Stanford Exploration Project, Report 110, September 18, 2001, pages 1–132

Estimation of systematic errors in tracked datasets using least squares crossing point analysis

Morgan Brown[1]

ABSTRACT

Geophysical data often contains both systematic and random errors. If left unchecked, the systematic errors can cause acquisition footprint in the final map. I present a method to estimate systematic error by analyzing measurements at points where two acquisition swaths cross. I then subtract the estimated systematic error from the data and generate a map with a familiar least squares formulation. I test the method on bathymetric data from the Sea of Galilee. Compared to two previous least squares formulations, my new method produces final maps which are relatively free of acquisition footprint, and which exhibit preservation of underlying bathymetric features.

INTRODUCTION

Simple tracked datasets, consisting of single-channel measurements (z) acquired by a moving instrument at surface points (x, y), have traditionally proven valuable test-beds for least squares estimation techniques at SEP, due to their small size and conceptual simplicity. Ben-Avraham et al.'s (1990) bathymetric survey of the Sea of Galilee (Lake Kinneret) has drawn a prolonged interest (Claerbout, 1999; Fomel and Claerbout, 1995; Fomel, 2001), primarily because errors in the data seriously inhibit the task of translating the data into a gridded map.

In practice, each point measurement contains random errors, due both to instrumental inaccuracy and to physical phenomena with time and spatial scales smaller than two neighboring measurements. Unfortunately, these "random" errors are often non-gaussian, violating a central assumption of estimation theory that data contain gaussian-distributed error. In the context of Galilee, the ship's pre-GPS radio location system often mis-positioned depth soundings on the earth's surface. The erroneous measurements are easily identified as spikes in locations where the true sea floor is nearly flat, but not where the sea floor dips steeply. Even more crucially, however, these data also contain systematic error, which I define as error which varies slowly–in time and space–over a single track of measurements. Causes may include tidal shifts, instrumental drift, and wind-induced bulging of the lake's surface.

(Claerbout, 1999) cast the translation of irregular point data into a regular gridded map as an "inverse interpolation" problem. The simplest 2-D formulation of this problem is sensitive both to non-gaussian and systematic errors as noted by Fomel and Claerbout (1995). To

[1]**email:** morgan@sep.stanford.edu

overcome both difficulties, they included a composite residual weighting operator consisting of: 1) diagonal weight estimated via Iteratively Reweighted Least Squares (IRLS) (Nichols, 1994; Guitton, 2000) to handle non-gaussian noise, and 2) a finite-difference first derivative operator to suppress correlated components of the residual. A similar technique was applied successfully to process a Geosat dataset with similar errors by Ecker and Berlioux (1995).

While Fomel and Claerbout's (1995) approach suppresses both acquisition footprint and non-gaussian noise, the authors note a loss of resolution in the final map of Galilee, relative to the simple inverse interpolation result. As noted by Claerbout (1999), when the systematic error varies from one acquisition track to another, a bank of prediction error filters, one for each acquisition track, makes a far better residual decorrelator. Karpushin and Brown (2001) implement this approach and report good suppression of acquisition footprint, with preservation of underlying bathymetric features.

In this paper, I take a somewhat different tack to the problem. I measure the difference in measured sea floor depth at "crossing points", or points in space where two acquisition tracks nearly cross. I then solve a least squares missing data problem to estimate the systematic error at all points between the crossing points, subtract the estimated systematic error from the original data, and use the simple inverse interpolation methodology of Claerbout (1999), supplemented with IRLS diagonal residual weights to suppress spikes, in order to make a final map. My approach generates maps of Galilee's seafloor which are generally free of acquisition footprint, and which also exhibit excellent preservation of subsea geologic features. Furthermore, my new approach visibly unbiases the residual, proving that the result is optimal from the standpoint of estimation theory.

METHODOLOGY

The simplest inverse interpolation approach outlined in (Claerbout, 1999) can be written in least squares fitting goals as follows.

$$\begin{aligned} \mathbf{Bm} - \mathbf{d} &\approx \mathbf{0} \\ \epsilon \mathbf{Am} &\approx \mathbf{0} \end{aligned} \tag{1}$$

$\mathbf{B}$ is nearest neighbor interpolation and maps a gridded model ($\mathbf{m}$) to the irregular data space ($\mathbf{d}$). The model grid is 860x500 points, while the data space consists of over 132,000 (x, y, z) triples. $\mathbf{A}$ is a model regularization operator, which penalizes model roughness. For all examples contained herein, $\mathbf{A} = \nabla$. ϵ balances the tradeoff between data fitting and spatial model smoothness.

To handle non-gaussian noise, Fomel and Claerbout (1995) implement an Iteratively Reweighted Least Squares (IRLS) scheme to nonlinearly estimate a residual weight which automatically reduces the importance of "bad" data in least squares estimation. Adding a diagonal residual weight to equation (1) gives

$$\begin{aligned} \mathbf{W}(\mathbf{Bm} - \mathbf{d}) &\approx \mathbf{0} \\ \epsilon \mathbf{Am} &\approx \mathbf{0}. \end{aligned} \tag{2}$$

To handle systematic errors between data tracks, Fomel and Claerbout (1995) supplement the residual weight in system (1) with a first derivative filter to decorrelate the residual. Lomask (1998) used a single prediction error filter (PEF). Karpushin and Brown (2001) use a bank of PEF's, one for each acquisition track. Whatever the case, we can refer to the differential operator as **D** and modify equation (2) to obtain a new system of equations:

$$\begin{aligned} \mathbf{WD}(\mathbf{Bm}-\mathbf{d}) &\approx \mathbf{0} \\ \epsilon\mathbf{Am} &\approx \mathbf{0}. \end{aligned} \tag{3}$$

W is the same as in equation (2), except for the addition of zero weights at track boundaries.

Estimation of systematic error from crossing points

It is fairly easy to compute "crossing points" from the raw data. I partition the data spatially into fairly small regions and, using a Fortran90 data structure, store the track index of each data point in the region. I decide that two data points cross if they belong to different tracks and are separated (spatially) by a distance less than a predefined threshold.

Unraveling the distribution of errors between the two crossing points is a much more difficult, and possibly intractable problem, in the absence of prior information. In general, each of the two crossing points contain an unknown combination of systematic and random error. In this case, I simply "split the difference" by assuming that, given a measured difference between two crossing points, the systematic error is distributed evenly between the two points.

Once I have a series of measured differences at crossing points, I estimate the systematic error at all points in the data space by solving a least squares missing data problem. I impose the assumption that the systematic error is smooth along acquisition tracks by adding a roughness penalty to the least squares objective function. In least squares fitting goals, these ideas are written as follows.

$$\begin{aligned} \mathbf{W_e}(\mathbf{Ke}-\mathbf{Ke_d}) &\approx \mathbf{0} \\ \nabla\mathbf{e} &\approx \mathbf{0} \end{aligned} \tag{4}$$

$\mathbf{e_d}$ is the measured differences at crossing points. **e** is the estimated systematic error. The reliability of the measured difference is inversely proportional to the spatial separation distance of the two crossing points, r. Arbitrarily, I choose the residual weight, $\mathbf{W_e}$, for a pair of crossing points as (1 - $(r/r_0)^2$), where r_0 is the predefined threshold which defines crossing points. **K** is a selector matrix. Figure 1 outlines the scheme.

Why Bother Directly Estimating Systematic Error?

If system (3) produces track-free maps, the skeptic might wonder, *why even bother directly estimating the systematic error in the data?* First, as mentioned earlier, system (3) leads to a loss of resolution in the final map. Second, in many cases we may have prior information

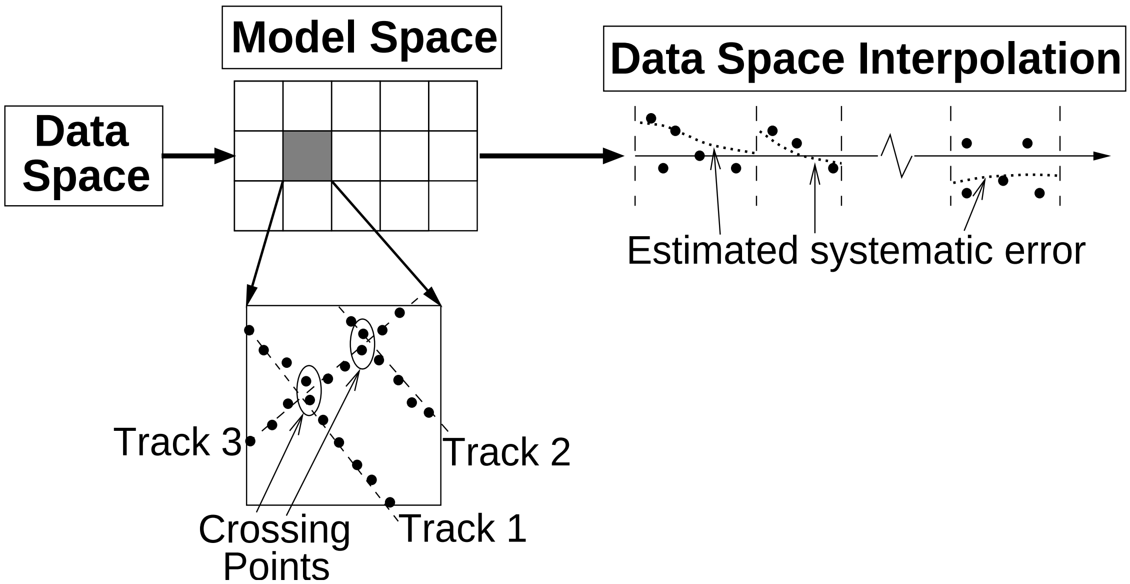

Figure 1: Algorithm for estimating systematic data error by inverting crossing point data. Data is partitioned into local regions, where crossing point differences are computed. The optimization of equation (4) operates in the data space. The measured crossing point differerences are the data; we use least squares to fit a smooth trend to the measured data. The optimization operates independently on individual tracks. morgan1-crosspoint [NR]

about the distribution and magnitude of the systematic error, which we could then include as an "inverse model covariance" (regularization operator) in system (4).

By directly estimating and subtracting systematic errors, we have more faith that the final map is an accurate representation of the true quantity. While the authors of the previously mentioned papers on Galilee and Madagascar were more interested in resolving the topographical features than the value of the underlying field, in many applications, the field itself is most important. Furthermore, when the systematic error is explainable by physical or other phenomena, we want to have control in its estimation.

Building a map with my new approach

My new approach can be summarized simply:

1. Estimate systematic error in data via system (4).

2. Subtract estimated systematic error from original data.

3. Generate final map using system (2).

RESULTS

In this section, I compare the results on the Galilee data of using systems (2) and (3) with my approach. I find the raw final maps uninterpretable, so instead, I will show roughened (first derivative) versions of the final maps, as well as the differences between the three final maps.

Figure 2 shows the final Galilee maps computed with the three discussed methods, and then roughened from southwest to northeast. The block-shaped artifacts on the periphery of the study area are due to the "Quadtree Pyramid" method used to generate a starting guess for the conjugate gradient iteration (Brown, 2000).

The result for system (3) in Figure 2 is free of the acquisition footprint which plagues the center panel (system (2)), but obscured by seemingly random noise. This noise is a result of track-end artifacts which "propagate", even though we have applied a zero weight on both sides of track boundaries. We could suppress the random noise by increasing ϵ, at the expense of resolution. The right panel of Figure 2 shows the result using my approach. Track artifacts are suppressed considerably, though not totally, compared to the center panel. The underlying geologic features look just as well resolved as in the center panel.

Figure 3 shows the difference between final maps generated by each of the three methods discussed in this paper. The left panel shows the difference between system (3) and system (2). We see track artifacts, as well as geologic features (238km north, 205km east), which confirms that first derivative along the track added to system (3) has caused some loss of resolution. The center panel of Figure 3 is the difference between the system (3) map and the map generated by my approach. First, we notice again the same loss of geologic features as in the left panel. Interestingly, we see correlated differences that are not along tracks. This tells us that although both approaches lead to white residuals, as we will see in Figure 4, they do not produce identical maps. This question needs to be answered. The right panel of Figure 3 shows the difference between my approach and system (2). As expected, we see considerable differences along tracks, but little to no geologic differences. This tells us that my approach has maintained the resolution inherent in system (2), while doing a good job of suppressing acquisition footprint.

In the northern region of the map, we notice mainly negative differences along east-to-west tracks, and positive differences along north-to-south tracks. In the southern region, this relationship is reversed, and the differences are noticeably smaller. The spatial regularity in differences implies that the systematic errors may well be correlated in time and/or space. Unfortunately, we do not have the times at which the samples were collected. We only know that they were collected between 1986 and 1987.

Figure 4 compares the data residuals for each of the three discussed methods. First, we see that IRLS alone (system (2)) produces a biased residual. The bias is most easily seen around sample 60,000 and sample 110,000. IRLS with track derivative (system (3)) produces an unbiased residual. On the bottom panel of Figure 4, we see that my approach leads to an effectively unbiased residual. Thus, from the viewpoint of optimization theory, my approach has achieved one half of the most important requirement of an optimal map: it is unbiased, although by inspection, not white. A better IRLS scheme should produce a more balanced,

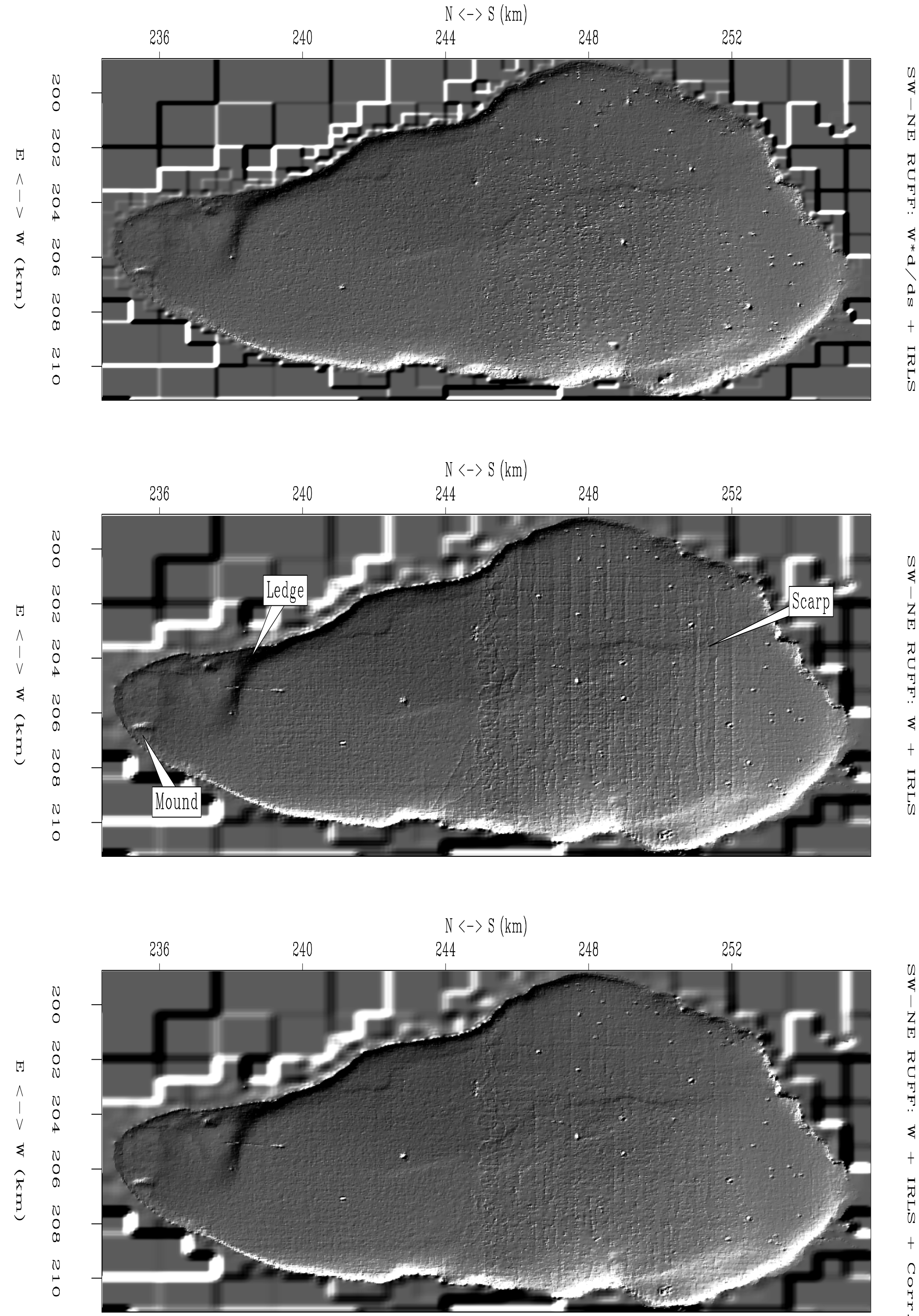

Figure 2: Final Galilee maps, roughened from southwest to northeast, using a simple difference filter. Left: IRLS with track derivative (system (3)). Center: IRLS only (system 2)). Right: My new approach–IRLS + crossing point correction. morgan1-galilee-ruff2 [ER,M]

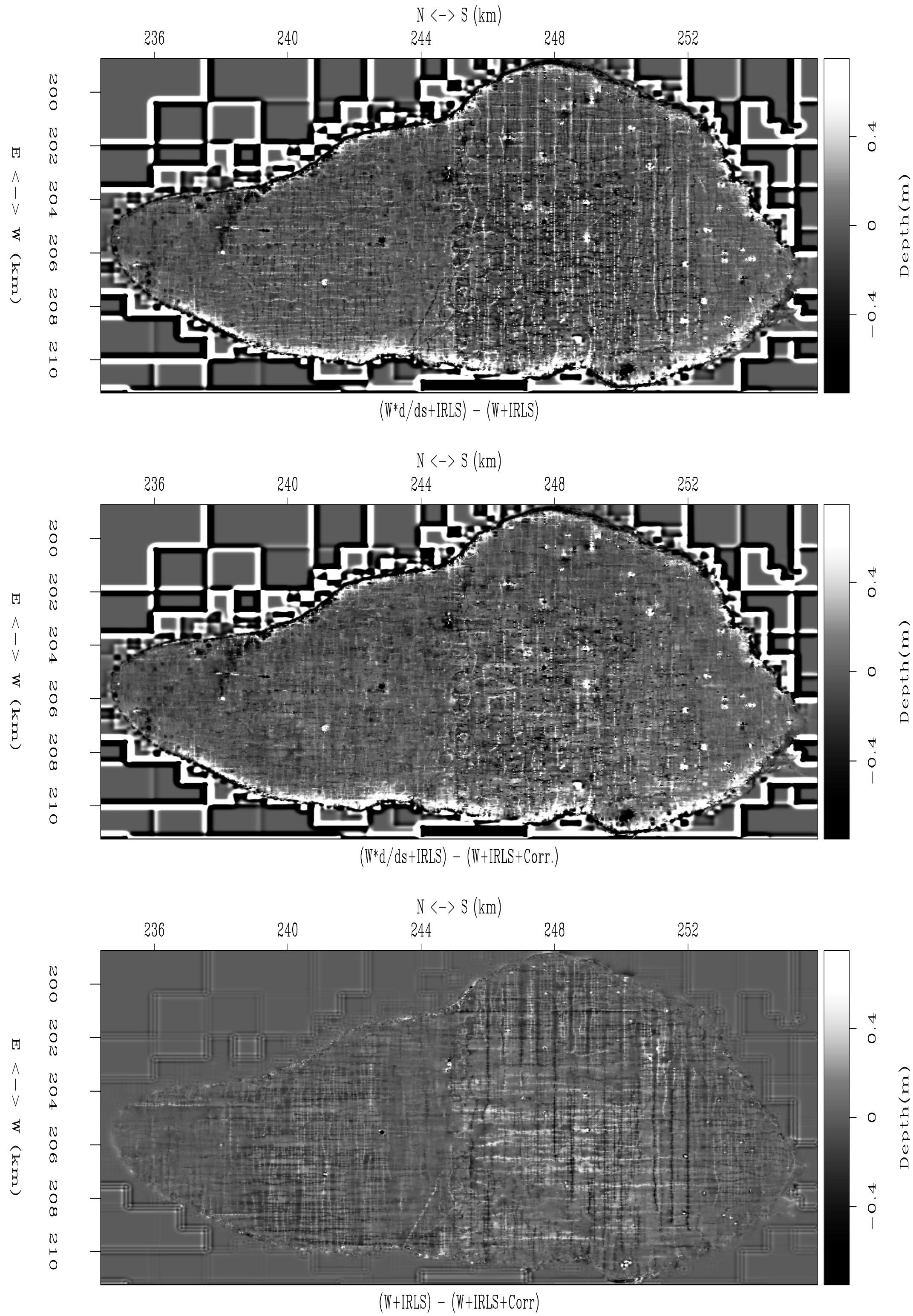

Figure 3: Difference between final Galilee maps produces by three discussed methods. Top: IRLS with track derivative (system (3)) minus IRLS only (system (2)). Center: IRLS with track derivative minus my new approach–IRLS + crossing point correction. Bottom: IRLS only minus my new approach. morgan1-galilee-diff [ER]

and hence white, residual for all three approaches.

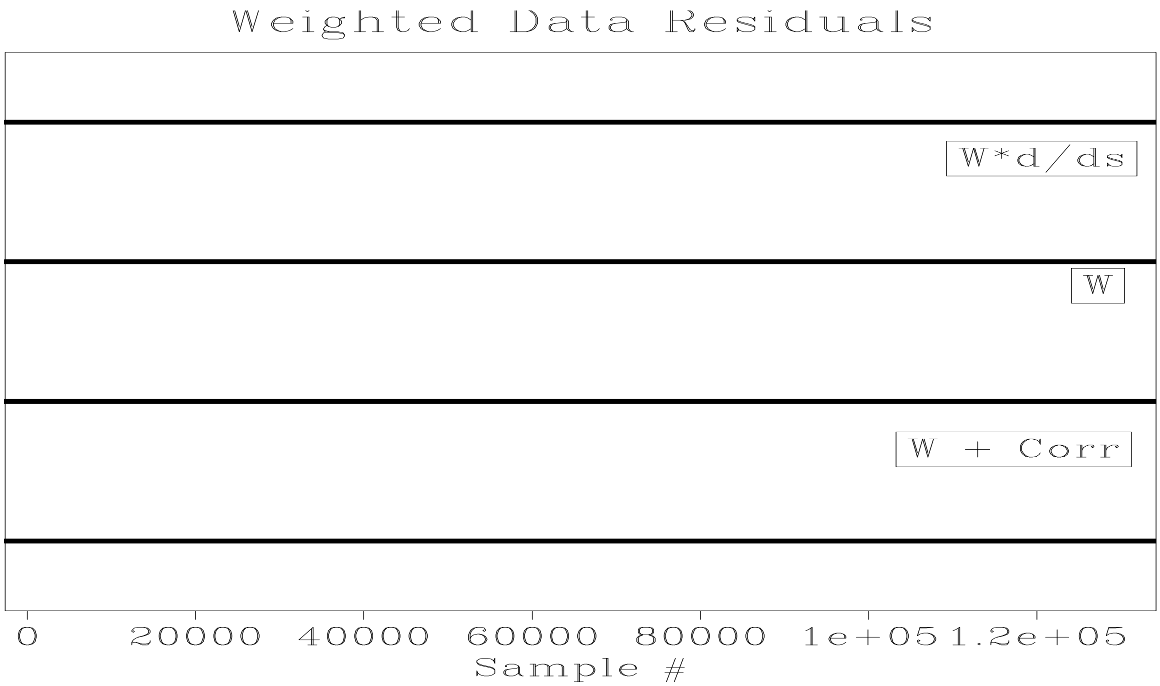

Figure 4: Residual error for the three discussed methods. Top: IRLS with track derivative (system (3)). Center: IRLS only (system (2)). Bottom: My approach. morgan1-galilee-resid [ER]

DISCUSSION

I presented a new method for building maps from tracked datasets. As discussed in previous works dealing with the Galilee bathymetry data, the biggest problem in building a map is systematic error between neighboring acquisition tracks. My method estimates systematic error between tracks by directly analyzing the difference between tracks at "crossing points" and using least squares optimization to estimate the error in points in between.

As we saw in tests on the Galilee data, my method effectively unbiases the data residual without a loss of resolution. My approach produces the most interpretable Galilee maps of the three approaches I discussed. Although the track artifacts are not removed completely, I believe the increased resolution relative to IRLS + track derivative approach weighs the balance in my approach's favor.

Looking to the future, I believe my approach offers a fundamental advantage over those of system (3)'s ilk. The philsophy behind system (3) assumes that we know that the residual is biased, but that we don't necessarily understand the data errors that cause the bias. In many cases, we *do* have strong prior information on the cause, magnitude, physics, and covariance of the systematic error. A good estimate of the systematic error may have interpretive value.

One weakness of my approach is that it reeks of "preprocessing". A more general and rigorous approach is that of Brown and Clapp (2001), which is basically an iterative variant of the approach that I've proposed here. Unfortunately, without prior information on the character of the systematic error, such an approach would be wasted.

REFERENCES

Ben-Avraham, Z., Amit, G., Golan, A., and Begin, Z., 1990, The bathymetry of Lake Kinneret: Israel J. Earth Sci, **39**, 77–84.

Brown, M., and Clapp, R. G., 2001, A least-squares approach for estimating integrated velocity models from multiple data types: SEP–**108**, 159–172.

Brown, M., 2000, Two strategies for sparse data interpolation: SEP–**105**, 49–66.

Claerbout, J., 1999, Geophysical estimation by example: Environmental soundings image enhancement: Stanford Exploration Project, `http://sepwww.stanford.edu/sep/prof/`.

Ecker, C., and Berlioux, A., 1995, Flying over the ocean southeast of Madagascar: SEP–**84**, 295–306.

Fomel, S., and Claerbout, J., 1995, Searching the Sea of Galilee: The splendors and miseries of iteratively reweighted least squares: SEP–**84**, 259–270.

Fomel, S., 2001, Three-dimensional seismic data regularization: Ph.D. thesis, Stanford University.

Guitton, A., 2000, Huber solver versus IRLS algorithm for quasi L1 inversion: SEP–**103**, 255–271.

Karpushin, A., and Brown, M., 2001, Whitening track residuals with PEFs in IRLS approach to the Sea of Galilee: SEP–**110**, 133–140.

Lomask, J., 1998, Madagascar revisited: A missing data problem: SEP–**97**, 207–216.

Nichols, D., 1994, Velocity-stack inversion using $\mathbf{L_p}$ norms: SEP–**82**, 1–16.