Resolution Determination through Level of Aggregation Analysis

Brian R. Calder*

***Abstract*—In order to accommodate significantly varying depths within a survey area, and the consequent data density changes, variable-resolution depth modeling technologies are now being deployed. A core question for such technologies is how to determine the appropriate spatially-varying resolution at which to estimate or model the seafloor in a computationally efficient manner. Current methods include conversion from roughly-estimated depth or data density to resolution, or spatially-recursive sub-division (typically via a quadtree) with an appropriate similarity metric, typically working on a coarse-to-fine basis (i.e., starting with the whole survey area, and working to finer scales as the resolution is determined). All of these methods require a preliminary pass through the source data, and make various assumptions about its structure. Computational efficiency and level of assumptions are therefore important for implementation.**

As an alternative to these techniques, this paper describes a fine-to-coarse method based on a "level of aggregation" metric which makes no assumptions about the structure of the data, allowing it to be used equally on acoustic, lidar, or random point data. This method is methodologically direct and simple, data adaptive, readily parallelized, and automatically determines both the rate at which resolution is changing and the final resolution within this structure.

The method is illustrated in the context of processing Riegl VQ-880-G high-resolution shallow lidar data, and mixed-sensor acoustic data from a NOAA survey, with particular attention to parallel and distributed implementation. A direct corollary of estimating resolution is the ability to assess whether a given data set can meet survey specifications, which effectively provides a measure of how "surveyed" an area is. This is illustrated on an archive collection of random data from the U.S. Atlantic Margin in the context of Seabed 2030.

***Index Terms*—Resolution Determination, Bathymetric Processing, Hydrography, Bathymetric Modeling, Digital Elevation Model, Seabed 2030**

I. Introduction

FOR hydrographic data processing systems that estimate the depth in the survey area, rather than selecting soundings, resolution of representation is a critical issue. While a depth estimate may be created at any point of interest, and therefore at arbitrary locations across an area as required by the surveyor, for reasons of efficiency in computation and visualization, they are typically arranged in a more structured order. Early examples of processing systems of this kind (e.g., GMT [1] in the geophysical community, or CUBE [2] for hydrography) typically used fixed resolution grids with the resolution being chosen by the processor with little or no guidance as to what was appropriate beyond some intuitive feel for what was achievable given the type of survey system.

Resolution is not, however, arbitrary. The type and configuration of the survey instrument plays a heavy role: a 400-beam MBES with a 130–140° swath is liable to support higher resolution reconstructions of the true shape of the seafloor than one with 20 beams over the same opening angle, for example. But the grid also represents a sampled version of the seafloor and therefore must obey Shannon's theorem [3], or suffer from aliasing: it is not possible to accurately represent 1 m cubic features with a grid resolution of 5 m. Crucially for hydographic practice, arbitrary selection of an inappropriate resolution may compromise the algorithm's ability to reliably represent hydrographically significant detail.

Nor is the appropriate resolution fixed for a given survey area. Different instruments used in different regions of the survey area, or the same instrument used in different depths could cause the achievable resolution to change. For survey areas with significant depth range, therefore, any choice of a fixed resolution at which to work is necessarily a compromise: it will almost surely be over-estimating the achievable resolution in some areas (leading to poor, or sparse, estimates) and under-estimating in others (leading to spatial aliasing and poorly defined features).

For these reasons, second-generation computer-assisted hydrographic workflows (e.g., CHRT [4], or CHARM [5]) attempt to use the data itself to predict an appropriate resolution at which to work, and support either mixed, or truly variable, resolution across the survey area. Setting aside the question of how to support a variable resolution data structure, the principal question is how to reliably and efficiently determine the appropriate estimation resolution at any given point.

Previous approaches to this problem have typically relied on translating some summary statistic of the data into an estimate of resolution; depth [6], data density [4], and complexity [5] have all been used. These proxies, however, have implementation difficulties (for example, reliably estimating area ensonified [7] to provide for data density estimates), and are only approximations to the problem, which leads to the necessity for empirical calibration constants in making the translation. They are often also unsuited to match modern survey standard specifications, which often specify a minimum number of observations required at each estimation node; the most commonly used data density-based estimate typically predicts the mean number of observations, rather than the minimum unless corrections are made, which may be dataset specific.

Addressing the same problem, this paper proposes an alternative method of determining resolution that minimizes the

*Corresponding author. Center for Coastal and Ocean Mapping & NOAA-UNH Joint Hydrogaphic Center (CCOM/JHC), University of New Hampshire, Durham NH 03824, USA. T: +1 603 862 0526. E: brc@ccom.unh.edu. ORCID: 0000-0001-9871-7824.

number of assumptions made about the data, is readily scalable, and which matches survey specifications for a minimum observation count. Starting with a simple count of observations in a high-resolution grid across the survey area, the algorithm works fine-to-coarse, aggregating at each point a sufficient number of high-resolution cells until the minimum number of observations required is achieved. Statistical analysis of this "level of aggregation" (LOA) allows the algorithm to determine the maximum size of an analysis box that is required to satisfy the lowest resolution that is mandated by the data; within each cell of a grid at this resolution, analysis of the LOA estimates then allows the algorithm to determine the final resolution of estimation at each point. This structure provides for several benefits over previous algorithms. First, since the algorithm is based purely on the count of observations, it does not rely on structure of the data and can be applied to MBES lidar, or even random point data. The structure of the estimation allows for parallel implementation of the problem, letting the algorithm to scale effectively without communications overheads. Finally, since the algorithm directly counts the observations achieved at each high-resolution grid cell, it is significantly better at achieving the minimum observation count to mandate reliable depth estimation.

This algorithm is illustrated on a number of datasets, including high-resolution topobathymetric lidar, MBES hydrographic data, and deep-water mixed random-point and MBES data in the context of Seabed 2030 [8]. These datasets illustrate the basic workings of the algorithm, the mechanisms to efficiently solve the problem at scale, and an alternative view on how to determine the level of survey completeness in a given area, respectively.

II. Methods

A. Level of Aggregation

Consider a fixed, high-resolution, zero initialized, grid imposed over the survey area, $C(i,j) \in \mathbb{Z}_{\geq 0},\ 0 \leq i < N, 0 \leq j < M$ with linear mapping to the projected coordinate space with $(x_i, y_j) = (Ri, Rj) + (x_0, y_0)$ for given origin point (x_0, y_0) and fixed resolution R. Each cell (i,j) therefore covers $\mathsf{A}(i,j) = [x_i, x_{i+1}) \times [y_j, y_{j+1}) \subset \mathbb{R}^2$; we assume that details of automatically extending this array to cover all of the data read are abstracted from this description, and that $C(i,j) \equiv 0$ outside of the given dimensions.

For each data point $\vec{s} = (x_s, y_s, z_s),\ 0 \leq s < S$, the grid of counts at $(i_s, j_s) = (\lfloor x_s/R \rfloor, \lfloor y_s/R \rfloor)$ is updated as

$$C(i_s, j_s) \leftarrow C(i_s, j_s) + 1 \tag{1}$$

(i.e., a simple accumulator). The level of aggregation (LOA) grid $L(i,j) \in \mathbb{Z}_{\geq 0},\ 0 \leq i < N, 0 \leq j < M$ is then defined as

$$L(i,j) = \min \left\{ \lambda : \sum_{c=i}^{i+\lambda} \sum_{r=j}^{j+\lambda} C(c,r) \geq n_{\text{req}} \right\},\ \lambda \in \mathbb{Z}_{\geq 0} \tag{2}$$

where n_{req} is the user's required minimum number of observations for stable depth estimation (which should include some allowance for blunders in the input data). The LOA computation is illustrated in Figure 1.

B. Refinement Estimation

The LOA grid $L(i,j)$ estimates directly the distance at each node that the algorithm must consider in order to accumulate enough observations to reliably estimate depth. It therefore directly estimates the grid resolution that can be supported by the data at each point. The spatial variation in the estimate, illustrated in Figure 2 for a lidar dataset, depends on the instrument, survey pattern, and bathymetry.

In order to provide for a computationally efficient data structure that supports variable resolution, the model of CHRT is adopted [4]. This requires that the code establishes a grid of appropriate fixed resolution, W, across the survey area which forms the basis for varying resolution: within each fixed W-resolution cell, the appropriate depth estimation resolution, $\Delta x(c,r)$, for the data is determined, and a refined grid at that resolution is constructed to fill the cell. In this way, the depth estimate resolution can be spatially variable (every W meters) while still maintaining what is essentially a collection of fixed resolution grids nested within the initial W-resolution grid cells. This structure is called a piecewise-constant sample spacing (PCSS) grid.

Instead of a user-defined W, the LOA analysis allows the spacing to be estimated directly. Consider the probability mass function (pmf) for the LOA of the survey area, $p(\lambda) \in \mathbb{R}, 0 \leq \lambda < \infty$, illustrated in Figure 3. W must be as small as possible so that $\Delta x(c,r)$ can adapt as quickly as possible to changing conditions, but must also be no less than the lowest resolution required by the data, since it is also the minimum refined resolution possible (i.e., the refinement grid for each cell must fit entirely within the cell). A reasonable choice for W is therefore the upper α-percentile (with $\alpha \approx 0.95 - 0.99$) of the probability mass function, $W = q_\alpha R$, where

$$q_\alpha = \min \left\{ x : \sum_{\lambda=0}^{x} p(\lambda) \geq \alpha \right\} \tag{3}$$

since this will ensure the maximum $\Delta x(c,r)$ for any part of the survey area will be accommodated while avoiding any estimation outliers from the LOA.

Once W is determined, the same techniques can be applied to the fixed, W-resolution, grid to determine a value for $\Delta x(c,r)$ in each cell. Let each cell (c,r) of this grid cover area $\mathsf{A}(c,r) = [x_0 + Wc, x_0 + W(c+1)) \times [y_0 + Wr, y_0 + W(r+1)) \subset \mathbb{R}^2$ in projected coordinates. The cells in $L(i,j)$ that intersect are the neighborhood of (c,r),

$$\mathsf{N}(c,r) = \{(i,j) : \mathsf{A}(i,j) \cap \mathsf{A}(c,r) \neq \emptyset\}. \tag{4}$$

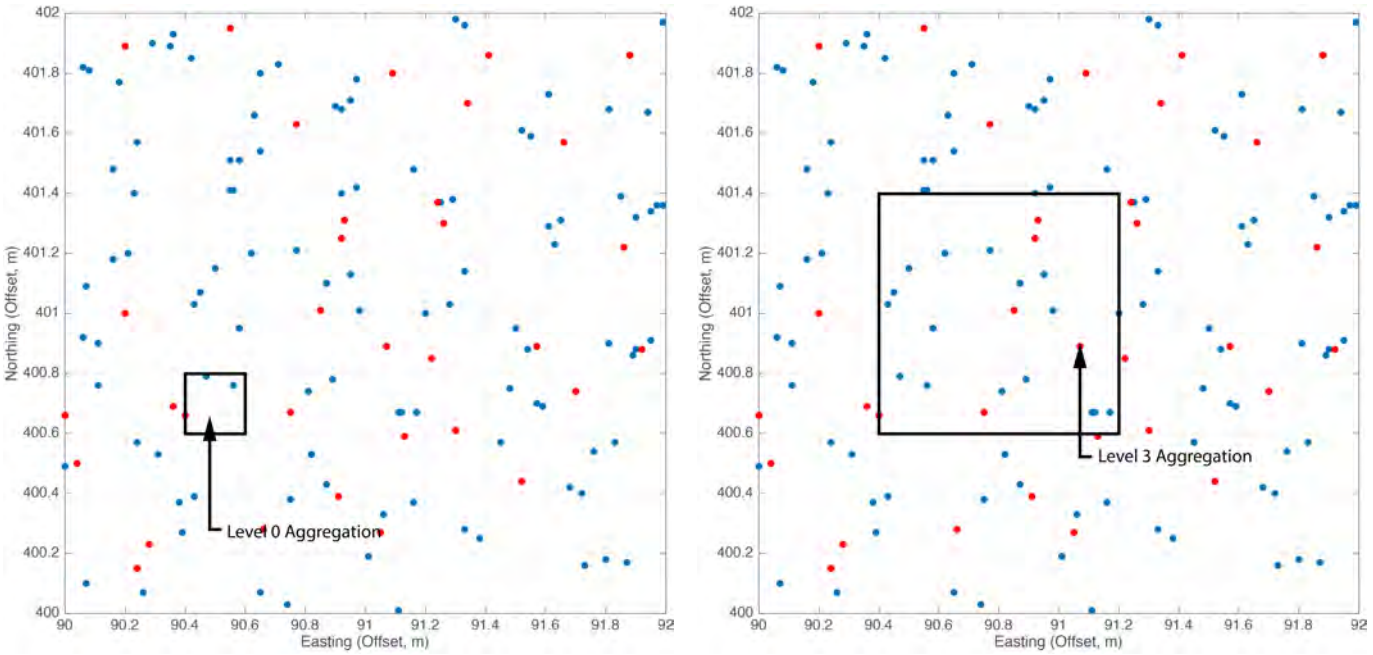
To determine $\Delta x(c,r)$, form the pmf of $L(i,j)$, $(i,j) \in \mathsf{N}(c,r)$ and then find the α-percentile for this pmf as before. As example refinement grid is illustrated in Figure 4.

III. Implementation

A. Level of Aggregation

Computing the LOA at each cell by evaluating Eqn. 2 would be extremely time consuming. Constructing the auxiliary function

$$\rho_\lambda(i,j) = L_\lambda(i,j) - n_{\text{req}} \tag{5}$$



(a) Base LOA (level 0)

(b) Incremented LOA (level 3)

Fig. 1. Estimation of Level of Aggregation. The LOA increases, taking in more and more soundings, until the number of soundings encompassed exceeds the user's minimum count (including any allowance for points expected to be lost to blunders if the computation is based on raw, rather than edited, data).

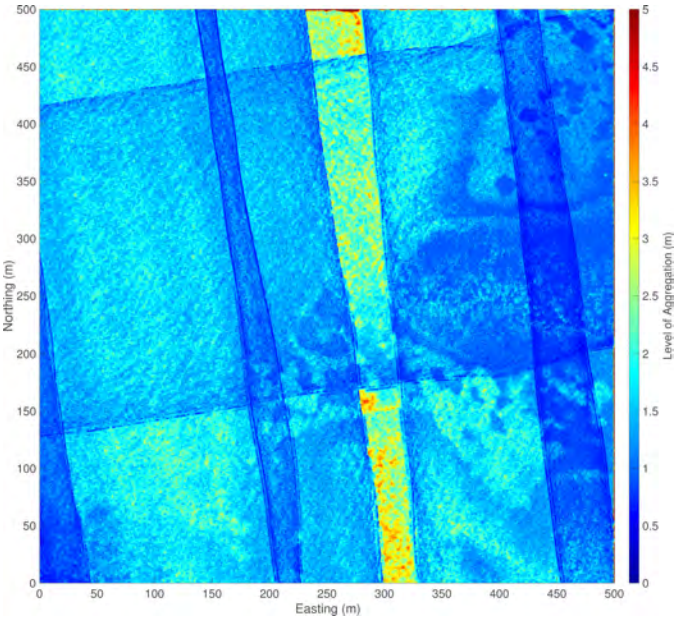


Fig. 2. Example of the LOA values, converted to meters using $R = 0.125\,\mathrm{m}$ and $n_{\mathrm{req}} = 5$, associated with a $500\,\mathrm{m} \times 500\,\mathrm{m}$ section of a lidar dataset from a Reigl VQ-880-G topobathymetric lidar in Key West, FL. Data courtesy of NOAA RSD.

demonstrates that this is essentially a root-finding problem, since the preferred solution is the value of λ just larger than the (single, because $L_\lambda(i,j)$ is monotone increasing) root of $\rho(i,j)$, Figure 5. Any appropriate root-finding algorithm may then be applied, but since gradients of $\rho(i,j)$ are not known, bisection search [9] is used in the example implementation. As a practical matter, care must be taken to select the next bisection point where the right-hand end of the solution range is a valid LOA estimate. In a simple implementation, the next point selected would tend to the right-hand side and therefore never converge. A number of *ad hoc* solutions are possible, but moving the right-hand endpoint to the left by a fixed amount and testing again for solution was found to be most efficient in practice. There is also some efficiency to be had by using a near, previous, solution (e.g., $L(i-1,j)$ or $L(i,j-1)$, as appropriate) as a starting point, since there is significant correlation between solutions.

The cost of evaluating $L_\lambda(i,j)$ can be ameliorated by recognizing that this can be implemented through a Summed Area Table [10] (SAT), which can be evaluated with complexity $O(1)$ rather than $O(N^2)$. The cost of constructing the SAT can be reduced by recognizing that each $L(i,j)$ is independent of all others given the counts table $C(i,j)$, and therefore that the computation can readily be done in parallel. For example, in a multi-threaded environment, each row $L(\cdot,j)$ could be handed to a different thread in turn (although for efficiency, a job quantum of a few rows is more useful).

B. Refinement Resolutions

As with the LOA computation, the refinement spacing W and depth estimate resolution $\Delta x(c,r)$ are independent given the $L(i,j)$, and therefore can be computed in parallel as before. There is a small efficiency to be had by accumulating the pmf of $L(i,j)$ as each cell is being evaluated in parallel, rather than doing so as a separate step. Two practical complications, however, occur.

First, the counts $C(i,j)$ are therefore the LOA $L(i,j)$ are spatially quantized at scale R. The pmf estimated to determine W and $\Delta x(c,r)$ is therefore also quantized [11] and can

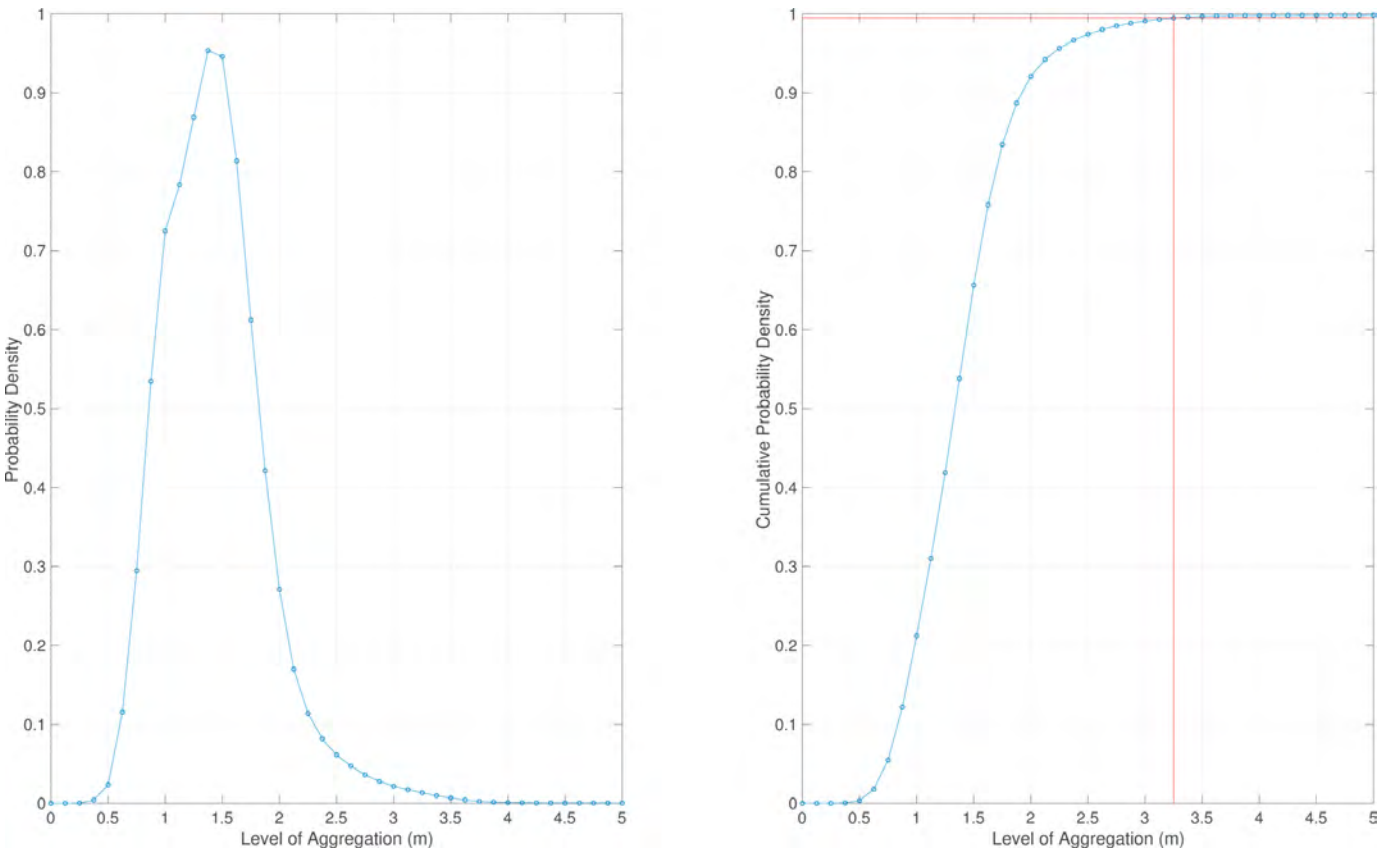


Fig. 3. Probability mass function for the lidar LOA shown in Figure 2, converted to meters using $R = 0.125\,\mathrm{m}$. Finding the 99% centile for the distribution can be used to set W, the analysis window width; analysis of the LOA within each W-resolution cell then sets the depth estimate resolution, $\Delta x(c, r)$.

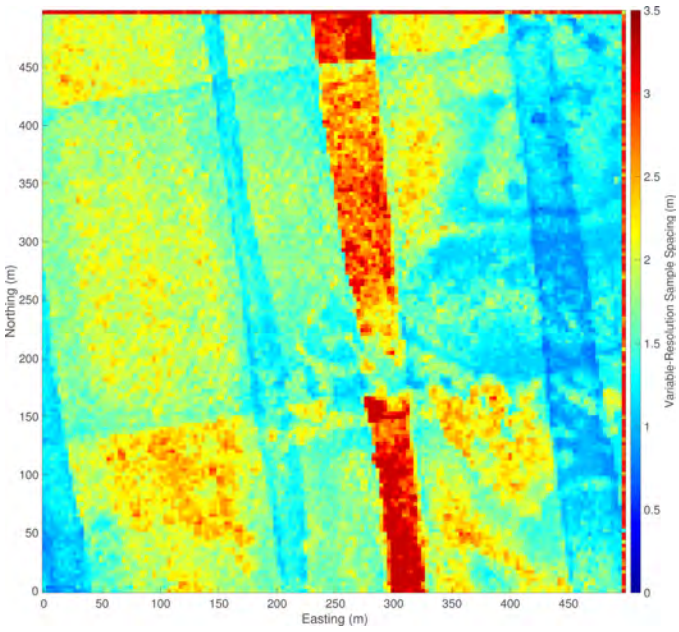


Fig. 4. Computed depth estimation resolutions for the lidar LOA shown in Figure 2.

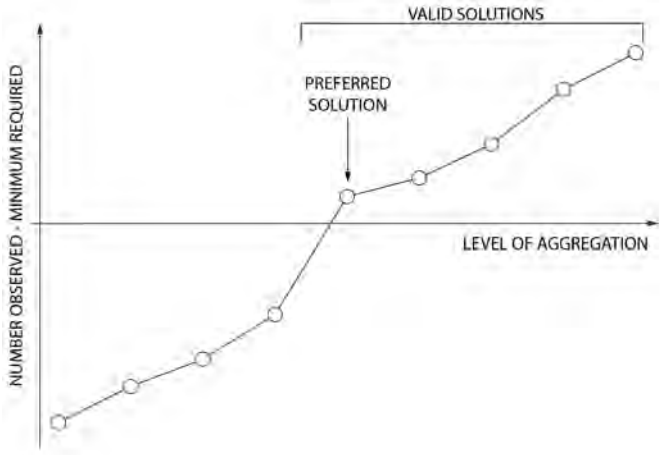


Fig. 5. Cartoon illustrating the root-finding solution for LOA determination, the range of valid solutions (i.e., meet or exceed the number of observations required), and the preferred solution, which is always immediately to the right of the root location.

suffer from aliasing. Choosing a sufficiently small R (e.g., one quarter of the smallest depth estimate spacing expected) resolves this issue.

Second, due to the quantization of $L(i, j)$, the value of q_α is likewise quantized, and it is difficult to determine a good estimate for W or $\Delta x(c, r)$. Interpolation of the pmf can be

used to resolve this; in theory, sinc interpolation is optimal, but in practice linear interpolation was found to be sufficient.

C. Large-Scale Structure

The preceding description assumes that the LOA for the entire survey areas if computed as a single entity, and that a global value for W is determined. While this is possible, it may not be optimal for regions with extreme depth range (e.g., from shoreline to mean ocean depth, as sometimes happens around volcanic islands). In such areas, the value of W determined would reflect the sparsest data available, and therefore would likely be relatively large. This has the consequence of reducing the rate at which resolution can be changed, which can be quite important in shallow areas, or on steep slopes.

A simple work-around for this is to break the survey area into smaller tiles, and determine a value of W for each tile. This also allows for spatial adaptation of R to reduce the effects of quantization in $C(i, j)$ described previously. Smaller tiles can also assist in active memory management [12].

IV. Applications

A. General Depth Estimation

The use of LOA to determine depth estimation resolution in lidar data has already been demonstrated in Figures 2–4 during development of the methods. Since the method relies solely on the observation counts, however, it is equally appropriate for MBES surveys, or mixed source surveys (SBES, MBES, and lidar), as illustrated in Figures 6–8. Note that obtaining stable resolution estimates such as these with prior methods [4] is quite difficult, and requires many special cases and approximations; here, it drops immediately from the computation, a significant advantage.

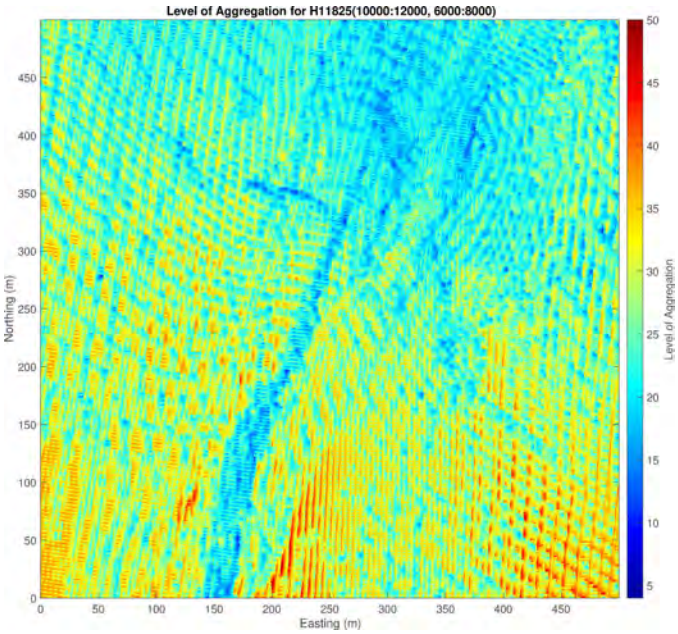


Fig. 6. Example LOA values from an intermediate depth region of MBES data in Ernest Sound, AK. Note the different patterns with respect to Figure 2 due to the differences between MBES and lidar instruments. Data courtesy of NOAA HSD.

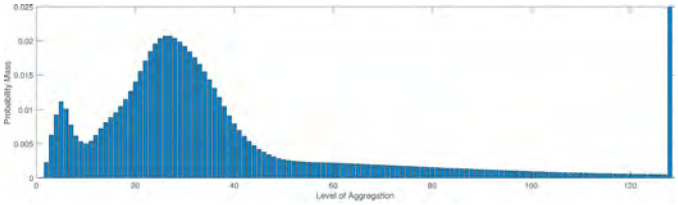


Fig. 7. Probability mass function for the LOA estimates in the whole dataset of Figure 6. The distinctly different pattern with respect to Figure 3 is due to the significant depth range in this dataset.

B. Survey Completeness

Although primarily intended to estimate supportable depth estimation resolutions in a spatially-varying manner, the LOA argument may also be reversed to address survey completeness. That is, given a survey specification for a depth resolution to be achieved, potentially as a function of depth, evaluation of $\Delta x(c, r)$ determines whether there is sufficient data available in the area to meet the specification. Practically, the survey would continue (with the estimates being updated as new data is added) until this was the case everywhere in the area of interest, or until it was demonstrated that adding more data had no material effect.

This method can also be used to assess data gaps. Consider, for example, the Seabed 2030 initiative [8], which aims to generate high-resolution maps of the whole world ocean by 2030. Recognizing that achievable resolution of gridded representation is a function of depth, Seabed 2030 requires a resolution of 100 m above 1,500 m depth, 200 m from 1,500–3,000 m, 400 m from 3,000–5,750 m, and 800m below 5,750 m.

Availability of data to support this can be readily addressed using the LOA analysis. Figure 9 shows the data inventory held at the National Centers for Environmental Information (NCEI) bathymetric databases for the region in an eight-degree area around Bermuda. While compelling, visualizations such as these are often misleading, since even a singlebeam trackline of passage soundings must be drawn as at least one pixel in order to be visible—which may be a significant size if the visualization is zoomed out to sufficiently small scale. This tends to make it look like there is more coverage of data than there really is. The estimated achievable resolution, Figure 10, at $W = 3,125$ m demonstrates readily just how much of the area is effectively unsurveyed. Using the depth estimated by the CHRT algorithm for the area, Figure 11, it is then possible to determine the Seabed 2030 required resolution, and thence for which areas of the region the database contains sufficient data to be considered complete, Figure 12.

Note that this estimate is intentionally pessimistic. The algorithm here has been configured for $n_{\text{min}} = 5$, with the assumption that 20% of the raw data used for the analysis will be blunders, which increases the total observation count required proportionately. This is significantly more than Seabed 2030 requires, but is not unreasonable when dealing with raw acoustic echosounding data (a formal assessment of data blunder levels in this case was not conducted). What this

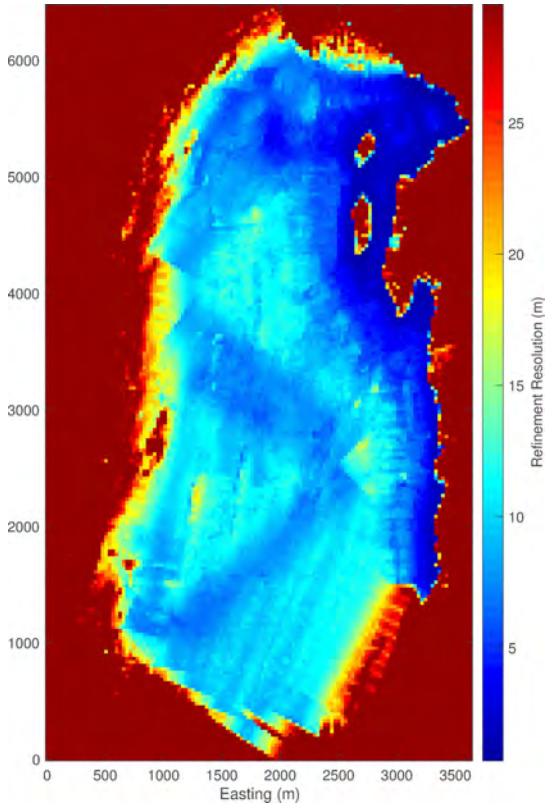


Fig. 8. Depth estimation refinement resolution computed by the algorithm for the LOA estimates in Figure 6. This is raw data, hence the "noise" around the edges; this method compares very well to previous best-in-breed estimates.

example demonstrates, however, is that even MBES data might not be sufficient for an area to be considered complete in the modern sense. For example, there are a number of cases where MBES tracklines are partially meeting the requirement, and even some where none of the data meets the current requirements (see, for example, the northwest corner of the area in Figure 12). This MBES data is a legacy of first-generation commercial deep-water MBES systems from the 1980s, where there were many fewer beams across the swath than is now typical, and therefore lower data density to support higher resolution depth estimation. Significantly, none of the areas covered by SBES lines are considered complete under these assumptions (although sparse depth estimates are still generated because the algorithm will construct a depth given even a single observation).

Finally, observe that the LOA solution, since it is computed from the count grid, can be readily updated as new data is added to the database, so long as the count grid is preserved. For scale, although allowing that computation time will vary significantly with implementation and hardware, the full refinement computation (post count-grid construction) in the example implementation was approximately 17.5 s (based on an Intel Core i7 processor, running at 4 GHz, with 32 GB memory, and an SSD-fronted hard disc), for Figure 10, with dimensions $N = 35,623$ and $M = 41,864$. Clearly, the algorithm can be implemented with low overhead (the time required to read the raw data in order to create the count grid is significantly longer than the computation time), which makes re-running the refinement resolution computation occasionally a low-cost event.

V. Summary

Careful selection of the resolution of representation is essential for a DTM to faithfully model the source data. Due to changes in depth, or instrument, however, this resolution is not necessarily constant across a survey area, and choosing a single resolution runs the risk of over- and under-sampling the surface in different areas. Either option has unfortunate repercussions.

This paper outlined a new method for estimating the appropriate resolution at which to process raw survey data, which has a number of advantages over other proposed methods. In particular, it directly assesses the stability of depth estimation in order to predict resolution, it automatically calibrates the resulting data structure to the data's behavior, it makes few assumptions about the structure of the raw data and can therefore be used for almost any sounding source, and it scales readily to a parallel implementation for efficiency.

In addition to its primary application in determining the resolution at which the data supports depth estimation, the algorithm can also be used to investigate whether the data so far collected meets the (possibly depth dependent) survey specification for resolution, and therefore can be used to assess completeness of surveys.

VI. Acknowledgements

This research was supported by NOAA grant NA15NOS-4000200. I am grateful to NOAA Remote Sensing Division, Hydrographic Surveys Division, and the National Centers for Environmental Infomation for provision of the data used in this paper.

References

[1] W. H. F. Smith and P. Wessel, "Gridding with continuous curvature splines in tension," *Geophysics*, vol. 55, no. 3, pp. 293–305, 1990.

[2] B. R. Calder and L. A. Mayer, "Automatic processing of high-rate, high-density multibeam echosounder data," *Geochem., Geophys. and Geosystems (G3) DID 10.1029/2002GC000486*, vol. 4, no. 6, 2003.

[3] J. G. Proakis and D. K. Manolakis, *Digital Signal Processing*, 4th ed. Prentice Hall, 2006.

[4] B. R. Calder and G. Rice, "Computationally efficient variable resoution depth estimation," *Computers and Geosciences*, vol. 106, pp. 49–59, 2017.

[5] N. Debese, R. Moitié, and N. Seube, "Multibeam echosounder data cleaning through a hierarchic adapative and robust local surfacing," *Computers and Geosciences*, vol. 46, pp. 330–339, 2012.

[6] T. CARIS, *HIPS and SIPS User Manual 2019*, 115 Waggoners Lane, Fredericton NB E3B 2L4, Canada.

[7] B. R. Calder and G. Rice, "Design and implementation of an extensible variable resolution bathymetric estimator," in *Proc. US Hydro. Conf.* Hydro. Soc. Am., April 2011.

[8] M. Jakobsson, G. Allen, S. Carbotte, R. Falconer, V. Ferrini, K. Marks, L. Mayer, M. Rovere, T. Schmitt, P. Weatherall, and R. Wigley, "Nippon Foundation-GEBCO Seabed 2030 Roadmap for Future Ocean Floor Mapping," Nippon Foundation-GEBCO, https://seabed2030.gebco.net/documents/seabed_2030_roadmap_v10_-low.pdf, Tech. Rep., June 2017.

[9] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes*, 3rd ed. Cambridge University Press, 2007.

[10] F. C. Crow, "Summed-area tables for texture mapping," *Proc. ACM SIGGRAPH*, vol. 18, no. 3, pp. 207–212, 1984.

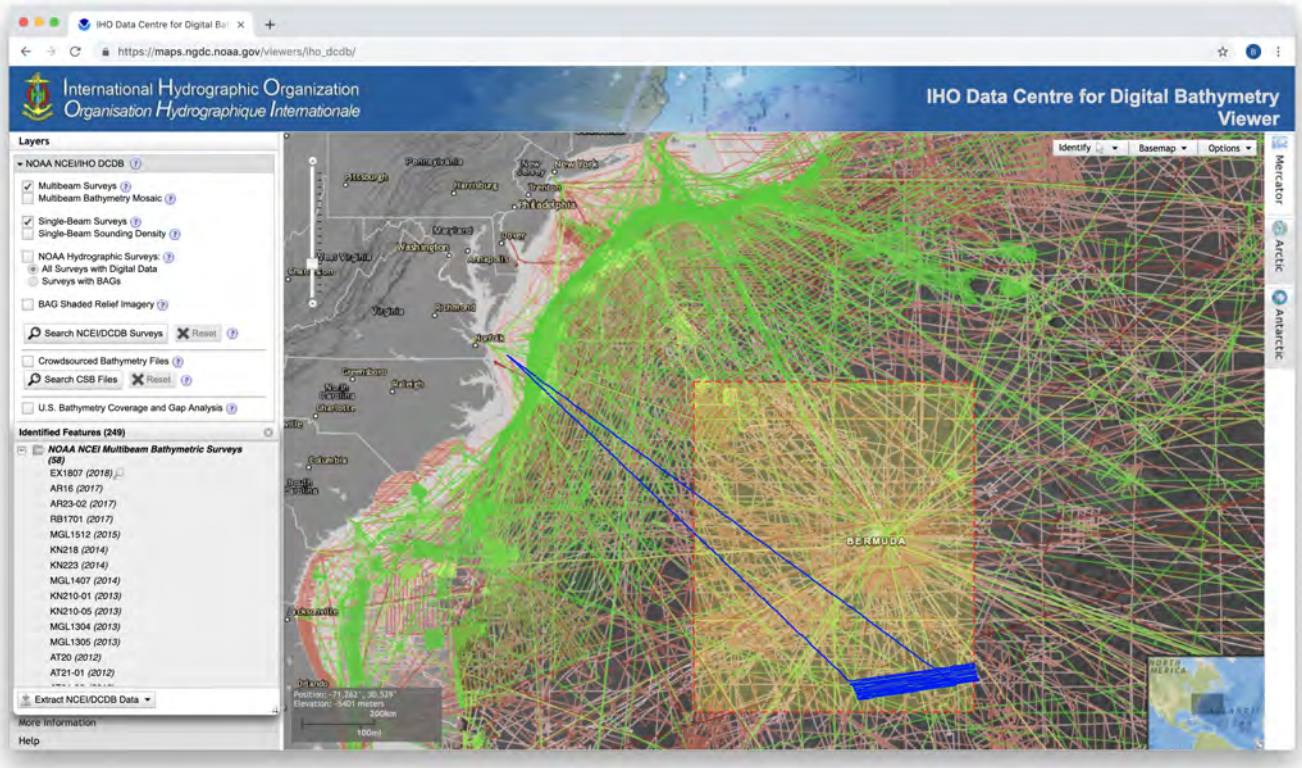


Fig. 9. Visualization of the data holdings at NCEI for an eight-degree (longitude and latitude) box approximately centered on Bermuda (62-70W, 28-36N).

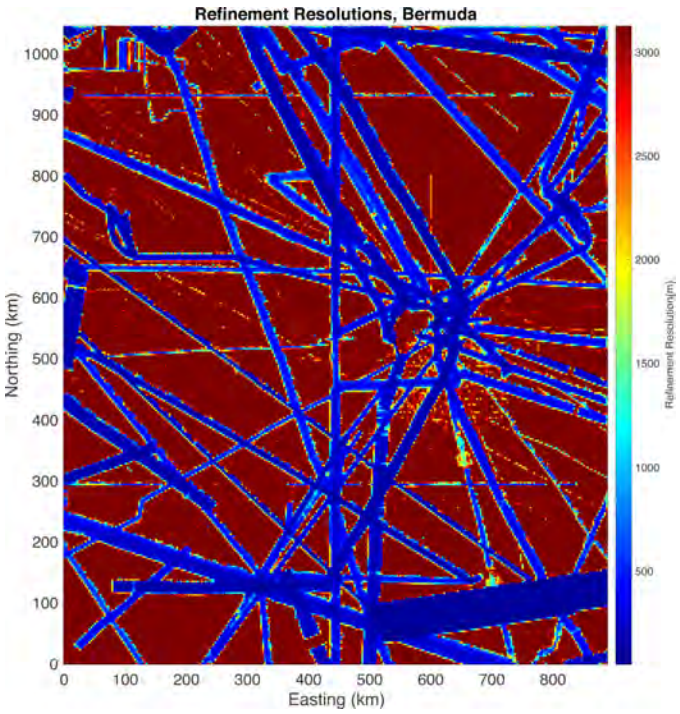


Fig. 10. Estimated achievable stable depth estimation resolution for Figure 9. Note that much of the area is set to 3,125 m, indicating that there is no data to support any depth estimate with the number of required observations ($b_{\text{req}} = 5$).

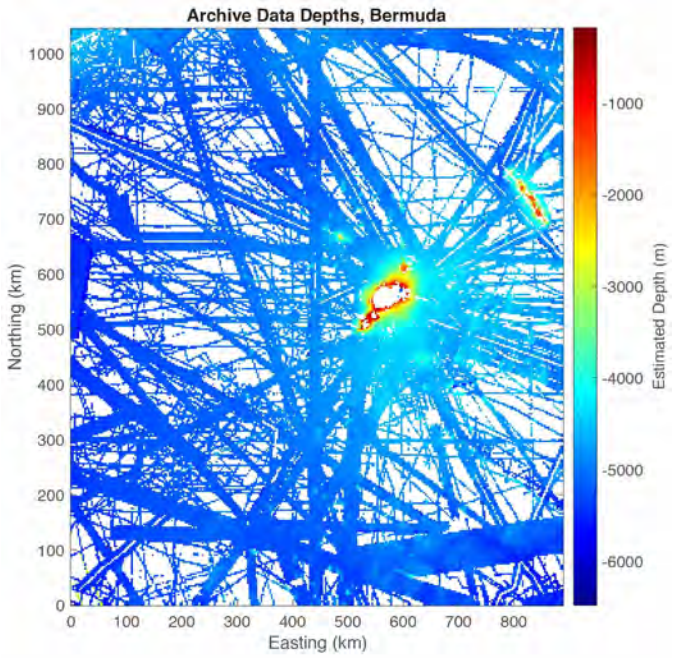


Fig. 11. Depth estimated from raw observations in the area of Figure 9. This is the primary CHRT reconstructed depth within each W-resolution cell over the area.

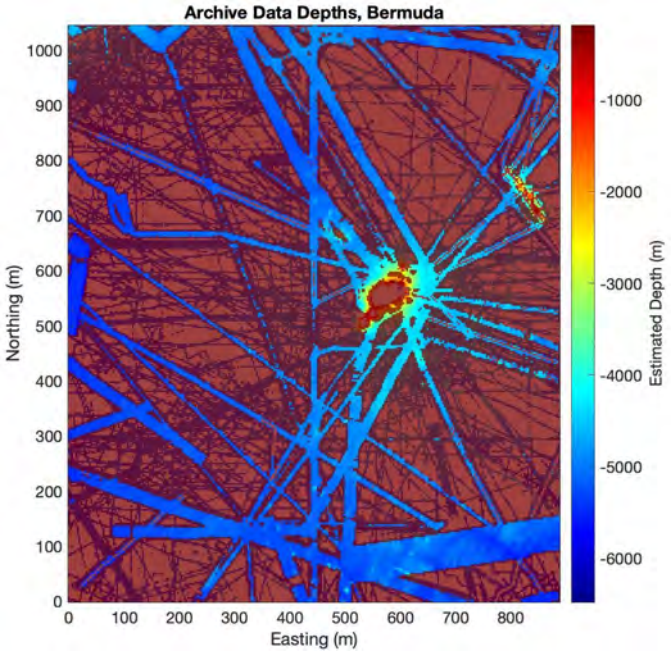


Fig. 12. Visualization of the area considered "complete" based on $n_{\text{req}} = 5$ observations and Seabed 2030 resolution requirements. Areas with transparent red tint (compare Figure 11) are incomplete.

[11] B. Widrow, "Statistical analysis of amplitude quantized sampled-data systems," in *Proc. AIEE Fall General Meeting*, no. 60-1240. AIEE, 1960.

[12] A. A. Yıldırım, D. Watson, D. Tarboton, and R. M. Wallace, "A virtual tile approach to raster-based calculations of large digital elevation models in a shared-memory system," *Computers and Geosciences*, vol. 82, pp. 78–88, 2015.