DOI 10.1590/s1982-21702019000400020

ORIGINAL ARTICLE

SPATIAL OUTLIERS DETECTION ALGORITHM (SODA) APPLIED TO MULTIBEAM BATHYMETRIC DATA PROCESSING

Italo Oliveira Ferreira[1] - ORCID: 0000-0002-4243-8225

Afonso de Paula dos Santos[1] - ORCID: 0000-0001-7248-4524

Júlio César de Oliveira[1] - ORCID: 0000-0003-0894-5597

Nilcilene das Graças Medeiros[1] - ORCID: 0000-0003-0839-3729

Paulo César Emiliano[2] - ORCID: 0000-0002-1314-9002

[1]Universidade Federal de Viçosa, Departamento de Engenharia Civil, Viçosa - MG, Brasil.
E-mail: italo.ferreira@ufv.br; afonso.santos@ufv.br; oliveirajc@ufv.br; Nilcilene.medeiros@ufv.br.

[2]Universidade Federal de Viçosa, Departamento de Estatística, Viçosa - MG, Brasil.
E-mail: paulo.emiliano@ufv.br.

Received in 27th July 2018
Accepted in 17th December 2018

Abstract:

The amount of data produced in an echo sounder has grown exponentially with scanning systems such as multibeam echo sounders and interferometric sonars, providing a considerable improvement in the submerged relief representation, especially to detect hazard objects to the navigator. However, the available processing algorithms did not follow this evolution; manual processing is usually necessary or at least constant intervention by an analyst, making this task arduous with a high subjectivity degree. In addition, statistical inconsistencies are common in most of the algorithms and filters available. Thus, SODA (Spatial Outliers Detection Algorithm) was recently presented, which is a methodology directed at first for echo sounder data treatment. The authors evaluated the algorithm efficiency using simulated data. Therefore, this article aimed to evaluate the SODA efficiency for real data treatment, with a multibeam echo sounder. A number of interesting results was obtained, reaffirming the methodology strength, regarding the search for spikes in echo sounder data.

Keywords: Spikes, Outliers, Multibeam Sonar, Multibeam Data Processing.

How to cite this article: FERREIRA, I. O.; SANTOS, A. P.; OLIVEIRA, J. C.; MEDEIROS, N. G.; EMILIANO, P. C. Spatial outliers detection algorithm (SODA) applied to multibeam bathymetric data processing. *Bulletin of Geodetic Sciences*. 25(4): e2019020, 2019.

1. Introduction

Hydrographic surveys deal with the configuration of deep ocean, rivers, lakes, dams, ports, among others. Generally, these surveys deal with the measurement of depths, tides, ocean currents, gravity, terrestrial magnetism and determination of chemical and physical properties of water, with the main goal of compiling data for generating charts and reliable nautical publications (Iho, 2005). Thus, performing these surveys requires specific physical environment knowledge, submarine acoustics, the numerous equipment and sensors, as well as the appropriate procedures to comply with the requirements and recommendations defined by national and international regulations (Iho, 2008; Jong et al., 2010; Dhn, 2014; Ferreira et al., 2015; Ferreira et al., 2016).

The main task of a hydrographic survey is collecting spatially georeferenced depths. This task occurs through so-called echo sounder surveys. In a current scenario, the products generated with swath system show high gain in resolution and accuracy, both in planimetric and altimetric terms (depth), and a large data densification, describing almost completely the submerged bottom (Iho, 2005; Usace, 2013; Maleika, 2015). This data density is even higher when sampling occurs in shallow water. An increasing number of hydrographic services have adopted multibeam technology as the main methodology for collecting echo sounder data for map production and updating (Iho, 2008; Instituto Hidrográfico, 2009; Linz, 2010; Noaa, 2011; Usace, 2013; Dhn, 2014).

Data collected by such technology, when used for making or updating nautical charts, as well as in related activities, should be free from abnormal data, that is, spikes and tops. However, bathymetric survey has numerous sources of uncertainty, as described, for example, in Iho (2008). In a depth measurement, spikes can depend, among other factors, on sonar quality and bottom detection algorithm (phase detection, amplitude, Fourier transform, etc.), secondary lobe detection, of multiple reflections, air bubbles of the transducer and reflections in the water column caused mainly by algae, fish shoals, deep scattering layer, thermal variations and suspended sediments (Urick, 1975; Jong et al., 2010)[1].

Manual processing of bathymetric data is still common today, especially in the phase of outliers elimination. However, as the increased density of collected data, this procedure became time consuming. Furthermore, when spike elimination occurs manually, that is, by a hydrographer, this stage shows a high degree of subjectivity, remaining dependent on the analyst, and therefore more sensitive to human errors. Less efficient systems are already capable of collecting thousands of points sets per hour of survey, which makes the data processing in its traditional form more costly than the hydrographic survey itself (Calder & Mayer, 2003; Calder & Smith, 2003; Bjørke & Nilsen, 2009; Vicente, 2011; Lu et al., 2010).

In response to the increase in the volume of data collected by the multibeam survey, researchers began to develop methodologies and algorithms of computer-aided processing in the 1990s in order to facilitate the hydrographer task (Vicente, 2011). According to Debese (2007), these algorithms estimate depth at a given location, and some are still able to evaluate qualitatively the estimation process. CUBE (Combined Uncertainty and Bathymetry Estimator) presented by Calder (2003) is one of the algorithms that stand out. This is one of the most promising algorithms for semi-automatic processing of multibeam data, which is implemented in numerous commercial processing packages.

Several authors have developed research in the detection area of anomalous values in bathymetry data from sonars. Ware et al. (1991), showed outliers detection process based on the analysis of the sample statistical properties. Following similar lines, Eeg (1995) proposed a statistical test to validate the size of the clusters that are used to detect spikes. Debese & Bisquay (1999), Motao et al. (1999), Debese (2007) and Debese et al. (2012) applied M-estimators. Calder & Mayer (2003) used the Kalman filter to process bathymetric data automatically. Similarly, Bottelier et al. (2005) used kriging techniques and Bjørke & Nilsen (2009) showed a technique for spike detecting based on the construction of trend surfaces. Lu et al. (2010) developed an algorithm based on the robust estimator LTS (Least Trimmed Squares).

[1]It consists of a layer of plankton that varies in depth throughout the day (IHO, 2005).

However, these methodologies are mostly difficult to apply, semi-automated or are implemented only in commercial packages. In addition, most of these techniques based on theoretical assumptions hardly met and/or verified, such as assuming that the variables are independent and belong to sets of variables normally distributed.

As an alternative to the available techniques, Ferreira et al. (2018) showed a new methodology for locating spikes called SODA (Spatial Outlier Detection Algorithm, in portuguese AEDO - Algoritmo Espacial de Detecção de Outliers). In the study, the authors described methodologically the developed algorithm and validated it through computational simulations. SODA has a series of robust and ideal characteristics, that is, it takes into account the autocorrelation and the non-normality of the data. In addition, it is a free computational code. However, tests with actual data had not been conducted.

Thus, this study aimed to evaluate the SODA efficiency from the real data processing. In this sense, we used bathymetric data collected through multibeam technology and preprocessed in agreement to specifications from national regulations. The depth and position data submitted to the SODA methodology were later compared to the traditional processing.

2. SODA – Spatial Outliers Detection Algorithm

SODA, presented by Ferreira (2018), is a methodology developed for spike detection in bathymetric point cloud, specifically multibeam echosounder data, interferometric sonars and airborne laser bathymetric systems. Although the focus of SODA is on bathymetric data, one can easily modify it to identify outliers in any set of georeferenced data, such as in Topography, Geodesy, Photogrammetry, Mining data, etc. (Ferreira, 2018).

The theoretical development of SODA bases on classical and geostatistical statistics theorems, which makes it supposedly robust and efficient. The operation principle consists in applying the steps described in the diagram below (Figure 1).

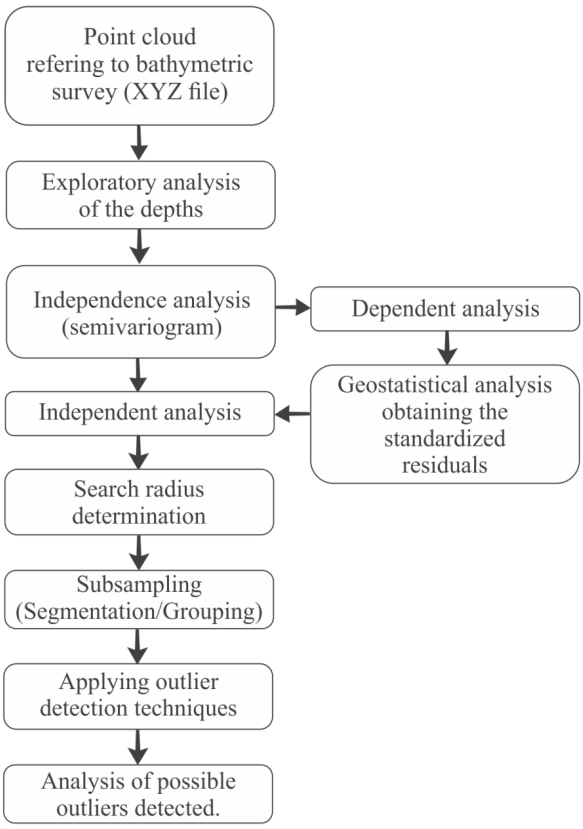

Figure 1: Flowchart of the logic process of SODA tool.

As can be seen in Figure 1, the first step of applying SODA consists in importing the points cloud. In this step, the three-dimensional coordinates in the XYZ format (Shapefile or text file) are imported, in which X and Y represent the planimetric coordinates that can be either local, projected or geodesic while Z denotes the reduced depth. An exploratory analysis is performed subsequently.

The next step, perhaps the most meaningful one to the entire process, is verifying the presence - or not - of spatial independence between the depth data, an assumed condition by the statistical supporting techniques used by SODA (Morettin & Bussab, 2004; Seo, 2006). For this evaluation, one used the semivariogram (MATHERON, 1965; FERREIRA et al., 2015).

Given the spatial nature of the bathymetry data, georeferenced depth sample is expected to be dependent. Regarding this, SODA uses Geostatistics tools to eliminate spatial autocorrelation and thus to continue with the procedures. In this step, a geostatistical modeling is carried out, aiming to generate the standardized residues (SRs), which are acknowledged as homogeneous, non-tendentious, independent and normal (Santos et al., 2017). By the complexity of a coherent geostatistical analysis, human intervention is required at this stage (Ferreira et al., 2013). Controversially, there may be cases where irregular sampling or even other factors lead to an independent sample, as proven by Ferreira (2018).

Once the spatial independence of the variable to be analyzed (depth or SR) has been verified, SODA application is continued. In the next step, the sample is spatially segmented in order to preserve the spatial characteristic of the analysis (local analysis). SODA employs **Segmentation in Circles**, in which the **Search Radius** is defined by the user or by a spatial analysis. In the latter case, a radius equal to three times the minimum distance between points is used.

From a given radius, the algorithm constructs a circle centered on each depth or SR, identifying and storing all the data inside this circle in subsamples. All analyses, from then onwards, are performed only on these subsamples. The independence characteristic of the data set is guaranteed for the subsamples by the following theorem: If $X_1, \cdots, X_k$ are independent random variables and $g_1(\cdot), \cdots, g_s(\cdot)$ are s functions such that $Y_j = g_j(X_j), j = 1, \cdots, k$ are random variables, then, $Y_1, \cdots, Y_s$ are independent. The demonstration of this theorem, as well as theoretic examples, can be found in Mood (1913) and Mood et al. (1974).

In a successive step, as discussed, local analyzes are performed in the sub-samples generated in the previous step. Thus, SODA employs three detection techniques of outliers: Adjusted Boxplot (Vandervieren & Hubert, 2004); Z-Score Modified (Iglewicz & Hoaglin, 1993) and δ -Method (Ferreira, 2018). The latter has been developed together with SODA and, as well as with other methods mentioned above, Ferreira (2018) prescribes it in detail.

The δ -Method, given by Equation 1, is a proposition consisting of a new outlier detection threshold, based on robust estimators with spatial characteristics.

$$Limiar = Q2_{Local} \pm c \cdot \delta \quad (1)$$

Where $Q2_{Local}$ is the median of the subsampled data as well as and are constants that depend on data variability. The constant assumes the value 1 for irregular terrains or artificial channels (high variability); 2 for slopes (average variability) and 3 for plain terrain (low variability). The value can be understood as a constant weight and must be inserted by the user. This constant is automatically determined by the algorithm through the evaluation of the Normalized Median Absolute Deviation Global ($NMAD_{Global}$) or Local ($NMAD_{Local}$). In other words, $\delta = 0{,}5 \cdot (NMAD_{Global} + NMAD_{local})$, if $NMAD_{Global} > NMAD_{local}$. Otherwise, $\delta = NMAD_{Global}$. When the detection thresholds for outliers are applied, the last SODA stage is performed. In this step, the probability of the data being outliers is determined ($P_{outlier}$) in each of the three techniques, based on the number of times the data were analyzed ($N^{o}_{analyzed}$) and the number of times they were considered outliers ($N^{o}_{outlier}$) (Equation 2):

$$P_{outlier}(\%) = \frac{N^{o}_{outlier}}{N^{o}_{analyzed}} \cdot 100 \quad (2)$$

For instance, considering that given any radius of search, the observation was sub-sampled 40 times. Hence, it was analyzed by the three detection techniques of outliers in these 40 times. In addition, one should note that out of the 40 times, the observation *i* was considered a spike by the δ -Method 10 times; hence, $P_{outlier} = (10/40) = 0{,}25$, or observation *i* is 25% likely to be a spike if the considered cutoff limit is the one given by the δ -Method. It is evident that SODA does this computation jointly and automatically alongside with the other thresholds.

Finally, once the user sets a standard P_{Limiar}, SODA spatially plots all observations, highlighting the detected spikes by the thresholds in use, i.e., all data where $P_{outlier} \geq P_{limiar}$. Afterwards, the hydrographer performs an inspection to confirm whereas the data flagged as spikes are in fact spikes and then deletes them. In all instances, new XYZ files are created for each technique previously discussed. To be more specific, SODA is associated to the *Adjusted Boxplot, Modified Z-Score* and δ -Method, respectively.

From the study presented by Ferreira (2018) SODA presents very interesting characteristics. When it comes to controlled environments, SODA has shown efficiency in detecting depths characterized as *spikes.* As for SODA's behavior in the presence of submerged structures – sandbars, rocks and shipwrecks-, there has been another promising outcome, comparing to other simulated results: SODA did not erroneously identify any of these features

3. Experiments and results

Real data that served as a basis for practical SODA evaluation were obtained from a partnership with the company A2 Marine Solution. The hydrographic survey occurred in April 2017, near the evolution basin of the Port Integrator Terminal Luiz Antônio Mesquita (TIPLAM). Evolutionary basin consists of the bordering area on berthing facilities, reserved for the necessary evolutions for the berthing and unberthing operations.

TIPLAM is located in the city of Santos, state of São Paulo, and currently moves about 2.5 million tonnes per year, being responsible for the discharge mainly of sulfur, phosphate rock, fertilizers and ammonia.

For bathymetric data collection, a swath system was used, composed by the multibeam echo sounders model Sonic 2022, by R2 Sonic, integrated inertial navigation system, model I2NS , by Applanix. The planning, execution and data analysis followed the recommendations of the NORMAM-25 (DHN, 2014) and S-44 (IHO, 2008) for category A and Special Order, respectively.

The surveys were firstly pre-processed in software Hysweep (Hypack, 2012), which consisted of the following steps:

- Conversion of data collected by the various sensors to the Hysweep format;
- Analysis of auxiliary sensor data (attitude, latency, sound speed, tide, etc.), aiming identifying possible faults. If necessary, interpolation or rejection of anomalous data;
- Junction of datagrams[2] ;
- Calculation of Total Propagated Uncertainty (Horizontal and Vertical);
- Calculation of three-dimensional coordinates in XYZ format (reduced georeferenced depths), and
- Duplicate data set filtering.

In this last step, were removed all duplicate data sets, mainly from overlapping data sets. This reduction of original sample allowed a considerable reduction of processing time. Subsequently, the study area was selected and the three-dimensional coordinates in the XYZ format (reduced georeferenced depths) were exported.

Figure 2 illustrates the study area.

[2]Complete and independent data entity. In this case, it refers to the data generated by the various sensors.

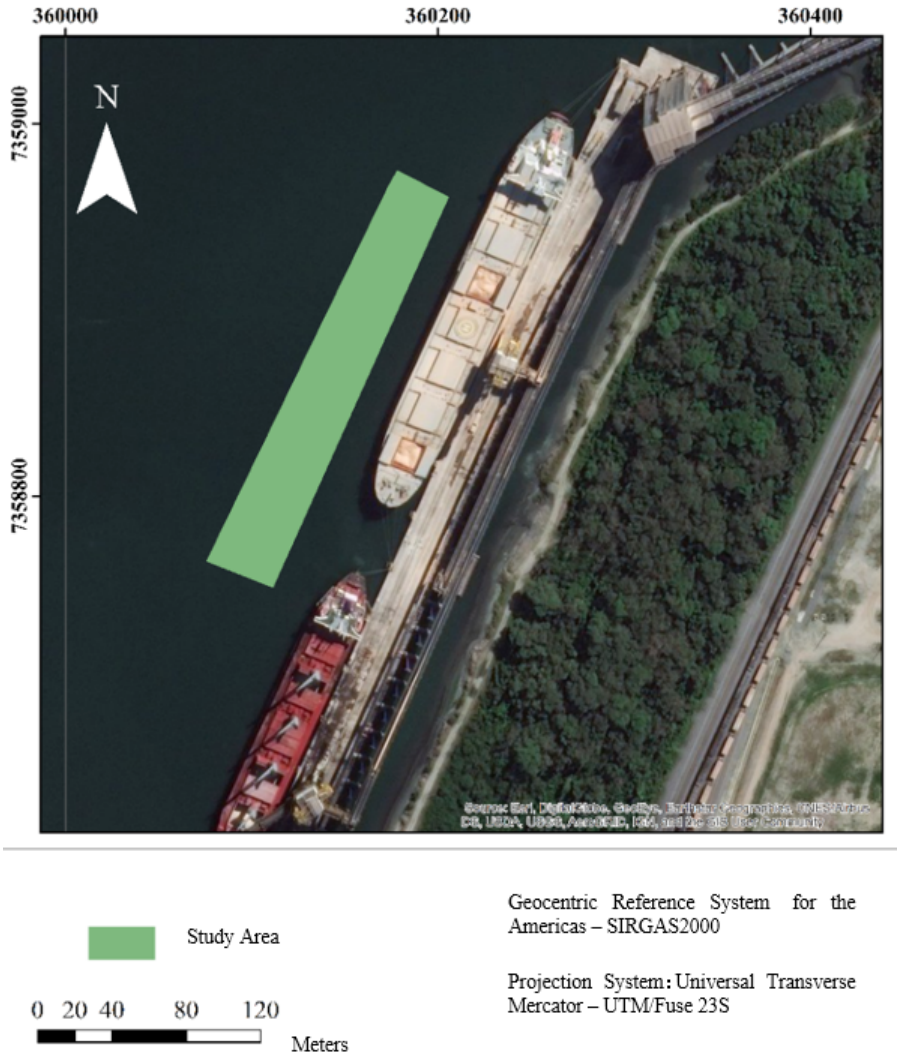

Figure 2: Study area.

With the spatial data set with their respective XYZ coordinates, SODA application may continue. Besides, the points were imported and submitted to exploratory analysis shown in Table 1.

Table 1: Descriptive statistics of the study area.

Number of Observations	8090
Average Depth (m)	14.854
Minimum Depth (m)	9.220
Maximum Depth (m)	15.690
Variance (m²)	0.1240
Kurtosis Coefficient	101.660
Asymmetry Coefficient	-6.900

From the variance (Table 1), it can be concluded that the data show low variability (WARRICK & NIELSEN, 1980). MoreFrom the variance value (Table 1), data show low variability (WARRICK & NIELSEN, 1980). Coefficients of asymmetry and kurtosis that quantify, respectively, the deviation of depth distribution in relation to a symmetrical

distribution and flatness degree of distribution, indicate an symmetrical distribution to the left (negative), leptokurtic and, in the first place, with a large concentration of values around the average. In conclusion, the distribution is non-normal and/or is affected by abnormal values.

Figure 3 shows some graphs that aid in the exploratory analysis and, as such, are constructed and generated by SODA.

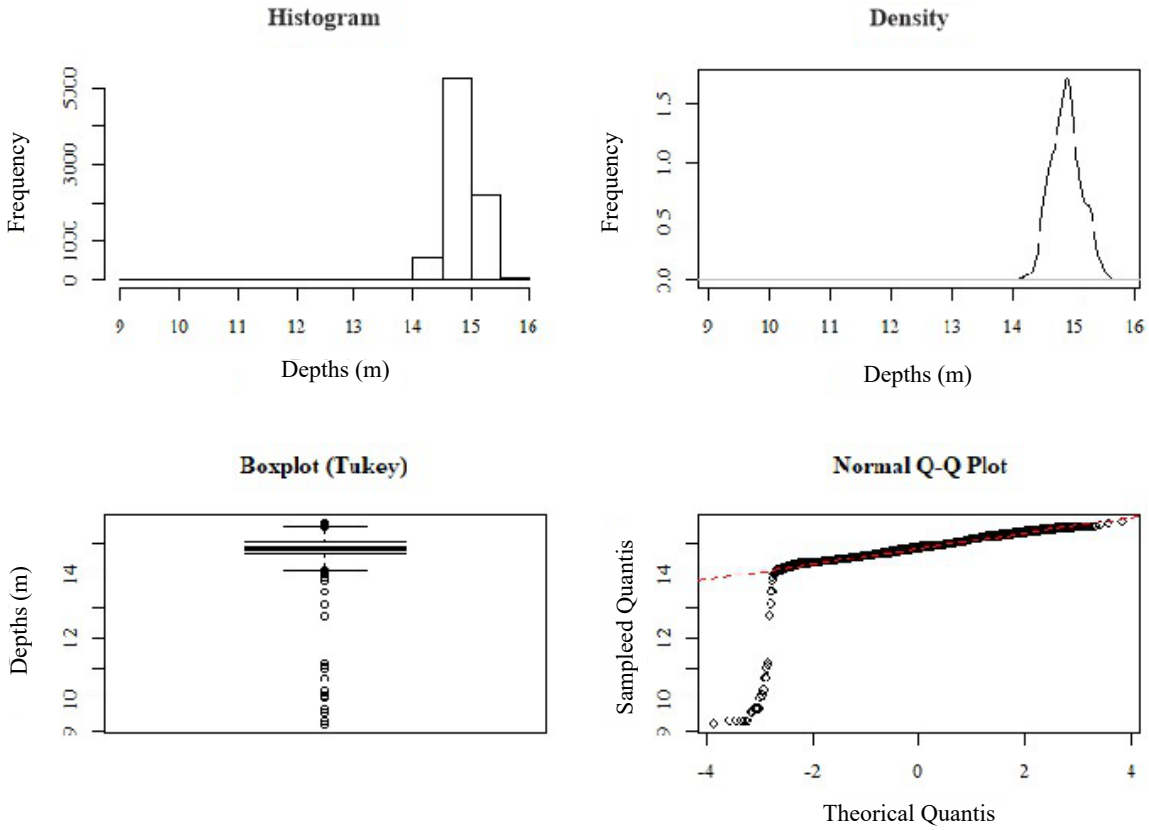

Figure 3: Exploratory graphical analysis.

The depth histogram together with the density curve graph are tools capable of providing indicative on data normality, which may or may not be proved by *Q-Q* Plot analysis. This consists of a graph that allows checking the adequacy of frequency data distribution (empirical/real) to any probability distribution, briefly, the quantiles of the empirical distribution function are plotted against the theoretical quantile of probability distribution, in this case, the normal distribution. If the empirical distribution is normal, the graph will be presented as a straight line (Höhle & Höhle, 2009). After the graphical analysis, we observed the non-normality of data, which could later be confirmed by univariate normality tests if spatial independence occurs.

Finally, we present the Tukey boxplot, in which we can see, in general, the presence of possible *outliers.* However, such an assertion cannot be confirmed since the *Tukey* method does not consider the spatial dependence structure of the depths. In addition, the cut-off thresholds are derived from the normal distribution. However, observing the histogram, there are some depths slightly apart from the average.

Figure 4 shows point cloud. In dark blue color, there is the presence of some depths disagreeing with its neighboring values. Some of the possible *spikes* coincide with those abnormal depths detected by the Tukey method.

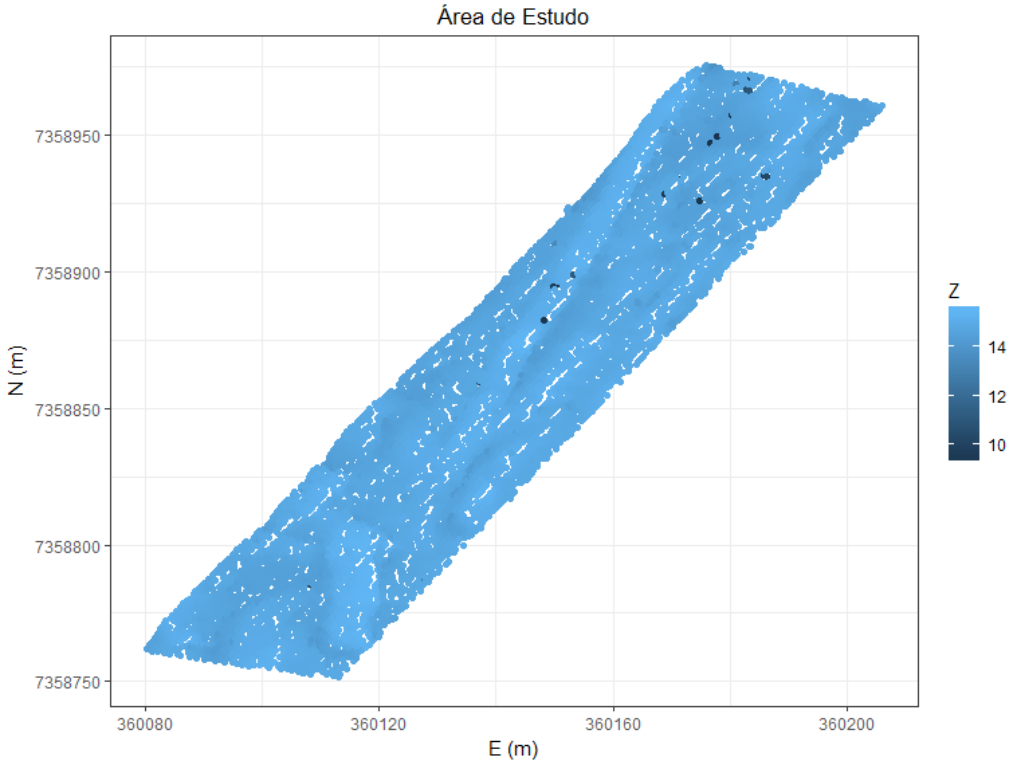

Figure 4: Spatial data sets of the study area.

According to the flow diagram shown in Figure 1, the spatial independence analysis is performed after the exploratory analysis. If the spatial independence is confirmed, the proposed methodology may continue, determining the search radius followed by segmenting the sample. Otherwise, a geostatistical analysis, with the objective to obtaining the standardized residues (SRs), must be performed. Then, all subsequent analyses are performed on the SRs. SODA builds three semivariograms that allow, through a visual analysis, to confirm or not, the spatial independence.

Figure 5 shows the semivariograms, where the spatial independence can be confirmed by the presence of pure nugget effect. This is a rare case, since an intrinsic characteristic of georeferenced data is its spatial autocorrelation. For this reason, it was necessary to redo this process even though the obtaining results have proven to be identical.

Determining the search radius is preferably carried out from a spatial analysis. In this case, a radius equals to three times the minimum distance is selected. In this study, one obtained a radius of 1.799 meters. From this radius, the algorithm segments and applies for each sub-sample the three proposed detection thresholds of *outliers*. Firstly, the *Adjusted Boxplot, Modified Z-Score* and δ *-Method* thresholds detected, respectively, 2 028 (25.07%), 1 204 (14.88%) and 161 (1.99%) possible spikes. As it is a flat area, the constant plain terrain of δ *-Method* was set to 3.

A comparison between the thresholds used by SODA in the study area shows a 95.65% agreement between the δ *-Method* and the *Modified Z-Score* threshold, i.e., out of the 161 possible *spikes* detected by the δ *-Method,* 154 were also detected by the *Modified Z-Score*. Similarly, the agreement between the also and the *Adjusted Boxplot* was 57.76% and between the *Modified Z-Score* and the *Adjusted Boxplot* of 43.35%.

In the next step, SODA refines the analysis by calculating the likelihood that the data is a spike for each of the three techniques. To execute this step, a has been used for the Adjusted Boxplot and δ *-Method* and $P_{threshold} = 0{,}5$ for the *Modified Z-Score*, as suggested by Ferreira (2018). After that, the proposed methodology associated with the *Adjusted Boxplot, Modified Z-Score* and δ *-Method* detected 66 (0.82%), 24 (0.30%) and 34 (0.42%) spikes, respectively. Out of the 24 *spikes* detected by the *Modified Z-Score* technique, 16 were also detected by the *Adjusted Boxplot* and by the δ *-Method*, showing an agreement of approximately 66%. The agreement between the δ *-Method* and the *Adjusted Boxplot* was nearly 38.23%, suggesting that the 16 outliers located in agreement with the *Modified Z-Score* are not the same.

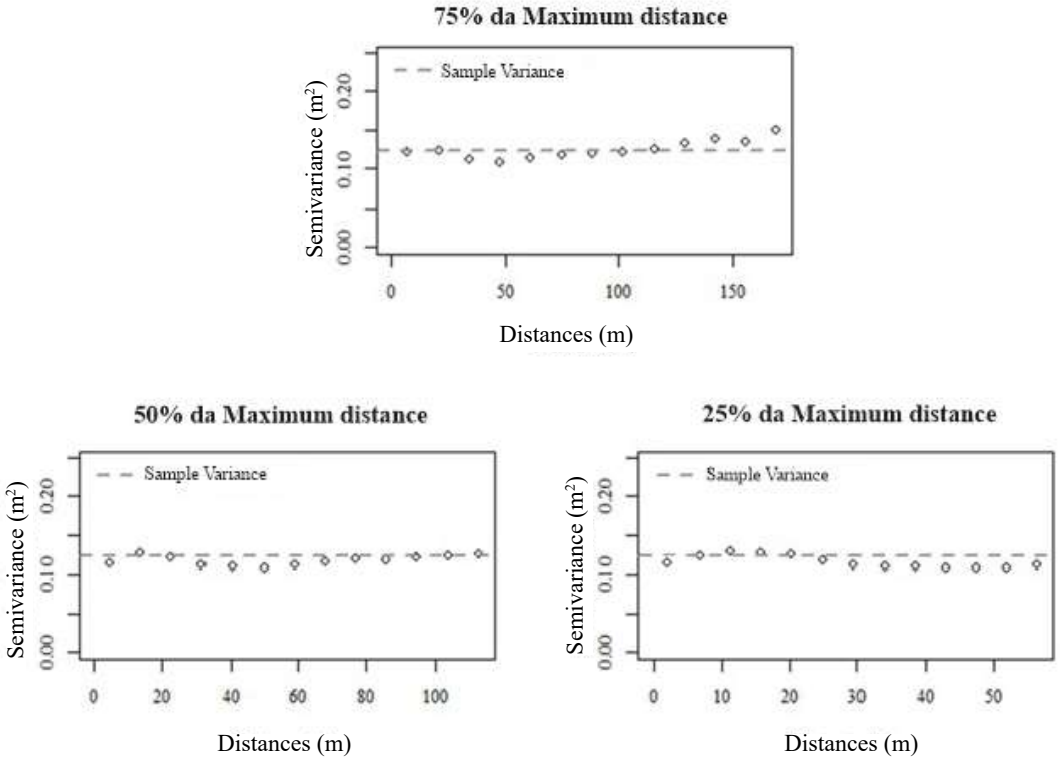

Figure 5: Experimental semivariograms with 75%, 50% and 25% maximum distance.

Figure 6 shows the study area, highlighted in red for the detected *spikes.*

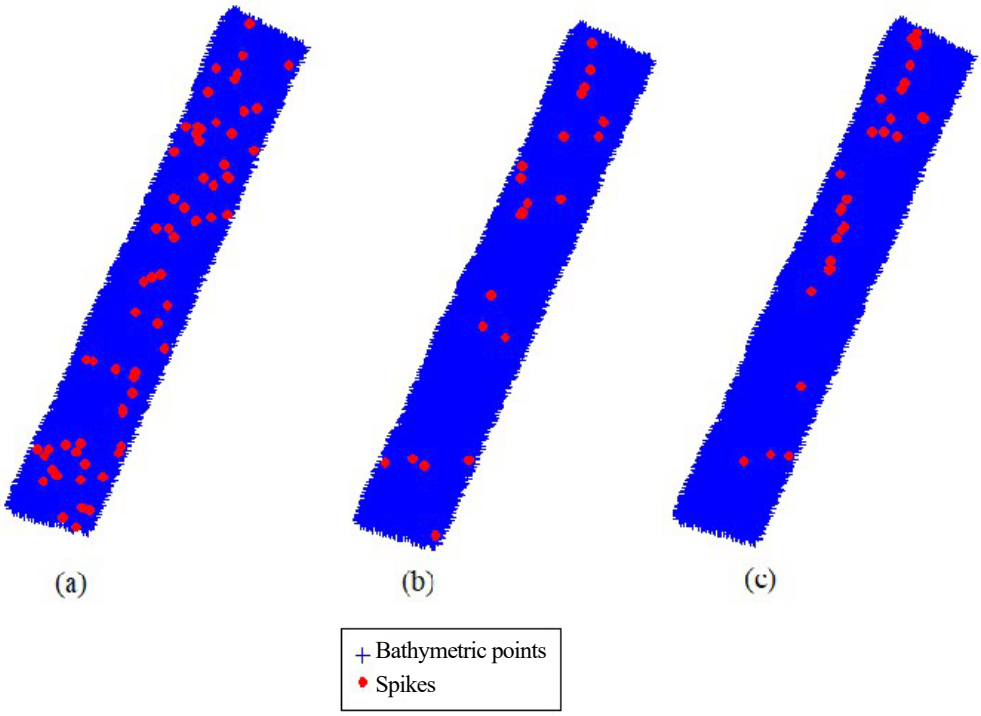

Figure 6: Spikes detected by SODA from the *Adjusted Boxplot* (a), *Modified Z-Score* (b) and δ *-Method* (c) thresholds.

It is noteworthy that for processing this data set, one used a machine with an operating system Windows 10, RAM of 8GB RAM (partially dedicated to software R (R Core Team, 2017)) and Intel® Core™ i7-4500U CPU @

1.80GHz 2.40 GHz processor. The processing time was approximately 9 minutes.

In order to evaluate thoroughly the robustness of the proposed methodology, and particularly of the *δ-Method*, the study area was submitted to manual processing. In this phase, the professional visualizes the data through a graphical interface, in which can be presented, among other information, the general and partial spatial data set, bathymetric profile and backscattering images, and decides, based on a qualitative and analytical judgment, what depth can be valid.

In general, the study area is relatively small and has a flat relief, which facilitates manual spike research. In addition, the surveys were collected by a high-quality multibeam system, which provided reasonably simple manual processing. However, collecting depths using low-quality multibeam echo sounders, interferometric sonars, or bathymetric LiDAR (Light Detection and Ranging) systems, usually produce very noisy data, making manual quality processing very slow. Obviously, in the shallow water or in large areas, the problem around manual processing is aggravated.

Through manual processing, 38 spurious depths were detected, approximately 0.47% spikes, as shown in Figure 7.

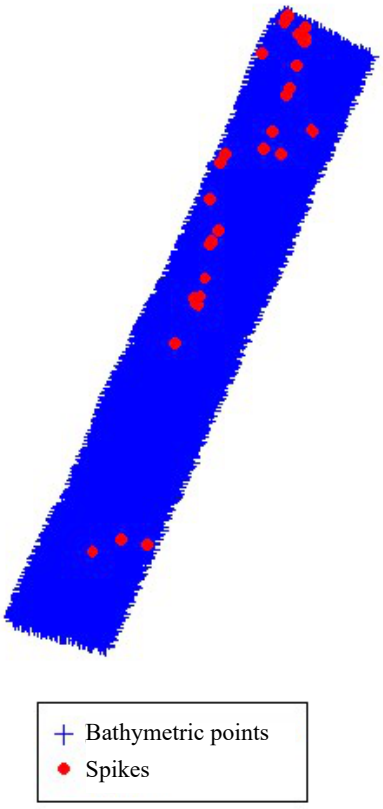

Figure 7: *Spikes* detected through manual processing.

Taking the manual processing as a reference, the proposed methodology, allied mainly to the *δ-Method* thresholds, was quite robust, obtaining an agreement of 82.35%, that is, of the 34 spikes located by the *δ-Method*, 28 were identified by manual processing. In other words, the hit percentage was 74%, since the *δ-Method* detected 28 of the 38 spikes identified in the data set.

The undetected data sets, almost entirely, are outliers with very low magnitudes, around 0.30 meters. Therefore, the failure to detect these data sets probably does not cause losses or gains to the final products.

This problem is due to two main reasons. The first relates to the border data sets. Figure 8 shows that SODA failed to detect a few *spikes* present at the extremities of the study area. This is due to the developed methodology, since spike searches occur on the subsamples formed from circles and, thus, the analyzes at the edges of the spatial data sets are

impaired by the reduced number of verifications (redundancy). In addition to that, there is the need of at least 7 points within each circle in order to apply outlier detection techniques (Ferreira, 2018).

The second reason is associated with spike agglomeration, which may mask the local analysis, and/or their presence in coverage failures. In these cases, analyzes redundancy also reduces and, consequently, there is a decrease in process reliability. In more extreme cases, spikes present in the coverage failures or in areas with low data set density may not be analyzed, since the sub-samples tend to show less than 7 depths. Thus, SODA is effective when the spatial data set is dense and has no coverage flaws.

Particularly analyzing the *δ-Method*, apparently, also showed problems in signaling 6 valid depths as spikes. When investigating these depths, some of them are present in areas with coverage failures and, as discussed, in these occasions the methodology may show flaws. The others (ID 6026 and ID 2184), when analyzed locally, showed distant depths from the neighbors, with a magnitude around 50 centimeters, indicating, therefore, failures in manual processing.

Other, as Table 2 shows, these depths obtained a $P_{outlier}$ of about 50%. This means that if $P_{outlier}$ was set by the analyst with a value of 51%, that is, $P_{outlier}$=51%, the method would find only 2 false spikes (ID 6026 and ID 2184).

Table 2: Valid depths marked as *spikes* by the δ *-Method*.

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

Adjusted Boxplot quantitatively overestimated spike detection, localizing depths that, compared to manual processing, are not characterized as spikes. Sixty-six probable spikes were detected, of which only 17 were identified through manual processing. Therefore, the agreement and percentage of accuracy of this threshold were, respectively, 25.76% and 44.74%.

The incorrect data set detected, when analyzed locally through partial gross bathymetric profiles, are depths that somewhat deviate slightly from the neighbors, showing themselves as local noises. However, in processing for navigational purposes, most points sets detected by this threshold are, in fact, valid depths. No probable spike detected by the threshold represents a direct danger to navigation and, thus, the exclusion of these data sets does not cause losses to the final generated products. This fact may represent gains, since the proposed method proved to be an excellent tool for smoothing bathymetric surfaces. Therefore, in cases in which the submerged relief softening is required, for example, when it is desired to construct contours or digital depth models, the application of this threshold can be very useful.

On the other hand, the *Adjusted Boxplot* threshold, even finding 66 probable outliers out of 38 possible, obtained only 44.74% accuracy. Analyzing the undetected *spikes,* the issue discussed previously occurred, that is, the methodology has low efficiency to locate *spikes* present at the extremities of the study area, at holidays and in areas with low points set density. Another important point was the failure to detect spurious depths in cases of spike agglomerations, such as that present in the central part of the study area (Figure 8).

Finally, the *Modified Z-Score* threshold underestimated spike detection. Twenty-four *spikes* were found, of which 16 are anomalous depths, that is, an agreement of 66.67% and a hit percentage of 42.15%. In general terms, this threshold was more efficient than the *Adjusted Boxplot.* Analyzing the undetected *spikes*, in accordance with

what happened previously, the *Modified Z-Score* showed the same problem, that is, agglomerate *spikes* in the extremities of the study area and in coverage failures were not detected. On the other hand, 8 valid depths were mistakenly marked as *spikes*, of which 6 were also detected by the *Adjusted Boxplot* and the other two (ID 6377 and ID 7801) show a difference in relation to the vicinity of about 20 centimeters.

Figure 8 shows *spikes* detected by manual processing and those located by the proposed methodology associated with outliers detection techniques.

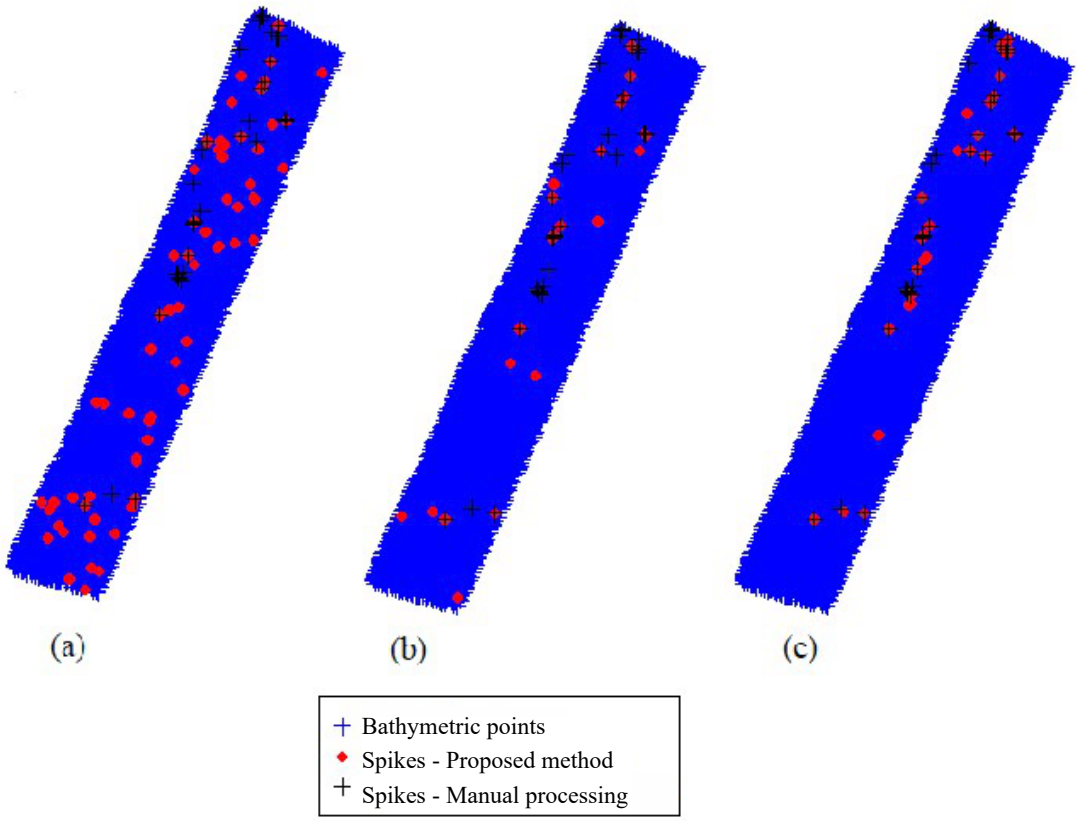

Figure 8: Excluded *Spikes* through manual processing and *Spikes* detected by SODA from the thresholds of the *Adjusted Boxplot* (a), *Modified Z-Score* (b) and δ *-Method* (c).

Given the above, the proposed methodology associated with *δ-Method* proved to be quite robust, obtaining about 74% accuracy in the real data processing, while the *Adjusted Boxplot* and *Z-Score Modified* thresholds reached 44.73% and 42.10%, respectively. Generally, both thresholds, when compared to manual processing, have proved to be powerful tools for locating and eliminating spikes. Some settings, such as defining the constant c for the *δ- Method*, are still required.

4. Final Remarks

This work aimed to evaluate the SODA methodology in the real bathymetric data processing. In the evaluation, we used bathymetric surveys lines with multibeam technology. Data were collected in a moderately small area with submerged bottom, which made manual processing reasonably easy. Data processed manually were used as reference for SODA validation.

The proposed strategy for spikes detection has proved to be a versatile alternative of great potential when compared to the methodologies currently used, especially for manual treatment.

The subjectivity degree of a process of manually identifying spikes reduced comparatively to the proposed method, since this is a semi-automatic approach, and the user should only indicate some parameters in the process, such as determining the constant c (*δ-Method*). In some cases, if the analyst prefers, he/she may choose to set the Search Radius and/or modify the P_{limiar}, instead of performing analyzes based on each personal experience. Thus, in addition to being a less subjective method, there is no need to analyze large data sets in the proposed approach, making the process less time consuming.

Even in the case of existing semiautomatic methods, which may be difficult to apply, these techniques are almost all based on theoretical assumptions difficult to answer and/or verified, such as assuming that the studied variables are independent and belong to normally distributed sets of variables, unlike the proposed tool that performs analysis of normality and spatial independence. That is, besides to minimize the analyst's interference, the use of theorems of classical statistics and Geostatistics theoretically strengthened SODA.

Among the thresholds used by SODA, the *δ-Method*, performed better when compared to the others. *Adjusted Boxplot* threshold overestimated the amount of *spikes* in the data set. However, SODA, associated with the *Adjusted Boxplot* technique, can be a useful tool for smoothing echo sounder models, with subsequent generation of contour lines. Unlike the results presented by Ferreira (2018), for real data, the Z-Score threshold underestimated the amount of spikes in the data set.

Although the focus has been on the use of multibeam survey data, the proposed methodology can be applied to airborne laser bathymetry systems, interferometric sonars and even related areas such as topography, geodesy, photogrammetry, mining, statistics, etc. Future tests and simulations are recommended in different types of relief, aiming to improve the performance of algorithms in terms of processing time, definition of the search radius for areas with coverage faults and definition of the constant of *Method δ*. Regarding the search radius determination, it is also suggest that different search radiuses be applied simultaneously during processing. This will allow for greater redundancy of analyzes as well as a possible improvement in the quality of processing results on spatial data set with coverage failures. Another problem to be solved is the question of analysis at the spatial data set extremities. However, computationally, the solution to this fact is still not feasible. A practical solution would always be to extrapolate the survey area during field surveys. Finally, the application of this methodology in related areas is desirable.

ACKNOWLEDGMENT

The authors would like to thank GEPLH (Study and Research Group on Hydrographic Surveys) and DEC (Department of Civil Engineering) for providing the realization of this survey.

AUTHOR´S CONTRIBUTION

The authors contribute equally.

References

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. & 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.

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.

FERREIRA, I. O. *Controle de qualidade em levantamentos hidrográficos*. Tese (Doutorado). Programa de Pós-Graduação em Engenharia Civil, Departamento de Engenharia Civil, Universidade Federal de Viçosa, Viçosa, Minas Gerais, 216p., 2018.

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.; 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.

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.

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, 24p., 2009.

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.

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.

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.

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.

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.

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.

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.

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.