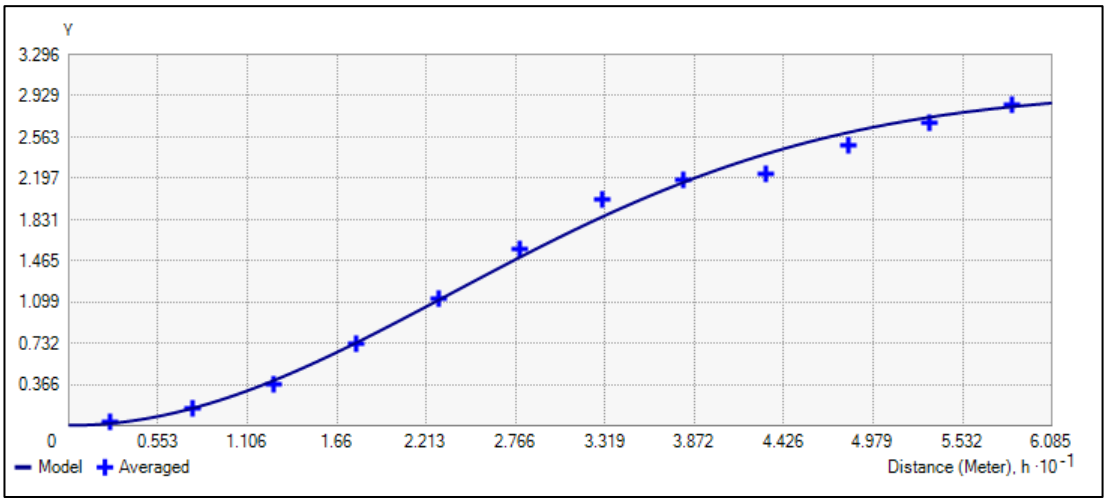

Figura 51 – Semivariograma das discrepâncias / amostra dp3_sp.

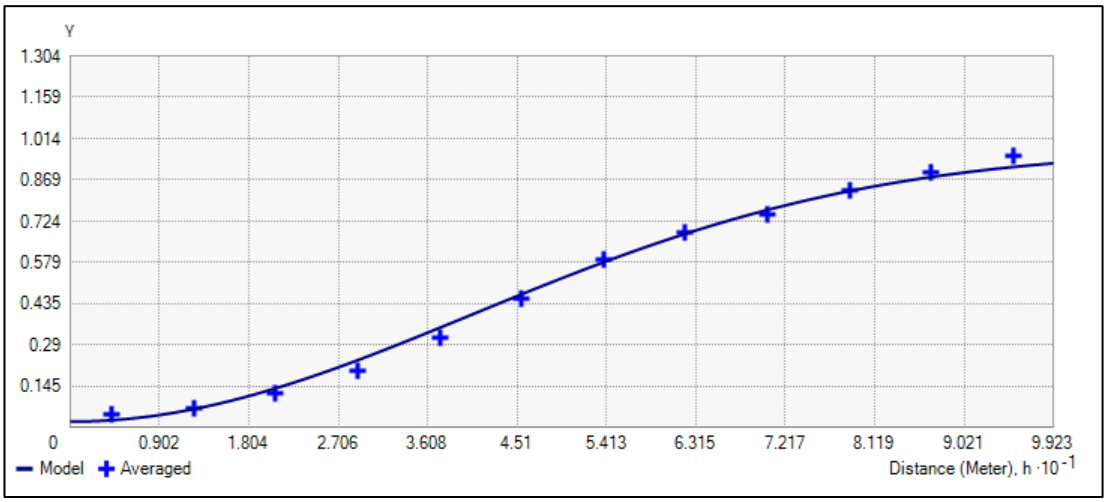

Figura 52 – Semivariograma das discrepâncias / amostra dp4_sp.