Multibeam outlier detection by clustering and topological persistence approach, ToMATo algorithm

Marceau Michel, Julian Le Deunf, Nathalie Debese, Laurène Bazinet, Loïc Dejoie

Multibeam outlier detection by clustering and topological persistence approach, ToMATo algorithm

Marceau MICHEL
ENSTA Bretagne
Brest, France
marceau.michel@ensta-bretagne.org

Julian LE DEUNF
Shom *(Service hydrographique et océanographique de la Marine)*
IMT Atlantique, Lab-STICC, UMR CNRS 6285
Brest, France
julian.le.deunf@shom.fr

Nathalie DEBESE
ENSTA Bretagne, Lab-STICC, UMR CNRS 6285
Brest, France
nathalie.debese@ensta-bretagne.fr

Laurène BAZINET
ENSTA Bretagne
Brest, France
laurene.bazinet@ensta-bretagne.org

Loïc DEJOIE
ENSTA Bretagne
Brest, France
loic.dejoie@ensta-bretagne.org

***Abstract*— The datasets acquired during hydrographic surveys contain outliers, i.e., soundings that do not describe the sea bottom. Many algorithms are developed to identify them. Here, we study unsupervised non-parametric algorithms with a density-based approach. These algorithms make no assumption about the data and identify outliers as the data furthest away from their neighbors. We asses the ToMATo method developed by INRIA in 2009 to detect outlier soundings from multibeam echosounder data. This clustering algorithm combines a mode-seeking phase with a cluster merging phase using topological persistence. After the theoretical presentation of the ToMATo algorithm, we evaluate its performance on four data sets representing a wide variety of seabeds. We compare this method with the well-known DBSCAN and LOF algorithms. Finally, we suggest an application of the ToMATo algorithm to multibeam data acquired in extra-detection mode, where topological persistence allows to form the most relevant clusters.**

Keywords—Data processing, Multibeam data, Outlier detection, ToMATo clustering

I. Introduction

Multibeam data processing is a critical task for the elaboration of nautical charts and its automation is challenging as shown in the review of methods presented in [1]. During the acquisition of bathymetric data, many sources of noise can disturb the acoustic sounder, resulting in soundings that do not describe the sea bottom. Three types of errors can be distinguished: systematic errors, outliers, and random noise. Systematic errors are mainly due to the system installation or the complexity of the environment: calibration procedure (e.g., patch test) prior to the survey and good practice in surveying should remove systematic errors. The random noise measurement, due to noise in the measurement process, is estimated and evaluated according to the minimum standards for hydrographic surveys [2]. Finally, outliers which can be caused by occasional sensor dysfunctions, human error, or environmental phenomena such as the presence of fish or plumes of hydrothermal mounts for example. As presented in [3], these outliers are removed from the dataset since they do not represent the seabed.

We applied the ToMATo (Topological Mode Analysis Tool) clustering method developed by INRIA in 2009 [4] to detect outliers in MBES bathymetric datasets. ToMATo was included in a bathymetric data processing pipeline and tested on reference datasets. Then, we compared the results of this method with those of two other clustering algorithms already known for the processing of bathymetric data: DBSCAN [5] and LOF [6]. Finally, we demonstrated the advantage of this method over existing methods by applying it to multibeam extra-detection data to identify objects in the water column.

II. ToMATo algorithm

ToMATo algorithm is a clustering algorithm, which combines a gradient ascent algorithm, with a cluster merging phase based on topological persistence. It uses the topological persistence method, introduced by H. Eldelsbrunner [7], which allows separating information of a topological nature from noise for data represented by functions on a topological space. The objective of the topological persistence approach is to detect clusters and merge unstable ones to regain stability. For the implementation of this method, we used the C++ package written by P. Skreaba and S. Oudot from INRIA available at https://geometrica.saclay.inria.fr/data/ToMATo/.

We illustrate the operation of the algorithm on a simulated ping (*Fig. 1*).

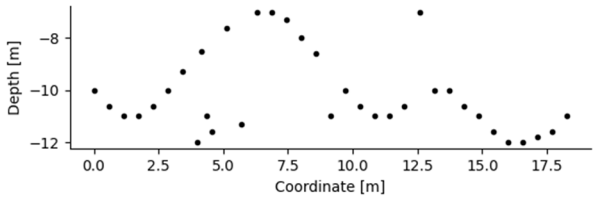

Fig. 1. Simulated ping before ToMATo clustering

A. Rips-graph and density estimation

ToMATo algorithm takes three inputs: the neighborhood graph, the density estimator, and the merging parameter τ.

1. Neighborhood graph

The ToMATo package proposes an implementation of the *δ-Rips graph* for the construction of the neighborhood graph.

This method connects two points in the graph when their distance is less than δ. In bathymetric data processing, δ can be seen as the data inspection scale. This data inspection scale is illustrated by the red circle in the figure below (*Fig. 2*) where soundings that lies at a distance less than δ are connected by a segment. An issue with multi-beam data is that one does not want to keep the same inspection scale along all directions. Indeed, the distance between two successive pings is often larger than the spacing between the soundings of a ping. Moreover, the vertical variations are often very small compared to the horizontal variations. Therefore, we propose to standardize the bathymetric data along each axis before proceeding to the construction of the neighborhood graph. This standardization has significantly improved the performance of ToMATo method for processing bathymetric data.

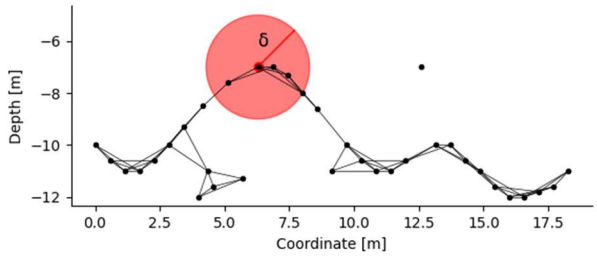

Fig. 2. Rips graph

2. Density estimator

ToMATo is less sensitive to the choice of the density function. Once again, we used the estimator available in the ToMATo package that uses a Gaussian kernel on a neighborhood, typically 50 nearest neighbors, to estimate the density. In the figure bellow (*Fig. 3*), the density is represented by the radius of the point.

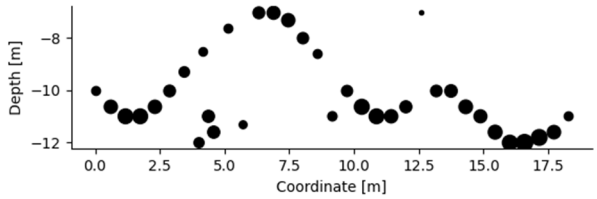

Fig. 3. Density estimator

3. Merging parameter

The merging parameter, or persistence threshold τ, is used in the cluster merging phase. Clusters of prominence less than τ are eventually merged into clusters of prominence as least τ. The prominence of a cluster is defined as "*the difference between its height and the level at which its basin of attraction meets the one of a higher peak (its parent in the hierarchy)*", where the basin of attraction of a peak is defined as all the points that can reach this peak by some greedy hill-climbing procedure [4]. More intuitively, the persistence threshold allows us to choose the peaks that we consider to be true to aggregate the clusters that are in their basin of attraction, and to eliminate the clusters that do not belong to it and whose persistence is too low (noise). In the case of bathymetric data, the highest peak prominence corresponds to the sea floor. The merging parameter τ is chosen after a first run on the ToMATo algorithm using the persistence diagram.

B. *Mode-seeking*

The mode-seeking phase is realized by a standard gradient ascent method in a neighborhood graph. The clustering algorithm by gradient ascent in a graph was initially proposed by Koontz in 1976 [8].

It consists in constructing a spanning forest of the neighborhood graph G by connecting each vertex v to its neighbor in G with the highest density estimator f. If all the neighbors of v have lower f values than v, then v is connected to itself and is declared as a peak of f: it thus becomes the root of one or several trees of the spanning forest. This construction is very sensitive to perturbations of the density function. In practice, this instability can become critical because density estimators tend to be noisy. In the figure bellow (*Fig. 4*), each cluster formed by this method is represented by a unique color.

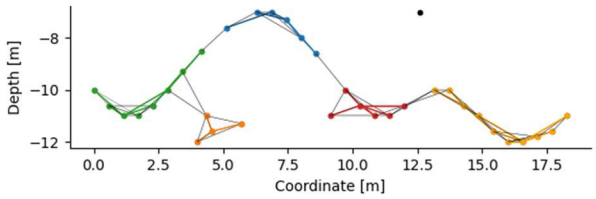

Fig. 4. Mode-seeking

C. *Merging*

The objective of the merging phase is to merge unstable clusters, produced during the mode-seeking phase, to regain stability. The persistence threshold can be chosen using the persistence diagram (PD) (*Fig. 5*). This diagram illustrates the life span of clusters: the ordinate axis corresponds to the creation time of the cluster, and the abscissa axis corresponds to its disappearance time, i.e., when it is merged with a new cluster. The highest prominence peaks are therefore located at the bottom right of this diagram.

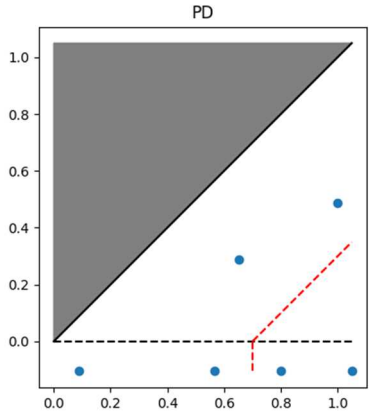

Fig. 5. Persistence Diagram

The persistence threshold merges each cluster of prominence lower than τ into its parent cluster. This merging is

done in the hierarchy order defined by the persistence value, computed on the fly in the mode-seeking phase. In the figure bellow (*Fig. 6*), the two clusters (orange and blue) correspond to the two peaks of prominence selected in the PD shown above.

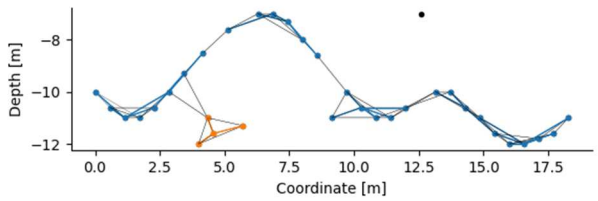

Fig. 6. Clusters Merging

III. Case Study in Real Bathymetric Datasets

A. Outliers' indentification

The identification of outliers is done by eliminating the soundings not belonging to the seabed cluster. Soundings that are not merged, and whose topological persistence is less than the persistence threshold τ, are called topological noise and correspond to outliers. One can also obtain dense clusters of point errors, thus having high persistence. It is then necessary to check the nature of these clusters to ensure that they are indeed outliers, which is immediate by visualizing the colored soundings (one color for each final cluster). On the previous figure (*Fig. 6*), the orange cluster formed during the mode-seeking phase is a cluster of outliers and should therefore be classified as such. Outliers are represented by red crosses in the following figure (*Fig. 7*).

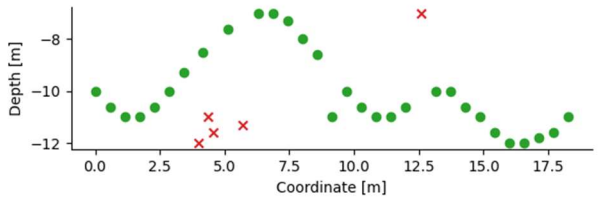

Fig. 7. Outliers' identification

B. Datasets

Experiments have been conducted on four datasets representing a wide variety of seafloor. The main characteristics of these four datasets are listed below:

- **Flat Bottom**: this survey was conducted on the 15th of November 2008, off the coast of Cadiz, by French research vessel Beautemps-Beaupré equipped with Kongsberg EM120 multibeam echosounder. The selected line contains 66,660 soundings. The bottom is flat, and outliers are observed mainly at nadir and at the edge of the swath.
- **Wreck**: this wreck is located off the point of Saint-Matthieu in western Brittany. This survey was carried out with the Kongsberg EM710 multibeam echosounder and the selected line contains 762,401 soundings.
- **Cliff**: this survey was conducted on the 14th of April 2000, near Minou lighthouse in Brest Bay, above a cliff. The sounder used is Kongsberg EM1002. The selected line contains 62,925 soundings.
- **Basse St Pierre**: this survey was conducted on the 27th of June 2011 by research vessel Panopée equipped with Teledyne RESON SeaBat 8125 multibeam echosounder. This survey above Basse St Pierre in Brest Bay is very noisy, with especially dense side lobes. The selected line contains 449,281 soundings.

C. Reference: manual processing

In this paper, we chose manual processing as reference. This processing is mainly subjective. It has a low sensitivity to point errors close to the background that the operator can relate to noise inherent to echosounder measurement. On the other hand, manual processing detects the most significant outliers and therefore those that most disturb the production of hydrographic products. It has a low false alarm rate.

We also note that the results of manual processing of the bathymetric data can vary from one operator to another. The table below illustrates the differences in the processing of multibeam data by two different hydrographers.

TABLE I. Results of the Manual Processing on Four Datasets Used

	Flat bottom	Wreck	Shelf	Basse St Pierre
Operator 1	1.40%	0.38%	0.68%	11.43%
Operator 2	0.71%	0.38%	0.65%	7.24%

D. Metrics

The results of the classification are given in the form of confusion matrices. "*Accepted*" and "*Rejected*" refer to the output of the manual processing.

We generally want to obtain a diagonal confusion matrix, i.e., a processing that is as close as possible to the reference. However, due to the characteristics of the manual processing previously detailed, a rather large number of false positives is expected.

TABLE II. Confusion Matrix

Number of soundings	Accepted pred	Rejected pred
Accepted	True Negative (TN)	False Positive (FP)
Rejected	False Negative (FN)	True Positive (TP)

In the field of statistical classification of data other metrics are also often used, for our study we focused on precision, sensitivity and specificity as presented below:

- $Precision = \frac{TP}{TP+FP}$

This indicator represents the proportion of relevant invalidations among all the soundings invalidated by the algorithm. It allows us to verify that we do not detect too

many soundings as invalid. The closer the precision is to 1, the less we invalidate excessively. On the contrary, the closer we are to 0, the less we generate FP compared to TP.

- $Sensibility = \frac{TP}{TP+FN}$

This indicator allows us to quantify the capacity of our algorithm to identify an invalid sounding as such. The closer we are to 1, the more our algorithm detect invalid soundings; on the contrary, the closer we are to 0, the more we generate FN compared to TP.

- $Specificity = \frac{TN}{FP+TN}$

This indicator is opposed to the sensitivity, it allows to quantify the capacity of our algorithm to identify a valid sounding as such. The closer we are to 1, the more we identify correctly valid soundings. On the contrary, the closer we are to 0, the more we generate FP compared to TN.

We also produce a 2D-map of the confusion matrix showing the location of TP in red, FP in orange, FN in blue, and TN in green. Indeed, the interest of combining global metrics, as presented previously, is to be able to locate with the 2D map precisely the errors and try to understand the origin of these wrong classifications (groups not statistically represented or outlier cluster too dense for example). In addition, it is important for navigation safety to ensure locally that an algorithm is working properly and has not been too destructive (e.g., deleting a high point of a wreck or trimming a slope break).

In this study we did not confront our results with the well-known CUBE algorithm [9]. Indeed, CUBE is an error-model based, direct digital terrain model (DTM) generator that doesn't classify the soundings between accepted and rejected. To allow an objective comparison it would be necessary to transform the classified soundings into a DTM which should be the most representative of the seabed (with an evaluation of the interpolation processes). This transformation has not been carried out here, but it would still be relevant to set up metrics in the future to compare with CUBE.

E. Results

For each of the four previously introduced datasets, the results are presented with a DTM, the confusion matrix and its 2-D representation and a brief analysis of the processing done by ToMATo.

1) Flat Bottom

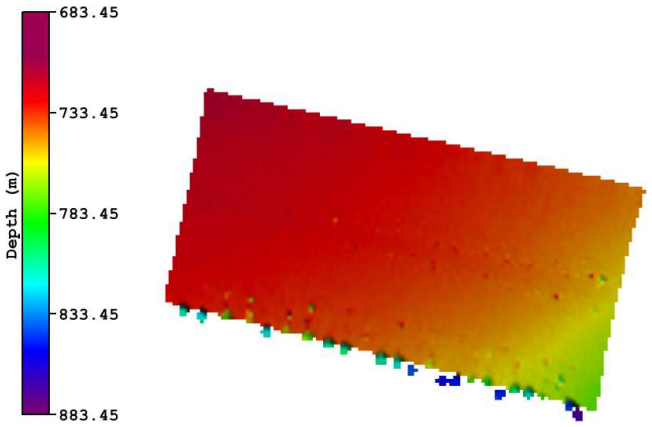

Fig. 8. DTM "Flat Bottom"

TABLE III. CONFUSION MATRIX "FLAT BOTTOM"

N_s = 66,660	Accepted pred	Rejected pred
Accepted	64,654 (96.99%)	1,066 (1.60%)
Rejected	87 (0.13%)	853 (1.28%)

ToMATo detects all outliers at the nadir and at the edge of the swath and is more sensitive to outliers than the manual processing. This dataset doesn't present any difficulties for the algorithm.

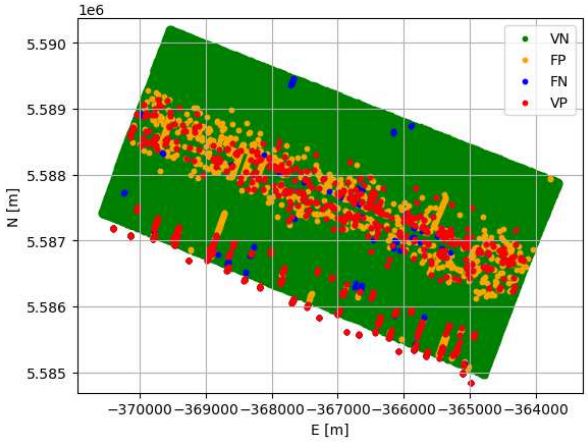

Fig. 9. 2D-map "Flat Bottom"

2) Wreck

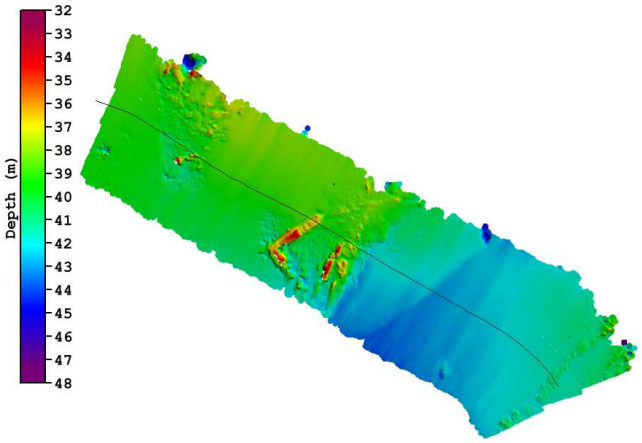

Fig. 10. DTM "Wreck"

TABLE IV. CONFUSION MATRIX "WRECK"

N_s = 762,401	Accepted pred	Rejected pred
Accepted	749,898 (98.36%)	9,378 (1.23%)
Rejected	762 (0.10%)	2,363 (0.31%)

ToMATo classifies lots of soundings as outliers near the wreck and rocky areas. This behavior is due to trade-off in the choice of the δ parameter: a too low δ makes the algorithm too sensitive when the variations of depth are important, while a too high δ makes the algorithm too insensitive when the bottom is flat.

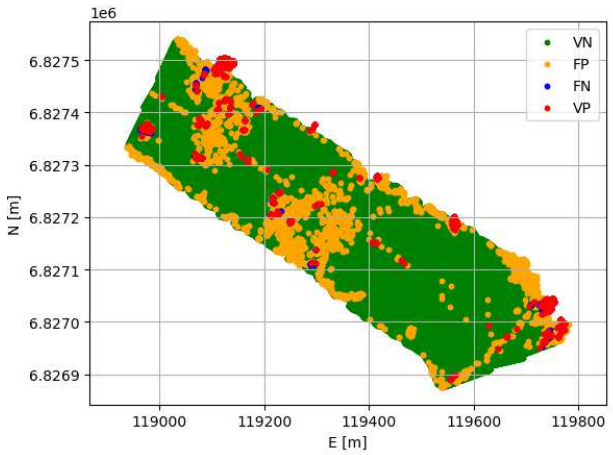

Fig. 11. 2D-map "Wreck"

3) Cliff

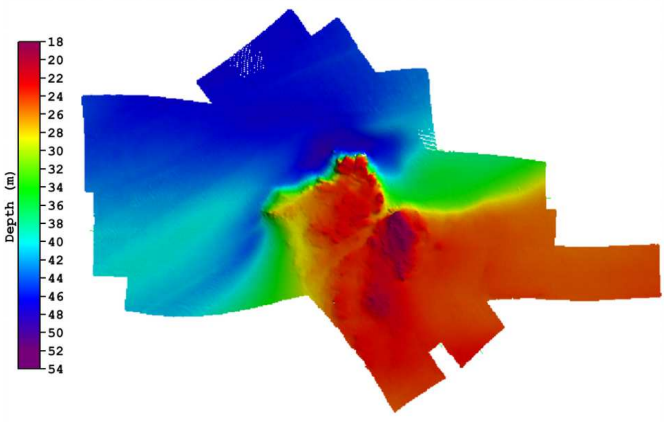

Fig. 12. DTM "Cliff"

TABLE V. CONFUSION MATRIX "CLIFF"

N_s = 62,925	Accepted pred	Rejected pred
Accepted	61,627 (97.94%)	866 (1.38%)
Rejected	326 (0.52%)	106 (0.17%)

High δ is required for the construction of the bottom cluster because of the cliff. Therefore, the ToMATo algorithm is less capable of detecting outliers on the shelf. Despite this important δ, lots of soundings on the cliff are identified as topological noise.

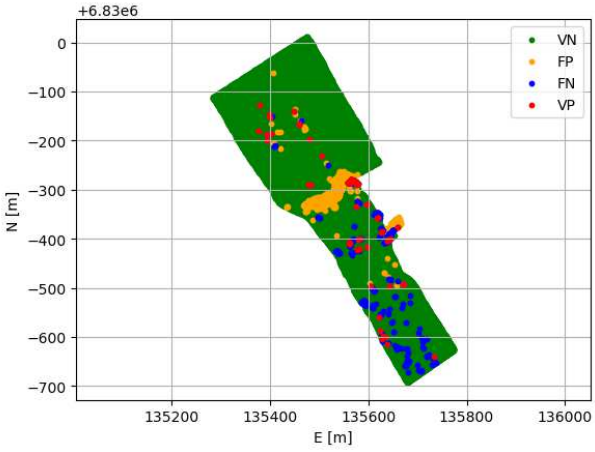

Fig. 13. 2D-map "Cliff"

4) Basse St Pierre

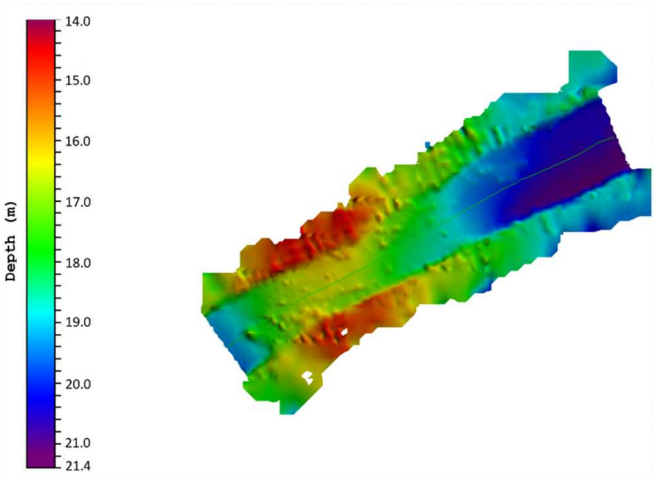

Fig. 14. DTM " Basse St Pierre "

TABLE VI. CONFUSION MATRIX " BASSE ST PIERRE "

N_s = 449,281	Accepted pred	Rejected pred
Accepted	352,732 (78.51%)	15,320 (3.41%)
Rejected	15,904 (3.54%)	65,325 (14.54%)

ToMATo is not affected by high density clusters of outliers on the two sides of the swath. In addition to eliminating these clusters, ToMATo remains highly sensitive to outliers on the sea bottom.

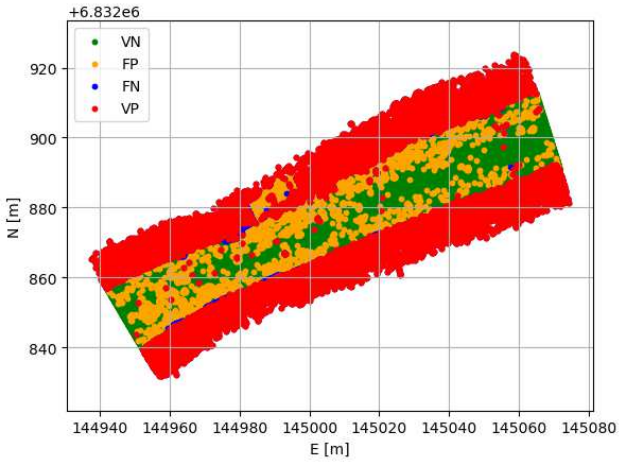

Fig. 15. 2D-map " Basse St Pierre "

F. Comparison with existing methods: DBSCAN & LOF

We compare the performances of the ToMATo method with those of two clustering algorithms already used for bathymetric data processing:

- The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was created in 1996 and relies on the density of clusters to perform partitioning [5].
- The Local Outlier Factor (LOF) algorithm was created in the 2000s and relies on the local density of observations [6]. If the density reveals a difference between the observed point and its neighbors, the point is considered as an anomaly.

These two algorithms have been implemented in python as part of previous work on the evaluation of algorithms for the identification of outliers in bathymetric data sets.

The following table presents the best parameters found for the three algorithms on each of the four data sets.

TABLE VII. PARAMETERS OF THE ALGORITHMS

	ToMATo	DBSCAN	LOF
Flat Bottom	$\delta = 4$ $\tau = 0.06$	$\varepsilon = 0.01$ $min_{samples} = 30$ $min_{cluster} = 50$	$N_{neighbors} = 8000$
Wreck	$\delta = 4.5$ $\tau = 0.07$	$\varepsilon = 0.1$ $min_{samples} = 2$ $min_{cluster} = 3$	$N_{neighbors} = 55$
Cliff	$\delta = 4$ $\tau = 0.08$	$\varepsilon = 0.1$ $min_{samples} = 10$ $min_{cluster} = 50$	$N_{neighbors} = 3000$
Basse St Pierre	$\delta = 3$ $\tau = 0.03$	$\varepsilon = 1$ $min_{samples} = 30$ $min_{cluster} = 50$	$N_{neighbors} = 12$

To compare the results of these three algorithms, we will used 3 common statistical indicators in statistical classification precision, sensibility and specificity as introduced in Section III-D.

TABLE VIII. COMPARISON WITH EXISTING METHODS

DTM	Algorithm	Precision	Sensibility	Specificity
Flat bottom	ToMATo	0.44	0.91	0.98
	DBSCAN	0.08	0.67	0.92
	LOF	0.25	0.93	0.96
Wreck	ToMATo	0.20	0.76	0.99
	DBSCAN	0.12	0.99	0.97
	LOF	0.07	0.95	0.95
Cliff	ToMATo	0.11	0.25	0.98
	DBSCAN	0.08	0.53	0.96
	LOF	0.19	0.43	0.99
Basse St Pierre	ToMATo	0.81	0.80	0.96
	DBSCAN	0.78	0.71	0.96
	LOF	0.96	0.56	1.00

First, it should be noted that the choice of parameters for the ToMATo algorithm is simpler than for the other two algorithms (Table VII). The standardization of the data results in having δ approximately equal to 4 for all datasets. The second parameter, τ, is selected using the persistence diagram. For the other two

algorithms, we observe that the parameters change considerably depending on the dataset processed.

Regarding the performance of the three algorithms, the selected statistical indicators do not allow us to clearly identify one algorithm as the best performing.

G. Application to multibeam extra detection

The extra-detection configuration allows the multibeam echosounder to acquire several soundings per beam, and thus in particular to study objects in the water column that backscatter less than the bottom.

This type of dataset defeats surface-oriented approaches as well as many statistical methods. Indeed, for surface-oriented approaches, the objects in the water column are necessarily far from the mathematical model representing the surface. They are therefore systematically identified as outliers. For statistical approaches, they identify outliers as the points belonging to the ends of the distribution tails [1]. If the objects in the water column are not simply removed, then statistical methods produce many choices (*hypothesis* in the CUBE framework [9]) that must be handled manually by a qualified operator, thus negating the value of automatic processing. The ToMATo method proposes a clustering of these objects. The topological persistence approach also allows to identify the most significant clusters, and thus to separate them from the noise.

This method has been tested on two datasets on which a mooring chain and submerged trees can be seen. On these two datasets, ToMATo was able to identify the sea bottom (in blue), the noise (in white) as well as the objects present in the water column (in color).

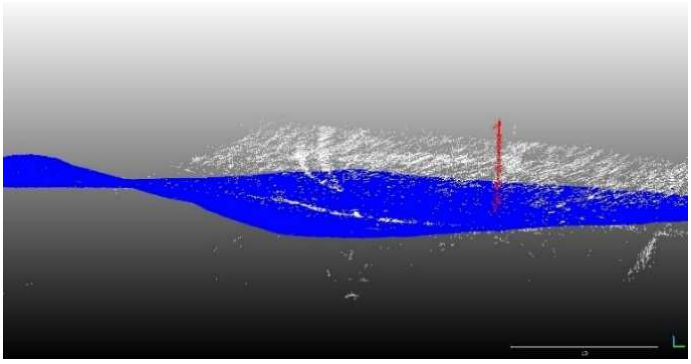

Fig. 16. "Mooring chain" – 5 October 2016, Brest Harbor, research vessel Panopée, Kongsberg EM2040C Extra-detection configuration

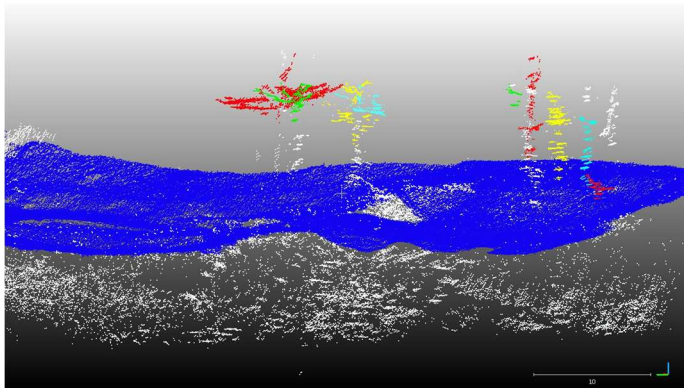

Fig. 17. "Underwater trees" – 18 October 2016, Guerlédan Lake, research vessel Panopée, Kongsberg EM2040C Extra-detection configuration.

IV. Conclusion and perspectives

The ToMATo method for processing multibeam bathymetric data yielded results comparable to other previously used clustering methods. Its topological persistence approach makes it robust to outlier clusters. Moreover, the standardization of the data makes its parameterization much easier than that of the other methods studied. On bathymetric data acquired in extra-detection, this method allows to identify efficiently the bottom and the objects present in the water column, reducing considerably the processing time of these data.

In a future study, it would be relevant to compare the ToMATo algorithm presented in this paper with another topological methodology [10] aimed at multibeam data cleaning, to examine the advantages and disadvantages of these two methods based on the same approach. It would also be interesting to compare the clustering performed by ToMATo on these extra-detection datasets to those obtained with DBSCAN or LOF. If these prove to be relevant, we could work on combining these three clustering methods to obtain a probability score, and thus reinforce the robustness of our processing. Finally, it should be noted that the ToMATo method is not limited to three dimensions. Other variables than depth can be included in the clustering process, such as intensity or quality factor.

Acknowledgment

Thanks to INRIA and especially to Steve Oudot for his advice in implementing the ToMATo algorithm in this bathymetric use case.

References

[1] J. Le Deunf, N. Debese, T. Schmitt et R. Billot, «A review of data cleaning approaches in a hydrographic framework with a focus on bathymetric multibeam echosounder datasets,» *Geosciences,* vol. 10, n° 7, p. 254, July 2020.

[2] IHO, *Standards for Hydrographic Surveys,* 6 éd., Monaco: International Hydrographic Bureau, 2008.

[3] N. Debese, Bathymétrie : Sondeurs, traitement des données, modèles numériques de terrain, Ellipses, 2013.

[4] F. Chazal, L. J. Guibas, S. Oudot et P. Skraba, «Persistence-Based Clustering in Riemannian Manifolds,» HAL, 2009.

[5] M. Ester, H.-P. Kriegel, J. Sander et X. Xu, «A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,» *AAAI,* pp. 226-231, 1996.

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng et J. Sander, «LOF: Identifying Density-Based Local Outliers,» chez *ACM SIGMOD Record*, 2000.

[7] H. Edelsbrunner, D. Letscher et A. Zomorodian, «Topological persistence and Simplification,» *Discrete & Computational Geometry,* n° 28, pp. 511-533, 2002.

[8] W. L. Koontz, P. M. Narendra et K. Fukunaga, «A graph-theoretic approach to nonparametric cluster analysis.,» *IEEE Trans. on Computers,* September 1976.

[9] B. Calder et L. Mayer, «Automatic processing of high-rate, high-density multibeam echosounder data,» *Geochemistry, Geophysics, Geosystems,* vol. 4, 2003.

[10] L. Arge, L. Larsen, T. Mølhave et F. van Walderveen, «Cleaning Massive Sonar Point Clouds,» *18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010*, San Jose, CA, USA, 2010.