DOI 10.1590/s1982-21702019000300014

ORIGINAL ARTICLE

ROBUST METHODOLOGY FOR DETECTION OF SPIKES IN MULTIBEAM ECHO SOUNDER DATA

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, Setor de Engenharia de Agrimensura e Cartográfica, 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 04th July 2018

Accepted in 18th April 2019

Abstract:

Currently, during the operation in shallow waters, scanning systems, such as multibeam systems, are capable of collecting thousands of points in a short time, promoting a greater coverage of the submerged bottom, with consequent increase in the detection capacity of objects. Although there has been an improvement in the accuracy of the depths collected, traditional processing, that is, manual, is still required. However, mainly due to the increased mass of data collected, manual processing has become extremely time-consuming and subjective, especially in the detection and elimination of spikes. Several algorithms are used to perform this task, but most of them are based on statistical assumptions hardly met and/or verified, such as spatial independence and normality. In this sense, the goal of this study is to present the SODA (Spatial Outlier Detection Algorithm) methodology, a new method for detection of spikes designed to treat bathymetric data collected through swath bathymetry systems. From computational simulation, promising results were obtained. SODA, in some cases, was capable to identify even 90% of spikes inserted on simulation, showing that the methodology is efficient and substantial to the bathymetric data treatment.

Keywords: Spikes; Outliers; Multibeam Echo Sounder; Multibeam Data Processing.

How to cite this article: FERREIRA, I. O., SANTOS, A. P., OLIVEIRA, J. C. MEDEIROS, N. G. and EMILIANO, P. C. Robust Methodology for Detection of Spikes in Multibeam Echo Sounder Data. *Bulletin of Geodetic Sciences*, 25(3): e2019014, 2019.

1. Introduction

The collection of depths is an essential task in several areas, especially those related to the production and updating of nautical cartography. Unlike electromagnetic waves, the acoustic waves present a good propagation in the aquatic environments and, for this reason, most of the sensors used in the depth determination use sound waves, such as: single-beam, multibeam echo sounders and interferometric SONAR (Sound Navigation and Ranging) (IHO 2005; Ferreira, et al. 2017a).

Despite the attenuation of the electromagnetic waves, LASER (Light Amplification by stimulated emission of radiation) probing systems have also been used in the bathymetric mapping, emphasizing, mainly, the great gain of productivity (Guenther et al. 1996; IHO 2005; Pastol 2011; Ellmer et al. 2014). The use of orbital images to estimate shallow water bathymetry has also been the subject of research (Gao 2009; Cheng et al. 2015; Moura et al. 2016; Ferreira et al. 2016a, 2016b).

However, in a current scenario, hydrographic surveys, especially those intended for cartographic updating, are restricted to the use of multibeam echo sounders and interferometric SONAR. In comparison with single beam sounders, these systems have a high gain in resolution and accuracy, both in planimetric and altimetric (depth) terms, and a large data densification, describing almost completely the submerged bottom, and improving the ability to detect objects (Cruz et al.2014; Maleika 2015). Less efficient multibeam systems are already capable of collecting more than 30 million points per hour in shallow water (Bjørke and Nilsen 2009).

While single beam bathymetry systems perform a single depth recording at each transmitted acoustic pulse (ping), resulting in a line of points immediately below the vessel's trajectory, the scanning system performs several depth measurements with the same ping, obtaining measurements of the water column in a swath perpendicular to the trajectory of the vessel. A growing number of hydrographic services have adopted multibeam technology as the main methodology for collecting bathymetric data for cartographic production (IHO 2008; Instituto Hidrográfico 2009; LINZ 2010; NOAA 2011; USACE 2013; DHN 2014). Interferometric Sonar systems are a relatively new technology, but they are likely to achieve results similar or superior to those of multibeam bathymetry, with advantages mainly in covering shallow water bottom (Cruz et al. 2014). However, this technology still lacks more detailed studies for validation in terms of building and updating charts and nautical publications.

Although beam echo sounding systems are the most widely used and bring improved resolution and accuracy of bathymetry, traditional data processing has been more time-consuming than the survey itself. Among the several phases of this process, stand out the detection, analysis and elimination of discrepant data (spikes) (Ware et al. 1992; Artilheiro 1998; Calder and Mayer 2003; Calder and Smith 2003; Bjørke and Nilsen 2009; Vicente 2011).

These discrepancies can be considered as outliers and thus undesirable in the set of data to be processed. The term outlier can be defined as an observation that, statistically, differs from the data set to which it belongs, that is, it is an atypical or inconsistent value (Mood et al. 1974; Santos et al. 2017). In this sense, outliers can be caused by gross errors, by systematic effects or, simply, by random effects, according for example, Santos et al. (2016). In hydrographic surveys, depths that are configured as outliers are known as spikes, while positioning errors are called tops. This work focuses specifically on the vertical component, and for this reason the term spike is

sometimes treated as synonymous with outlier. Figure 1 illustrates a bathymetric profile in the presence of spikes.

In the bathymetric survey, the anomalous values are mainly caused by the poor performance of the algorithms used by the echo sounder for bottom detection (detection by phase, amplitude, Fourier transform, etc.), detection by side lobes, multiple reflections, presence of bubbles of air in the face of the set of transducers, by reflections in the water column and, even, by equipment operating simultaneously in the same frequency (Urick 1975; Jong et al. 2010).

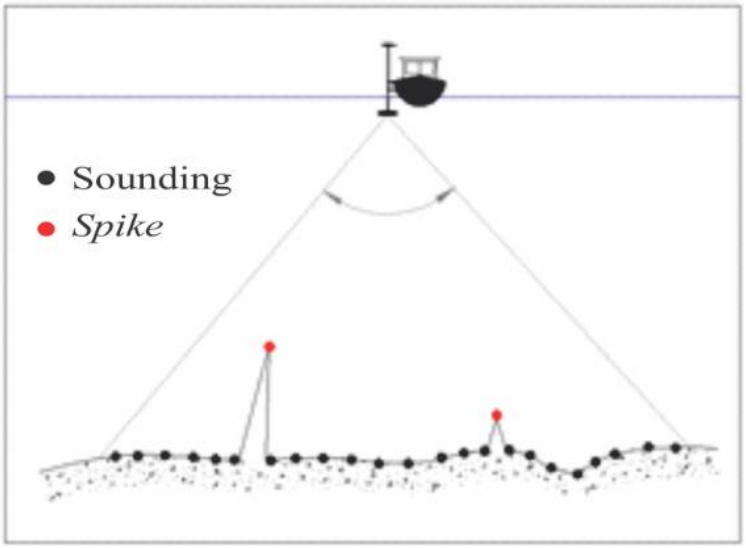

Figure 1 – Bathymetric profile with spikes.

Generally, the detection, analysis and elimination of spikes are performed manually by the surveyor who, visualizing the data through a graphical interface, supposedly decides with a degree of subjectivity which survey may or may not be considered an abnormal value. A priori this task may seem simple, since spikes are random points that do not represent the bottom surface, visibly diverging from it and should in these cases be eliminated. However, due to the large volume of data from a multibeam survey, this task became very time-consuming and, in a way, even more subjective (Ware et al. 1992; Calder and Smith 2003). It is important to note that the analysis of anomalous points should not be ambiguous, that is, they are sometimes interpreted as spurious data, or as belonging to the bottom surface.

Thus, in order to facilitate the task of the surveyor, several authors have developed research in the area of spike detection in bathymetric survey data, such as: Ware et al. (1991; 1992); Eeg (1995); Motao et al. (1999); Debese (2001); Calder and Mayer (2003); Bottelier et al. (2005); Debese (2007); Bjørke and Nilsen (2009); Lu et al. (2010) and Debese et al. (2012).

The first algorithms were based on the generation of bathymetric surfaces, mainly obtained from polynomial functions or weighted averages, followed by the use of filters for the detection and elimination of outliers (Ware et al. 1991; Eeg 1995). With the increase of the computational technology, more robust algorithms were developed, based on M-estimators (Debese and Bisquay 1999; Motao et al. 1999; Debese 2007; Debese et al. 2012), Kalman filters (Calder and Mayer 2003), Kriging techniques (Bottelier et al. 2005), trend surfaces (BJØRKE and Nilsen 2009) and LTS (Least Trimmed Squares) estimator (Lu et al.2010).

Among the various procedures, the CUBE (Combined Uncertainty and Bathymetry Estimator) algorithm presented by Calder (2003) performs well. This algorithm is implemented in the main hydrographic packages and is perhaps the most used semiautomated tool in multibeam data processing (Vicente 2011), including spike research.

However, these methodologies are mostly difficult to apply, semi-automated or are implemented only in commercial packages. Moreover, most of these methods are based on theoretical presuppositions hardly met and almost never verified. According to Vicente (2011), the problem of algorithms, except in the case of CUBE (Calder 2003; Calder and Mayer 2003), remains in their inability to estimate the uncertainty associated with reduced depth.

A technique based on the analysis of standardized residuals was presented by Santos et al. (2017) for terrestrial altimetry data. The methodology, although not automated, since it requires a geostatistical analysis, was very efficient for detection of outliers. However, the nature of the technique imposes a certain pattern of subjectivity on the process, especially in the semivariogram modeling stage, which, according to Ferreira et al. (2013), is a crucial phase of the geostatistical modeling process and should not be automated or neglected.

From the perspective of classical statistics, one of the most commonly used tools for detection of outliers in a univariate continuous data set is the Boxplot or Box Diagram (Tukey 1977; Chambers et al. 1983; Hoaglin et al. 1983). Another commonly used method is the Modified Z-Score, which, unlike the traditional Z-Score, uses robust statistics such as median and absolute deviation, which can ensure that cut-off values are not affected, precisely because of the presence of outliers (Iglewicz and Hoaglin 1993). Several other methods can be applied to detect anomalous values in univariate data sets, composed of continuous quantitative variables, as summarized in Seo (2006).

The problem of applying these methodologies lies on the fact that, besides disregarding the spatial location of the analyzed data, they assume as basic assumptions that observations are independent and identically distributed random variables (Mood et al. 1974; Morettin and Bussab 2004; Seo 2006), indispensable presuppositions for a classic and coherent statistical treatment, but hardly met or theoretically proven. Moreover, in most of these techniques, cut-off values for outlier detections are derived from the normal distribution, which reduces the efficiency of the methods when the sample distribution is asymmetric (Hubert and Vandervieren 2008). However, they are mechanisms of simple application and analysis.

Thus, one can envisage the possibility of developing and applying methods for automated spike detection through the use of these mechanisms in bathymetry data, provided that the methodologies developed take into account the basic statistical assumptions and the spatial dependence structure, inherent to spatially continuous data. Geostatistics is a potential support tool, given its ideal characteristics, that is, spatial modeling with no trend and minimal variance, attributes that can support any outliers detection techniques (Vieira 2000; Matheron 1965). These characteristics were confirmed by Ferreira et al. (2013, 2015 and 2017b) during studies for modeling bathymetric surfaces. Thus, Geostatistics can be used as a tool to support the techniques and algorithms developed in this study.

In view of the above, the main goal of this study is to propose a new methodology for the detection of spikes for bathymetric data collected by beam sounding systems, called SODA (Spatial Outlier Detection Algorithm, in portuguese AEDO - Algorítimo Espacial de Detecção de Outliers). The proposed method employs three outlier detection techniques or thresholds, namely the Adjusted Boxplot, Modified Z-Score and the δ Method. The latter was developed in conjunction with SODA.

Aiming to strengthen the methodology, all the theoretical basis is based on the theorems of classical statistics and Geostatistics.

2. Proposition of the method

The proposed method is based primarily on classical statistics and geostatistics theorems. The entire methodology, including the innovative part, was implemented in free software R (R Core Team 2017). For geostatistical analysis, when necessary, the geoR package, developed by Ribeiro Júnior and Diggle (2001), is used.

Figure 2 illustrates the proposed methodology, called, in this work, SODA. The first step is importing the point cloud (spatial data set). In this phase, the developed algorithm is able to import the three-dimensional coordinates in XYZ format (Shapefile or text file), where X and Y represent, respectively, the positional coordinates, be they local, projected or geodesic, and Z denotes another positional coordinate which represents the reduced depth. In cases where the user decides to import a text file, it is necessary to inform the adopted projection system

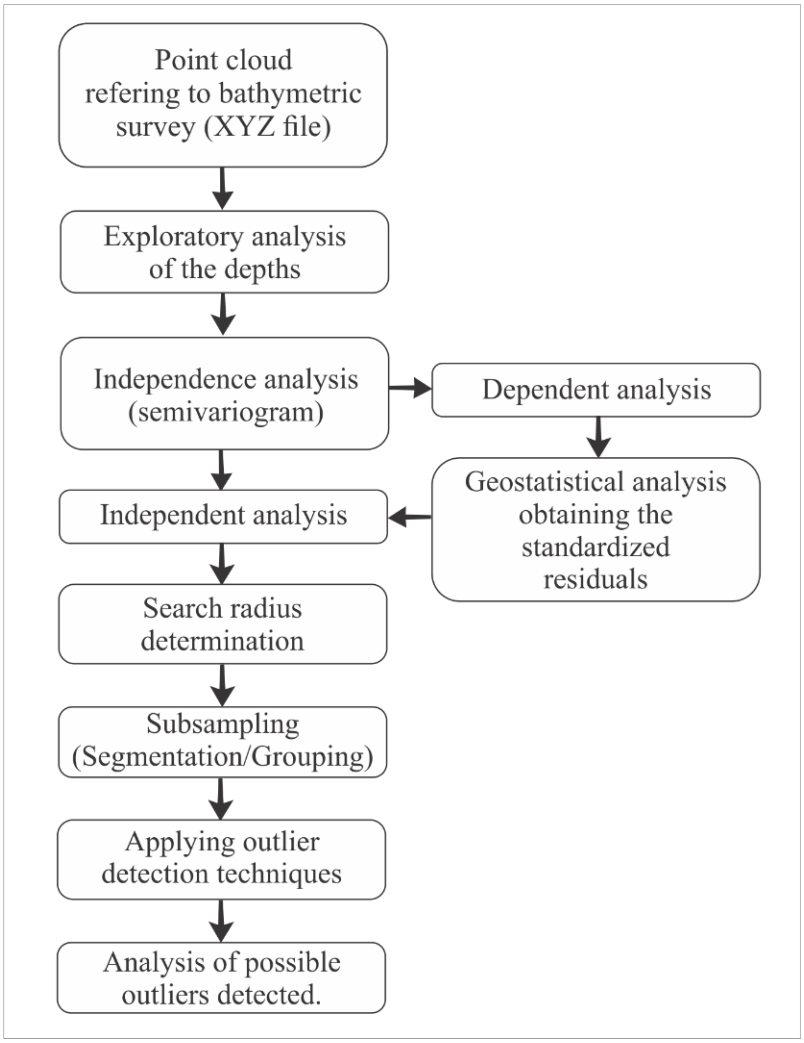

Figure 2 – Flowchart of the proposed methodology for detection of spikes in bathymetry data collected from scanning systems.

Afterwards, the exploratory analysis of the depth data is carried out, an indispensable phase in any statistical and/or geostatistical analysis (Ferreira et al. 2013). Basically, in this step, the method proposes the construction and interpretation of graphs (histograms, Q-Q Plot etc.) and statistics, such as: mean, standard deviation, minimum, maximum, asymmetry and kurtosis coefficients, among others.

The next step is to check for spatial independence between the depth data, a condition assumed by the outliers detection techniques used in this study (Morettin and Bussab 2004; Seo 2006). For this, due to its efficiency, it is suggested the use of semivariogram, a tool used by Geostatistics to evaluate the spatial autocorrelation of the data (Matheron 1965; Ferreira et al. 2015).

The semivariogram of the data (Figure 3), hereafter referred to as experimental semivariogram, is a graph constructed by the function of semivariance versus each value h, where h is the Euclidean distance between the sampled depths. This graph is also known as variogram (Bachmaier and Backes 2011).

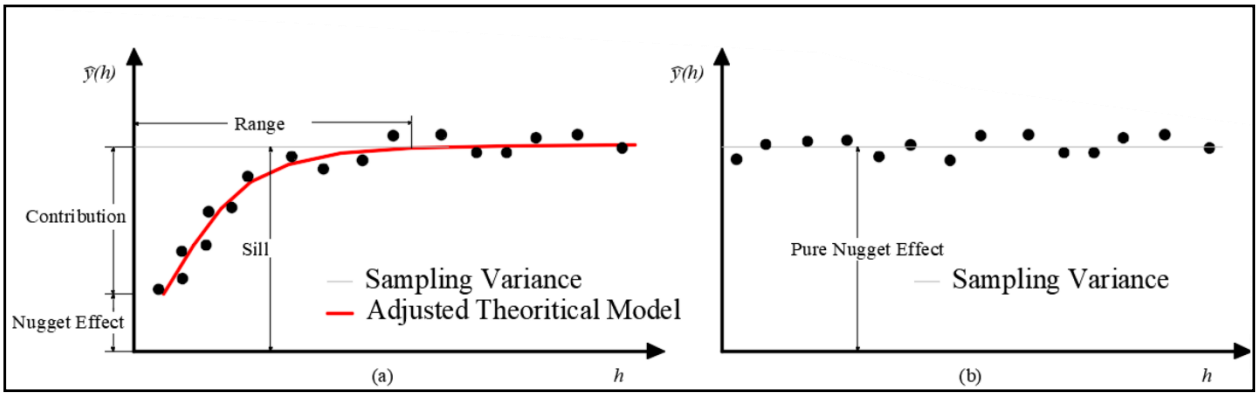

Figure 3 – Example of semivariograms for spatially dependent (a) and spatially independent (b) data (Ferreira 2018).

According to Matheron (1965), the semivariance function is defined as half the mathematical expectation of the square of the difference between the realizations of two variables located in space, separated by the distance h.

Among the estimators of semivariance, we highlight the method based on moments, given by Equation 1:

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

Where $\hat{y}(h)$ is the estimated value of the semivariance for distance h and $N(h)$ is the number of pairs of values $Z(x_i)$ and $Z(x_i + h)$, separated by a distance h. It is expected that $\hat{y}(h)$ increases with distance h, to a maximum value, theoretically the sampling variance, which stabilizes at a sill corresponding to the distance within which the samples are spatially correlated, this distance will be called range. According to Equation 1, one can easily conclude that $\hat{y}(0) = 0$. However, it is common for most of the variables studied, the experimental semivariogram to present a discontinuity for distances smaller than the smaller sampling distance, thus, $\hat{y}(0) \neq 0$. This

phenomenon is known as nugget effect. The difference between the sill and the nugget effect is called contribution (Figure 3a). Finally, in cases where the depth shows no spatial autocorrelation, the semivariogram will only show the pure nugget effect (Figure 3b).

In order to construct the experimental semivariogram, a step distance must be defined to select the depth pairs, and a limit distance for the growth of the steps. According to Santos (2015), one should choose a limit distance (maximum distance in 2D domain) that best represents the spatial the maximum distance between the points (Ribeiro Júnior and Diggle 2001; Diggle and Ribeiro Júnior 2007).

In this sense, SODA calculates the maximum distance between the depths and, based on this information, constructs three semivariograms, the first one with a range equal to 75% of the maximum distance; the second with 50% of the maximum distance and the third with 25%. With these graphs, the analyst can decide on the existence or not of spatial dependence between the data, that is, if at least one of the semivariograms does not present pure nugget effect, the spatial autocorrelation is confirmed. In this step, the algorithm is also able to generate the Monte Carlo envelope (Monte Carlo simulation) to confirm, in an explanatory way, the existence of spatial autocorrelation (Isaaks 1990).

If the spatial dependence is verified, it is suggested to use Geostatistics for the correct statistical treatment of the data. The choice of geostatistics as a support methodology is based on its ideal characteristics. According to Vieira (2000), Geostatistics, in addition to considering the spatial dependence structure of the data, is capable of modeling and predicting without trend and with minimum variance, being, therefore, a very efficient support tool to treat geospatial data.

The semivariogram is the basic support tool for geostatistical techniques and is therefore the most important step of the analysis. The geostatistical inference is based on the assumption of three hypotheses of stationarity, first and second order stationarity and the semivariogram stationarity (Matheron 1965; Ferreira et al. 2013). However, as Vieira (2000) affirms, it is commonly assumed only the intrinsic or semivariogram stationarity hypothesis, that is, it is assumed that the variogram exists and is stationary for the variable in the study area.

When the semivariogram shows an identical behavior for all directions, it is said to be isotropic, otherwise it is said to be anisotropic. When anisotropy is detected, it must be corrected, usually through linear transformations, since it prevents the existence of stationarity, a necessary condition for accuracy of the analysis and estimates for the area under study (Isaaks 1990; Vieira 2000; Ferreira et al.2013; 2015).

Once the experimental semivariogram is obtained, one can then adjust it through theoretical models. This adjustment consists of modeling the spatial dependence itself, so it must be done with caution. Uncertainties in this adjustment will lead to prediction uncertainties (Ferreira et al.2013). With the adjusted theoretical model, values can be predicted in non-sampled sites, considering the spatial variability of the data (Vieira 2000; Santos 2015).

There are several isotropic models in the literature, which contemplate semivariograms with and without sill. Among the models without sill, stands out the power model and among the most those with sill (the most common), stand out the exponential model, the spherical model and the Gaussian model (Vieira 2000). After modeling the semivariogram, we can predict non-sampled values, without bias and minimum variance, through the geostatistical interpolation method called kriging. Further details on geostatistical modeling can be found, for example, in Vieira (2000) and Ferreira et al. (2013).

After the geostatistical modeling, the process of leave-one-out cross-validation is performed which, according to Ferreira et al. (2013), is the procedure that quantifies the uncertainties inherent to modeling and prediction process, due to the assumptions made or, more commonly, to the fit of the model. This technique consists of temporarily withdrawing a sampled value and predicting the value using the theoretical model adjusted to the other sampled values. At the end, modeling residuals are obtained, that is, the difference between the observed values and their corresponding predicted (Vieira 2000). From these residues one can evaluate the quality of the estimate.

According to Santos et al. (2017), these residuals are known as white noise, random noise or random walk and in their standardized form, hereinafter referred to as SR (Standardized Residual), have important statistical characteristics, namely: follow normal distribution with zero mean and unit variance, are independent, unbiased and homogeneous.

After confirming the spatial independence, either from the depths or from the SRs, we proceed with the application of the proposed methodology (Figure 2). Thus, the next step is the segmentation of the sample, which aims, first and foremost, to preserve the spatial characteristic of the analysis (local analysis). This subsampling also allows for a considerable reduction of machine processing time.

As already discussed, the methodologies for outlier detection based on classical statistics assumes that observations are independent random variables and identically distributed (Morettin and Bussab 2004; Seo 2006). Thus, the subsampling step proposed in this study is based on 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 theoretical examples, can be found in Mood (1913) and Mood et al. (1974).

From this perspective, SODA applies a segmentation called, in this study, ***Segmentation in Circles*** (Figure 4). Thus, the algorithm generates a centered circle, which a 2D domain is defined, at each depth or SR, identifying and storing all the data present inside the circle into subsamples. All analysis, from that moment on, is then performed only on these subsamples.

The circle radius or ***Search Radius*** may be defined by the user or based on spatial analysis. It is emphasized that this greatness is closely linked to the bottom morphology. As the submerged relief is not visible, the determination of this radius by the analyst becomes quite subjective. For the time being, it is only known that in those places where the presence of a flat relief is clear, one can adopt larger circles.

Alternatively, it is suggested that the radius is equivalent to three times the minimum distance. In this case, the algorithm is able to compute the smallest distance between points and assign three times that value to the radius of the circle. Such a suggestion, a priori, has no theoretical basis, therefore comes from experimentation, and aims to eliminate the intervention of the analyst, automating the process. It is based on the assumption that the point cloud, acquired from a beam echo sounding system, is dense and without holidays. Thus, this radius is able to guarantee a local investigation, with subsamples containing enough points for analysis. The Figure 4 below illustrates the procedure.

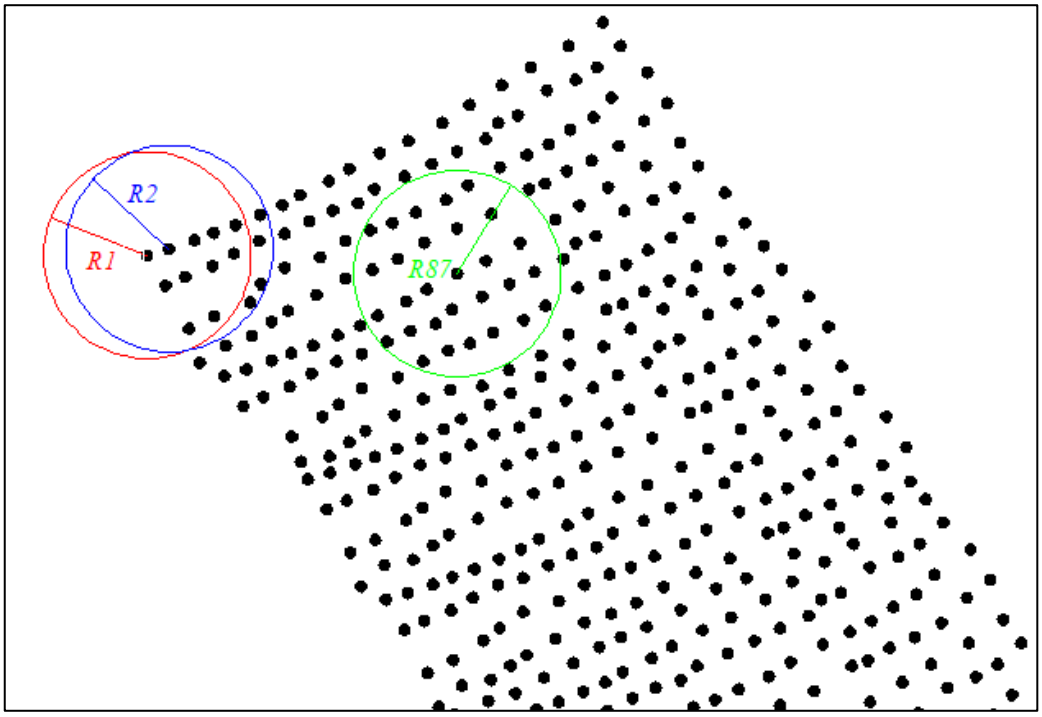

Figure 4 – Example of the Segmentation in Circles technique employed by SODA

Figure 4 presents a cloud of points containing bathymetry data collected by a beam echo sound system. The search radius has been set by the user. The circle generated for point 1 in red captured 11 points, while the circle generated for point 2 in blue captured exactly 15 points. Note that, of these 15 points, 11 coincide with those captured by the previous circle. It is Suggested that in SODA, depths or SRs will be analyzed more than once, logically depending on the search radius and the density of the cloud of points. This fact brings gains to the proposed methodology, as will be seen later, and therefore, the algorithm stores this information, aiming to use them later.

In green, the circle generated for point 87, containing within it 29 points. It is important to ensure that subsamples have sufficient points for consistent statistical analysis. Thus, the algorithm verifies the number of points present in each subsample. If the number is less than 7 (empirical basis), the subsample is disregarded in subsequent analyses.

With the subsamples, SODA applies three outlier detection techniques: Adjusted Boxplot (Vandervieren and Hubert 2004); the modified Z-Score (Iglewicz and Hoaglin 1993) and the δ Method, proposed in this work. The δ Method was inspired in parts by the technique proposed by Lu et al. (2010), which consists of applying spike detection thresholds based on the global and local sampling variance, where the local variance refers to variance of subsamples. In this method, if the global variance is greater than the local variance, the cutoff value is set to be equal to the global variance, otherwise the threshold is set to $0.5 \cdot (\sigma^2_{Global} + \sigma^2_{Local})$, where σ^2 is the variance. And so, any observation that has greater residual, in absolute value, than the cutoff value, is considered a spike. However, as already discussed, the standard deviation and, therefore, the sampling variance of the data set, are dispersion measures that are not resistant to outliers.

On the other hand, the theory of errors states that when a normal distribution can be assumed, 68.3% of the data evaluated are within the range $\mu \pm \sigma$; 95% are within the range $\mu \pm 1.96\sigma$ and 99.7% of the data evaluated are within the range $\mu \pm 3\sigma$ (Mood 1913; Mood et al. 1974). Based

on these conjectures, it is very common, especially in the geodesic sciences, to eliminate outliers by applying the threshold $3 \cdot \sigma$ (unbiased data) or $3 \cdot RMSE$ (biased data), where $RMSE$ is the root mean square error (Mikhail and Ackerman 1976; Cooper 1987; Höhle and Höhle 2009).

Therefore, the δ Method is a proposition consisting of a new spatial threshold of outlier detection, based on robust estimators, given by Equation 2:

$$Threshold = Q2_{Local} \pm c \cdot \delta \tag{2}$$

Where $Q2_{Local}$ is the median of the subsampled data and c and δ are constants that depend on data variability. The constant c assumes the value 1 for irregular reliefs or artificial channels (high variability); 2 for undulating reliefs (medium variability) and 3 for flat reliefs (low variability). This value can be understood as a weight of the constant δ and must be entered by the user. The constant δ is determined automatically by the algorithm through the evaluation of the Global Normalized Median Absolut Deviation ($NMAD_{Global}$) or Local ($NMAD_{local}$), that is, $\delta = 0.5 \cdot (NMAD_{Global} + NMAD_{local})$, if $NMAD_{Global} > NMAD_{local}$, or otherwise, $\delta = NMAD_{Global}$.

In view of the above, in the hypothesis that SODA uses the thresholds set by the δ *Method*, as well as by the *Modified Z-Score*, it is indirectly assumed that the subsamples have a normal distribution. This is due to two main factors. The first is that it is not possible to perform hypothesis tests to determine the probability distribution of each subsample, and the second factor, even used to justify the first, lies in the fact that the normal distribution is the most important continuous probability distribution and, for this reason, used in most applied statistical techniques (Mood 1913; Mood et al. 1974). Thus, the *Adjusted Boxplot* may have advantages, since it intrinsically considers the possible asymmetry of the sampling distribution.

After applying the outlier detection thresholds, in the next step, the proposed method determines the probability of the data being an outlier ($P_{outlier}$) in each of the three techniques, based on the number of times the data was analyzed ($N^{o}_{analyzed}$) and the number of times it was considered an outlier ($N^{o}_{outlier}$) (Equation 3):

$$P_{outlier}(\%) = \frac{N^{o}_{outlier}}{N^{o}_{analyzed}} \cdot 100 \tag{3}$$

For example, consider that, given any search radius, the observation i was subsampled 20 times (Figure 4). Thus, it was analyzed by the three techniques of detection of outliers in these 20 times. Also consider that, among the 20 times, in 10 of them observation i was considered an outlier by the δ *Method*, hence $P_{outlier} = (10/20) = 0.5$, that is, observation i has 50% probability of being an outlier if the cutoff limit considered is that given by δ *Method*. The demonstration for the other thresholds adopted by SODA is identical. Table 1 illustrates the information for this step.

Table 1 – Analysis of the probability of depth *i* be a *spike*.

Point	$N^{o}_{analyzed}$	$N^{o}_{outlier}$	$P_{outlier}$ (%)
1	6	6	100.00
2	15	5	33.33
⋮	⋮	⋮	⋮
n	27	20	74.07

Finally, by defining a default $P_{threshold}$ by the user, SODA spatially plots all observations, highlighting the spikes detected by the thresholds used, that is, all outliers with $P_{outlier} \geq P_{threshold}$. The user then performs a visual inspection to confirm the spikes and subsequently eliminate them. In all cases, new XYZ files for each technique are created, that is, the SODA associated with, respectively, the *Adjusted Boxplot*, *Modified Z-Score* and δ *Method*.

This last step requires extra caution in the sense that if there is any doubt about the possible spike, one should refine the analyses and, depending on the purpose of the survey, return to the sounding area to conduct a hazard survey. It is very common, depending on the density of sounding, the analyst confuses marine features or even sunken objects with spikes and thus, mistakenly treat them as such.

3. Experiments and results

Aiming to evaluate the robustness of SODA, as well as to make adjustments, we used the computational simulation. A study area similar to a navigation channel was constructed using simulated data, as shown in Figure 5.

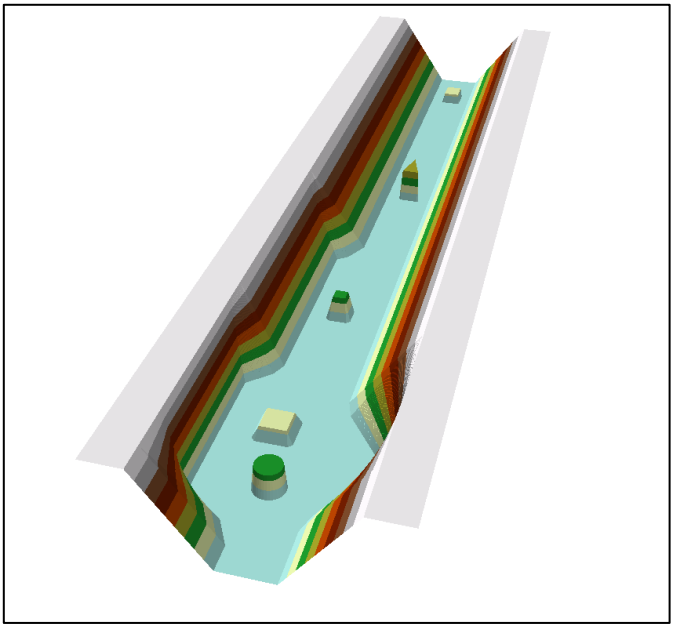

Figure 5 – Three-dimensional bathymetric surface constructed through computational simulation.

The simulated bathymetric surface has an area of 1,600m² (80 x 20m) with submerged relief varying from 8 to 15 m, two lateral slops with average slants of 135% and five underwater structures, located in the channel bed, which reproduce dangers to navigation, such as: sandbars, rocks and hull of the sunken ship. These features are represented by known geometric solids (parallelepiped, cone trunk, pyramid trunk, etc.) and have heights varying between 1 and 3 m. From this surface, the data set was composed of 40,000 bathymetric points, initially without outliers, 20 cm spaced apart, that is, 25 points/m².

Table 2 summarizes the descriptive statistics information of the study area.

Table 2 – Descriptive statistics of the simulated study area.

Number of Observations	40,000
Average depth (m)	11.607
Minimum Depth (m)	8.000
Maximum Depth (m)	15.000
Variance (m²)	8.6130
Coefficient of kurtosis	1.300
Coefficient of skewness	-0.050

Since they are simulated data, the phases of exploratory analysis and sample independence are not detailed, however, it is evident that the dataset is spatially independent. Therefore, the proposed method was applied directly to the depths. The search radius, as explained in section 2, was defined as three times the minimum distance between points, i.e., 0.60 m. From this radius, the original sample was segmented into 40,000 subsamples and the outlier detection thresholds were applied. In the first step, the *Adjusted Boxplot*, *Modified Z-Score* and δ *Method*, located, respectively, 3,476 (8.69%), 7,742 (19.35%) and 453 (1.13%) possible spikes. Since it is a navigation channel, the constant c of δ *Method* has been set to the unit value.

Aiming to evaluate the agreement between the methods, a comparative analysis of the possible spikes located by each threshold was performed, and it was concluded that of the 453 points detected by the δ *Method*, 391 were also detected by the *Modified Z-Score* and 182 by the *Adjusted Boxplot*. That is, taking the smallest set of data as reference, there is a concordance of, respectively, 86.31% and 40.18%. The concordance between the *Adjusted Boxplot* and the *Modified Z-Score* was 98.36%, that is, of the 3,476 possible spikes located by the *Adjusted Boxplot* threshold, about 3,419 were also detected by the *Modified Z-Score* threshold.

In sequence, the analysis was refined by calculating the probability that the data was a spike for each of the three techniques, based on the number of times the depth was analyzed and the number of times it was considered an outlier. For the execution of this step, a priori, it is suggested a $P_{threshold} = 50\%$, that is, if $P_{outlier} \geq 0.5$, the depth analyzed is considered a spike.

After this step, the *Adjusted Boxplot* and δ *Method* thresholds, as expected, did not locate any spikes. On the other hand, the *Modified Z-Score* method, in a wrong way, signaled 287 (0.72%) points as possible outliers, among them the depths of the submerged structures, as shown in Figure 6a, where the network of bathymetric points is plotted in blue and spikes highlighted in red. In the case of real data processing, the elimination of such sounding data representing hazards

to navigation could cause serious problems, such as ship and boat stranding, damages to the hull of ships and even a shipwreck.

Analyzing the metadata of the detected outliers, it was noticed the need to make an adjustment in the $P_{threshold}$ of this specific cutoff value. After some tests and simulations, an optimum value of 80% was reached, that is, $P_{threshold} = 0.8$ (Figure 6b). Thus, based on the simulated data, it is recommended a $P_{threshold} = 0.5$ for the *Adjusted Boxplot* and δ *Method* and a $P_{threshold} = 0.8$ for the *Modified Z-Score*.

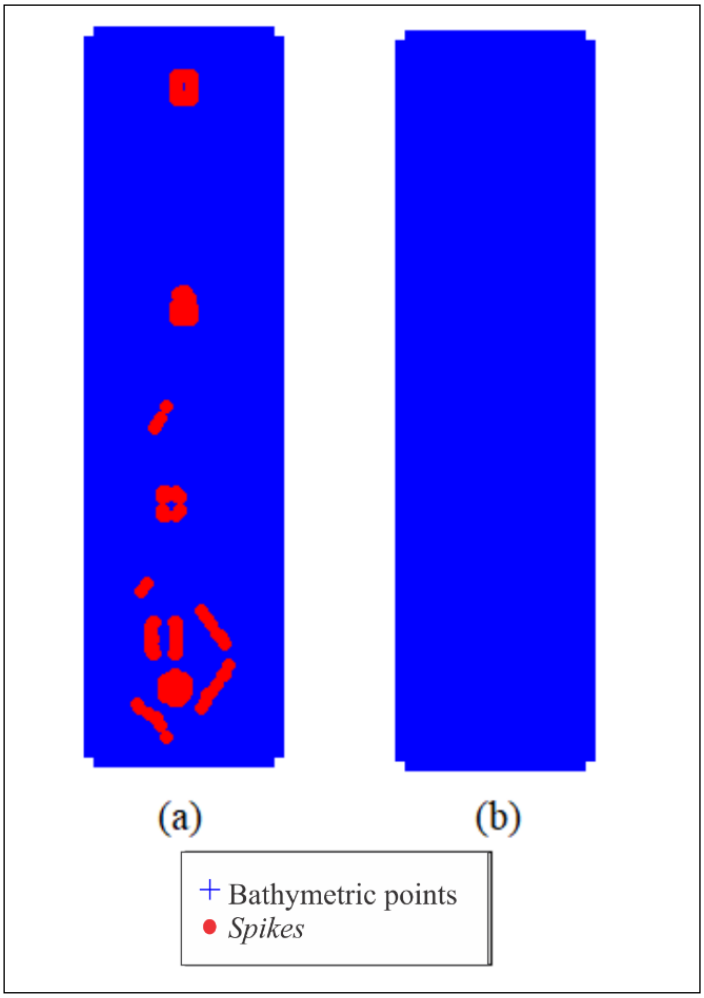

Figure 6 – *Spikes* detected by SODA from *Modified Z-Score* thresholds, for $P_{threshold} = 0.5$ (a) and $P_{threshold} = 0.8$ (b).

Following the simulations, ten spikes were introduced into the data set, with horizontal positions chosen randomly. The magnitude of these spikes was determined based on practical knowledge acquired from the real data processing (Table 3). The reduced number of spikes will allow later a more thorough examination.

Table 3 – *Spikes* randomly inserted into the data set.

Point ID	$X\ (m)$	$Y\ (m)$	$Z\ (m)$	Z_{spike} (	*Spike* Magnitude (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

Applying the methodology proposed in a similar way to the previous, at first, the *Adjusted Boxplot*, *Modified Z-Score*, and δ *Method*, located, respectively, 3,458 (8.65%), 7,749 (19.37%) and 458 (1.15%) possible spikes. Of the 458 spurious depths determined, a priori, by the δ *Method*, 187 were also located by the *Adjusted Boxplot* (40.83%) and 396 by the *Modified Z-Score* (86.46%). While the agreement between the Adjusted Boxplot and the Modified Z-Score was approximately 98%, this is, of all the points detected by the both techniques, just 2% was different.

Later, $P_{threshold} = 0.5$ was defined for the *Adjusted Boxplot* and δ *Method* and $P_{threshold} = 0.8$ for the *Modified Z-Score*. The results are summarized in Table 4.

Table 4 – Result of data processing of the simulated study area.

Point ID	*Spike* Magnitude (m)	*Adjusted Boxplot* ($P_{outlier}(\%) \geq 50$)	*Modified Z-Score* ($P_{outlier}(\%) \geq 80$)	δ *Method* ($P_{outlier}(\%) \geq 50$)
2751	1.50	Detected	Detected	Non-detected
2659	2.00	Detected	Detected	Non-detected
15051	4.00	Detected	Detected	Detected
15260	3.00	Detected	Detected	Detected
17573	0.60	Non-detected	Non-detected	Non-detected
19611	2.00	Detected	Detected	Non-detected
23557	4.80	Detected	Detected	Detected
25943	5.00	Non-detected	Detected	Detected
35141	3.00	Detected	Detected	Detected
38738	4.00	Detected	Detected	Detected

The *Adjusted Boxplot* threshold detected all inserted spikes, except for ID points 17573 and 25943, whose error magnitude is respectively 0.60 and 5 meters. Thus, the percentage of success was 80%, that is, of the 10 inserted spikes, the *Adjusted Boxplot* threshold located 8.

Point ID 17573 was not detected due, in particular, to its low magnitude for the relief surveyed. However, of the 29 times that this point was analyzed, it presented itself as a spike on 11 occasions, that is, a 38% probability. On the other hand, it is clear that the failure to detect the ID 25943 point is not related to the magnitude of the error or to the applied threshold, since spikes of lower magnitudes were located. Thus, this fact may be closely related to the neighborhood of the analyzed outlier. The point 25943 is positioned, horizontally, on a submerged structure, near the edge, that is, on the crest of the slope. However, as can be seen in Table 5, this point had a $P_{outlier}(\%) = 48\%$ very close to the adopted $P_{threshold}$. All other points, considered outliers in the first step of the SODA method, obtained $P_{outlier}$ less than 28%.

The *Modified Z-Score* obtained a percentage of success of 90%, that is, it was able to detect all spikes, except point ID 17573, which has, as mentioned above, an error with a magnitude much lower than those experienced in hydrographic practice. This point had a $P_{outlier}(\%) = 21\%$. Of the other inserted spikes, 8 of them reached $P_{outlier} = 100\%$, which shows the efficiency of this threshold. On the other hand, approximately 130 points obtained a $P_{outlier}$ varying between 60% and 79%, many of them representing submerged structures, suggesting greater care in subsequent analyses.

Finally, the δ *Method* detected 60% of the inserted spikes, all with a $P_{outlier} = 100\%$. Points ID 2751, 2659, 17573 and 19611 obtained a $P_{outlier} = 0\%$ and thus were not detected (Table 4). Analyzing Table 3, it is easy to notice that the failure in the location of these points is related to the magnitude of the errors, that is, the δ *Method* was able to detect, for the relief in question, only the spikes with magnitude greater than 2 meters. In this analysis, it should be noted that the threshold in question is based on a robust data variability estimator, which may be ineffective for very regular data, such as the analyzed set, which have several exactly flat data, i.e., with the same depth and consequently $NMAD = 0$.

All other possible spikes had a probability of less than 28%, except for 6 points, which had a $P_{outlier}$ ranging from 35% to 48%. These points represent the crest of the submerged structure of triangular flat shape, with a height of 1 meter.

Table 5 and Figure 7 summarize and illustrate, respectively, the information discussed.

Table 5 – $P_{outlier}(\%)$ of data of the study area.

Point ID	*Adjusted Boxplot*	*Modified Z-Score*	δ *Method*
	$P_{outlier}$	$P_{outlier}$	$P_{outlier}$
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%

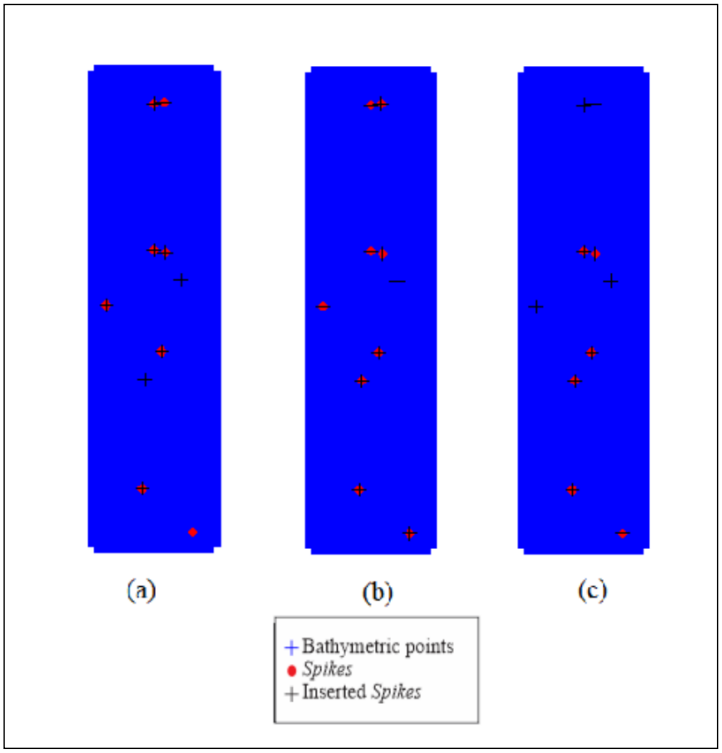

Figure 7 – Inserted *Spikes* and *Spikes* detected by SODA from thresholds of the *Adjusted Boxplot* (a), *Modified Z-Score* (b) and δ *Method* (c).

In general, the proposed methodology, i.e., SODA presented efficiency and versatility. Although the δ *Method* presented only 60% accuracy, it was able to detect all spikes with magnitude greater than 2 meters and, perhaps the most important point, no submerged structure belonging to the channel relief was signaled as spike, preserving, in these cases, navigation safety. Similarly, the other thresholds used by the proposed methodology also achieved optimum results.

It should be noted that the implemented algorithm performed all the processing of this data set in approximately 2 hours and 45 minutes, using a machine with Windows 10 operating system, 8GB RAM (partially dedicated to software R) and Intel® Core™ i7-4500U CPU @ 1.80GHz 2.40 GHz processor.

4. Conclusions

It can be concluded from the obtained results that the initial objectives were met, since the proposed methodology presented robustness in the detection of spikes for the simulated bathymetric data. SODA, implemented with the support techniques: *Adjusted Boxplot, Modified Z-Score* and δ *Method*, showed characteristics of great interest to the support of outlier identification techniques.

Data used in this research were simulated by similar practices found in a beam echo sounder survey, that is, considering natural and artificial aspects of submerged relief modeling, inserting navigation hazards such as: sandbars, rocks and hull of the sunken ship, usually found in

submerged areas. These data were established with the purpose of validating the proposed methodology, since it allows for analysis in a controlled environment.

It was verified that the performance of SODA associated with the *Modified Z-Score* threshold was superior, compared to the others, with identification of 90% of the outliers introduced in the simulation. It should be noted that the unidentified spike had an error of magnitude well below the values experienced in hydrographic practice. The *Adjusted Boxplot* method also presented a satisfactory result, considering an 80% success in identifying the simulated outliers. On the other hand, the δ *Method*, although presenting only a 60% accuracy rate, was able to detect spikes with a magnitude greater than 2m. In all cases, the methodology proved to be effective in terms of possible erroneous identification of submerged structures, known belonging to the channel relief, a very interesting result when regarding the construction of bathymetric models for navigation.

Importantly, the use of theorems of classical statistics and Geostatistics was fundamental to strengthen the methodology used. Finally, it is recommended for future studies, the use of real data to analyze the performance of SODA.

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

Artilheiro, F. M. F. 1998. 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.

Bachmaier M and Backes, M. 2011. Variogram or Semivariogram? Variance or Semivariance? Allan Variance or Introducing a New Term? Mathematical Geosciences, v. 43, n. 6, p. 735-740.

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

Bottelier, P. et al. 2005. Distinguishing features from outliers in automatic Kriging-based filtering of MBES data: a comparative study. Springer Berlin Heidelberg.

Calder, B. R. and Mayer, L. A. 2003. Automatic processing of high-rate, high-density multibeam echosounder data. Geochemistry, Geophysics, Geosystems, v. 4, n. 6.

Calder, B. R. and Smith, S. 2003. 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.

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

Chambers, J. M. et al. 1983. Graphical Methods for Data Analysis. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Cheng, L. et al. 2015. 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.

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

Cruz, J. et al. 2014. Benefícios da utilização de sondadores interferométricos. 3as Jornadas de Engenharia Hidrográfica. Instituto Hidrográfico Português, Lisboa, Portugal.

Debese, N. 2001.Use of a robust estimator for automatic detection of isolated erros appearing in the bathymetric data. International Hydrographic Review. V. 2, no. 2, September.

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

Debese, N.; Moitié, R.; Seube, N. 2012. Multibeam echosounder data cleaning through a hierarchic adaptive and robust local surfacing. Computers and Geosciences, v. 46, p. 330-339.

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

Diggle, P. J. and Ribeiro Júnior, P. J. 2007. Model-based Geostatistics. New York: Springer, 229p.

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

Ellmer, W. et al. 2014. Feasibility of Laser Bathymetry for Hydrographic Surveys on the Baltic Sea. The International Hydrographic Review, n. 12, p. 33-50.

Ferreira, Í. O. et al. 2016a .Modelo de incerteza para sondadores de feixe simples. Revista Brasileira de Cartografia, v. 68, n. 5, p. 863-881.

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

Ferreira, Í. O.; Rodrigues, D. D.; Santos, G. R. 2015.Coleta, processamento e análise de dados batimétricos. 1ª ed. Saarbrucken: Novas Edições Acadêmicas, v. 1, 100p.

Ferreira, Í. O. et al. 2017b. In bathymetric surfaces: IDW or Kriging? Boletim de Ciências Geodésicas, v. 23, n. 3, p. 493-508.

Ferreira, Í. O.; Santos, G. R.; Rodrigues, D. D. 2013.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.

Ferreira, Í. O. et al. 2016b. 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.

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

Guenther, G. C.; Thomas, R. W. L. ; Larocque, P. E. 1996. 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.

Hoaglin, D. C.; Mosteller, F.; Tukey, J. W. 1983. Understanding robust and exploratory data analysis. New York: Wiley, 433p.

Höhle, J. and Höhle, M. 2009. 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.

Hubert, M. and Vandervieren, E. 2008. An adjusted boxplot for skewed distributions. Journal of Computational statistics and data analysis, v. 52, n. 12, p. 5186-5201.

Iglewicz, B. and Hoaglin, D. 1993. How to detect and handle outliers. Milwaukee, Wis.: ASQC Quality Press, 87p.

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

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

Instituto hidrográfico. 2009. Especificação Técnica para Produção de cartografia hidrográfica. Marinha Portuguesa, Lisboa, Portugal, v 0.0, 24p.

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

Jong, C.D. et al. 2010. Hydrography. 2ª ed. Delft University Press: VSSD, 354p.

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

Lu, D. et al. 2010.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.

Maleika, W. 2015. 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.

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

Mikhail, E. and Ackerman, F. 1976. Observations and Least Squares. University Press of America, 497p.

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

Mood, A. M.; Graybill, F. A.; Boes, D. C. 1974. Introduction to the Theory of Statistics. McGraw-Hill International, 577p.

Morettin, P. A. and Bussab, W. O. 2004. Estatística básica. 5ª ed. São Paulo: Editora Saraiva, 526p.

Motao, H. et al. 1999. Robust method for the detection of abnormal data in hydrography. The International Hydrographic Review, v. 76, n. 2, p. 93-102.

Moura, A.; Guerreiro, R.; Monteiro, C. 2016.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.

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

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

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

Ribeiro Júnior, P.J. and Diggle, P.J. 2001. GeoR: a package for geostatistical analysis. R-News. v. 1, p. 15-18.

Rousseeuw, P. J. and Croux, C. 1993. Alternatives to the median absolute deviation. Journal of the American Statistical association, v. 88, n. 424, p. 1273-1283.

Santos, A. M. R. T.2017. Detection of inconsistencies in geospatial data with geostatistics. Boletim de Ciências Geodésicas**,** v. 23, n. 2, p. 296-308.

Santos, A. P. 2015. 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.

Santos, A. P. 2016. 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.

Seo, S. 2006. 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.

Tukey, J.W. 1977. Exploratory Data Analysis. Princeton, Ed. Pearson.

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

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

Vandervieren, E. and Hubert, M. 2004. An adjusted boxplot for skewed distributions. Proceedings in Computational Statistics, p. 1933-1940.

Vicente, J. P. D. 2011.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.

Vieira, S. R. 2000. 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.

Ware, C. et al. 1992. A System for Cleaning High Volume Bathymetry. The International Hydrographic Review, v. 69, n. 2, p. 77-94.

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