Localized Regression Analysis as a Method for Detecting Erroneous Measurements in Geospatial Databases, with Application to Gravity Databases

Qian Wen, Nigel Hicks, G. Randy Keller, Ann Q. Gates, and Vladik Kreinovich*

Pan-American Center for Earth and Environmental Studies,
University of Texas at El Paso, El Paso, TX 79968, USA

Abstract

Geospatial databases generally consist of measurements related to points (or pixels in the case of raster data), lines, and polygons. In recent years, the size and complexity of these databases have increased significantly and they often contain erroneous measurements or noise. In this paper, we address the problem of detecting erroneous and suspicious values in a database consisting of point measurements. We use a database of measurements of anomalies in the Earth's gravity field that we have complied as a test case, and we found that the standard methods of detecting erroneous measurements – based on regression analysis – do not work well. As a result, experts use manual methods to clean such databases that are very time-consuming. In this paper, we propose a (natural) "localized" version of regression analysis as a technique for automatic cleaning of the database and illustrate its efficiency in the case of this gravity database. We believe that this approach will prove to be useful when dealing with many other types of point data.

Keywords: Geospatial Databases; Error Detection; Gravity Data

1 Introduction

1.1 Motivation for the research

In many application areas, researchers and practitioners have collected a large amount of geospatial data. For example, geophysicists measure values of the gravity and magnetic fields, elevation, and reflectivity of electromagnetic energy for a broad range of wavelengths (visible, infrared, and radar) at different points; see, e.g., (Sharma 1997). Each type of data is usually stored in a large geospatial database.

Based on these measurements, geophysicists generate maps and images and derive geophysical models that fit these measurements. The geophysical use of gravity database compiled at UTEP is illustrated for a variety of scales in Cordell et al. (1982), Fliedner et al. (1996), Adams and Keller (1996), Tesha et al. (1997), Simiyu and Keller (1997), Grauch et al. (1999), Rodriguez-Pineda et al. (1999), and Keller (2001).

The main problem with the existing geospatial databases is that they are known to contain many erroneous points (e.g., Goodchild and Gopal, 1989; McCain and William, 1994; Scott, 1994). For example, there are several reasons why gravity measurements can be erroneous which include:

- there can be measurement errors in gravity and in elevation;
- there can be transcription errors;
- there may be an error in the instrument calibration;

*Corresponding author; email vladik@cs.utep.edu

- finally, there may be base station problems (gravity measurements are always relative to some "known" value).

Erroneous values can corrupt the results of data processing and analysis. In addition, many existing databases contain data from hundreds or even thousands of sources that may not be consistent with each other. Therefore before processing the measurements, it is important to clean them by eliminating obvious errors and by marking suspicious data points.

Elimination of erroneous data points is an important part of assuring and improving the quality of geospatial data, as recommended by the US Federal Standard (FGDC, 1998).

The main goal of this research was to investigate how to "clean" a gravity database. At present, the cleaning of gravity databases is done mainly "by hand", by a professional geophysicist looking at both the raw measurements and the preliminary results of processing these raw data. This expert also uses using experience and other information such as geologic maps. There are many useful map overlay and statistical techniques to help with this manual analysis (e.g., Scott, 1994), but even with these techniques, the manual cleaning is very *time-consuming* and *subjective*.

To overcome these two problems – i.e., to make the cleaning process less time-consuming and less subjective – it is necessary to design an *automated* method for eliminating erroneous measurements. We have addressed this problem here studying the case of measurements of the Earth's gravity field.

1.2 Contents and structure of the paper

The paper is structured as follows: In Section 2, we present an overview of our case study: gravity measurements and the gravity database. In Section 3, we describe the main idea of the existing methods for automatic cleaning of geospatial databases. These methods are mainly based on regression analysis (e.g., Goodchild and Gopal, 1989; Scott, 1994). These methods work well for many types of geospatial databases. For example, (Scott, 1994) describes the result of applying regression techniques for cleaning a groundwater database. However, we show that these methods do not work well for gravity databases where the errors are relatively subtle and thus that a modified method is needed.

In Section 4, we analyze the reasons why the existing regression-based techniques do not work well for databases such as the gravity database and use the results of our analysis to design a new method. This method exploits a natural idea of "localization" – that we should put more emphasis on comparing each measurement with the measurements in the neighboring locations (e.g., Anselin and Getis, 1992; Scott, 1994)). We show how this idea can be incorporated into regression analysis.

The resulting localized regression analysis method is summarized in Section 5. In Section 6, we describe the results of testing the new method on an actual gravity database. Future work is outlined in Section 7.

2 Case Study: Gravity Database

2.1 Why gravity measurements are important

Gravity measurements are one of the most important sources of geophysical and geological information. There are two reasons for this importance. First, in contrast to more widely used geophysical data like remote sensing images, that mainly reflect the conditions of the Earth's surface, gravitation comes from the whole Earth (e.g., Heiskanen and Meinesz, 1958; Heiskanen and Moritz, 1967). Thus gravity data contain valuable information about much deeper geophysical structures. Second, in contrast to many types of geophysical data, which usually cover a reasonably local area, gravity measurements cover broad areas and thus provide important regional information.

2.2 How gravity measurements are stored in the corresponding database

The accumulated gravity measurement data are stored at several research centers around the world. One of these data storage centers is located at the University of Texas at El Paso (UTEP). This center contains gravity measurements collected throughout the United States and Mexico and parts of Africa.

Since the gravity value is determined by the integral of the Earth's mass distribution over its entire volume, the measured value is almost the same in all the places. Using straightforward physical principles (Newton's law of gravitation), we can calculate variations caused by differences in latitude and elevation and obtain an almost perfect prediction of the measurements. Specific information about each site is therefore provided by the difference between the measured gravity value and the gravity value predicted on the basis of the known latitude and elevation. This difference is called the *Bouguer anomaly* (BA), and large anomalies are on the order of 1 part in 10,000 of the Earth's gravity field.

Accordingly, for each measurement, the corresponding record in the gravity database at UTEP the following fields:

- geographical coordinates of the point:
    - latitude (field 1);
    - longitude (field 2);
    - elevation (field 6);
- measured gravity value (field 7);
- Bouguer anomaly value (field 3);
- field 8 indicates who performed this particular measurement.

Fields 4 and 5 contain auxiliary correction values that take terrain into consideration:

- terrain correction for the outer zones in the Hammer (1939) system (radius > 0.9 km) (field 4), and
- terrain correction for the inner ($r < 0.9$ km) zones (field 5).

Terrain corrections are difficult to calculate and are a source of error in their own right especially in areas with high topographic relief. Thus, our database, as well as many others, does not have terrain correction values for all stations. A detailed treatment of this complication is beyond the scope of this paper. The terrain effects on neighboring stations are similar in magnitude and were thus ignored for the purpose of this study.

2.3 Test data suites used in this research

In the present research, we used two data sets:

- a data set which contains all the measurements from the region around El Paso, with latitude from 32.5 to 33 and longitude from -109.5 to -109;
- a data set which contains all the measurement from Mojave desert and surrounding region, with latitude from 33 to 38 and longitude from -118.5 to -111.

The first data set contains 550 measurements, the second data set contains 63,144 measurements.

We asked experts to look at these two data sets. According to the experts, the first data set does not contain any erroneous measurements, while the second data set contains several measurements which are clearly "dirty" (erroneous).

2.4 The need for additional "seeded" erroneous measurements

The existing gravity databases have already been cleaned "by hand". Because of recent research efforts and crude statistical methods, it is not surprising that there are no erroneous measurements in the El Paso region, but our research efforts recently detected erroneous values in the Mojave desert region in southeastern California and adjacent areas.

It is, of course, important to make sure that the automatic cleaning methods detect all the remaining erroneous measurements, but the main objective of the desired automatic cleaning system is to detect all the original erroneous measurements – most of which have been already eliminated by hand.

Since these originally detected erroneous measurements have not been preserved, we asked the experts who helped to clean the original database to provide us with simulated ("seeded") erroneous measurements, so that we will be able to check that they are indeed eliminated by our techniques.

In short, El Paso was a "clean" region, so it was seeded for testing the methods. After we tested our method with the seeded data, we moved to the Mojave desert region.

2.5 "Seeded" erroneous measurements used in this research

We added three seeded measurements to the El Paso region. Here are these measurements (latitude and longitude are measured in degrees, Bouguer anomaly in mGals, i.e., in 10^{-3} cm/s^2):

- latitude 32.5, longitude -109.3, Bouguer anomaly -200;
- latitude 32.6, longitude -109.4, Bouguer anomaly -100;
- latitude 32.9, longitude -109.2, Bouguer anomaly -50.

These values contrast with values in the region which mostly range from -140 to -180.

3 Regression Analysis as an Approach to Cleaning Geospatial Databases

3.1 Detecting outliers by regression analysis: general idea

The main idea of detecting outliers by using regression analysis is as follows: Based on the measurements, we can usually conclude that the value of the measured quantity y depends on the values of other physical quantities $p, \ldots, q$ at this location, e.g., on the elevation, latitude, and/or longitude. In other words, we can usually conclude that for each location, y is approximately equal to $f(p, \ldots, q)$ for some known function $f(p, \ldots, q)$. Since this dependence is confirmed by numerous experimental data points, we can conclude that this dependence is not a mathematical artifact, it is actually a physically meaningful dependence.

The fact that y is approximately equal to $f(p, \ldots, q)$ is usually reformulated as follows: the *approximation error* $e = y - f(p, \ldots, q)$ (also called *residual error*) is (reasonably) small.

What happens if one of the measurements y' is actually erroneous, i.e., quite different from the actual value y of the measured quantity?

- Since we concluded that the dependence $y \approx f(p, \ldots, q)$ has a physical meaning, the difference $y - f(p, \ldots, q)$ between the actual (unknown) value y of the measured quantity and the value $f(p, \ldots, q)$ should be reasonably small.
- However, since at this particular location, the measurement y' is drastically different from the actual value y of the measured quantity, the actually computed difference $e' = y' - f(p, \ldots, q)$ will be drastically different from the expected small difference $e = y - f(p, \ldots, q)$.

A number that is drastically different from a small number is not, by itself, small. Thus, for an erroneous measurement y', the residual error $e' = y' - f(p, \ldots, q)$ is much larger than usual.

In view of this conclusion, we can detect all the erroneous measurements as follows:

- first, we extract, from the measurements, the dependence $y \approx f(p, \ldots, q)$; this extraction is usually called regression;
- second, we compare the values of the residual errors $e = y - f(p, \ldots, q)$ at different locations; if at some location, the value of the residual is much larger than for all the others, this means that the measurement corresponding to this location is, most probably, erroneous.

In different physical situations, we can have dependencies $y \approx f(p, \ldots, q)$of different complexity. Let us describe possible cases starting from the simplest possible (degenerate) one: when y does not depend on any of the known physical quantities $p, \ldots, q$ (i.e., when the function $f(p, \ldots, q)$ is simply a constant).

3.2 Degenerate case: why it is important

At first glance, many things in nature are related to each other, so it may seem that the case when the measured value does not depend on any other predicable physical characteristic should be extremely rare. However, in practice, such cases are very frequent. Let us explain why.

To explain this fact, let us start with a simple example. Suppose that, before Ohm discovered his law, we are recording, for different electric circuits, the values of current I, resistance R, and voltage V. As a result, we get a huge database of records. When we analyze the data stored in this databases, we realize that for each of these records, the voltage is simply equal to the product $I \cdot R$. After we found this out, storing the values of voltage does not make sense any more. If we want to know the voltage, we can always multiply the measured values of current and resistance, and get the desired value of V.

Later on, it turns out that Ohm's law is only an approximation. For some materials, the actual voltage V is slightly different from the predicted value $I \cdot R$. In this case, it makes sense to measure the voltage, but instead of storing the actual value of voltage, it makes more sense to store and analyze the "voltage anomaly", i.e., the difference $V - I \cdot R$ between the measured and the predicted values of voltage.

We described this simple example to illustrate what happens in our case study of a gravity database. The measured gravity value depends on such physical characteristics of the location as its latitude and elevation. However, we have a very good idea of how the measured gravity value should depend on these characteristics. The existing models of gravitational field predict the measured gravity values with an accuracy of 0.02%. In other words, 99.98% of the measured value can be predicted without any measurement. The only part of the original gravity measurement that carries useful geophysical information is the *difference* between the measured and predicted gravity values – i.e., the Bouguer anomaly.

The dependence of the gravity value on the latitude, elevation, etc., is already captured by the calculations that predict the gravity value. Thus, not surprisingly, the residual differences between the measured and predicted values (Bouguer anomalies) do not seem to depend on the location (expect on a very broad scale where Bouguer anomalies negatively correlate with elevation).

For some data types (e.g., for groundwater elevation measurements), we do not originally know of any dependence on location. As a result, regression methods can uncover some dependence. As we gather and analyze more data, this dependence becomes better and better known. Eventually, we will know this dependence so well that, instead of analyzing the original measurements, we will be able to simply analyze the differences between the measured and predicted values.

3.3 Regression analysis in the degenerate case: towards a detailed description

In many real-life situations, data are normally distributed, with a mean a and a standard deviation σ. In this case, most measurements x_i lie within a certain number of standard deviations from the mean. For example, 99.9% of all the data lies within the "3 sigma" interval $[a - 3\sigma, a + 3\sigma]$; all but $10^{-6}\%$ of the data lies within the "six sigma" interval $[a-6\sigma, a+6\sigma]$, etc. (see, e.g., (Devore, 1999), (Wadsworth, 1993)). Thus:

- if a data point x_i is outside the three sigma interval, then this data point is most probably erroneous;
- if a data point x_i is outside the six sigma interval, then this data point is definitely erroneous.

3.4 Regression analysis in the degenerate case: an algorithm

In view of the above, it seems natural to apply this idea to a geospatial database. Namely:

- we estimate the mean a as the average of all the measured values $x_1, \ldots, x_n$:

$$a = \frac{x_1 + \ldots + x_n}{n};$$

- then, we estimate the standard deviation σ as the square root of the average of the squared difference $(x_i - a)^2$:

$$\sigma = \sqrt{\frac{(x_1 - a)^2 + \ldots + (x_n - a)^2}{n}};$$

- we then select two multiples of σ, $k \cdot \sigma$ and $K \cdot \sigma$, so that:
  - if for some measurement x_i, $|x_i - a| > K \cdot \sigma$, we declare this measurement x_i erroneous;
  - if for some measurement x_i, $|x_i - a| > k \cdot \sigma$, we mark the measurement x_i as suspicious.

The choice of the multiples of sigma depends on the size of the database:

- If the database contains 550 measurements, then, since the probability of a more than three sigma deviation is less than 10^{-3}, we should reasonably expect that no measurement is outside the corresponding interval. Thus, if $|x_i - a| > 3\sigma$, we expect x_i to be erroneous. In other words, for such a database, we take $K = 3$.

- If the database contains 63,144 measurements, then, to get the expected number of outside values to be around 0.5 (less than 1), we should select $K = 5$; for $K = 5$, the probability of a more than five sigma deviation is less than 10^{-5}. Thus, we should reasonably expect that no measurement is outside the corresponding interval $[a - 5\sigma, a + 5\sigma]$. Thus, if $|x_i - a| > 5\sigma$, we expect x_i to be erroneous. In other words, for such a database, we take $K = 5$.

3.5 Regression analysis in the degenerate case: robust versions

The main idea of detecting an erroneous measurements is that for these measurements are drastically different from the average of non-erroneous measurements. However, originally, we do not know which measurements are erroneous and which are not; therefore, we take the average of all measurements, an average that combines both the non-erroneous and the erroneous measurements.

If there are very few erroneous measurements, then their presence in the sum defining the average does not affect the result too much, so the average of all the measurements is approximately the same as the average of all non-erroneous measurements.

In some geospatial databases, however, a significant portion of measurements are erroneous. In such situations, if there are quite a few seriously biased measurements, the average value a will also be seriously biased. For such databases, we need methods of estimating the average which are *robust*, i.e., which do not change much if we add a few erroneous measurements; see, e.g., (Iglewicz and Hogalin, 1993), (S-Plus, 1989), and (S-Plus, 1991).

For example, we can use the following method:

- first, we apply the above-described method to estimate the average and standard deviation and to mark the outliers;

- second, if any outliers are indeed marked, we eliminate them from the original database, re-calculate the average and the standard deviation based only on the remaining measurements, and check if any more measurements will be thus marked as outliers;

- if any new outliers are marked, we eliminate them and repeat the procedure again and again until no new outliers are detected.

There are many other robust algorithms. For example, if we know the approximate percentage of erroneous measurements (e.g., 10%), then, instead of taking the average of all measurements, we can do the following:

- sort all the measurements from the smallest to the largest;

- when we find an average, we ignore the (10%/2=)5% of the smallest and 5% of the largest as maybe erroneous, and estimate a as the arithmetic average of the 90% middle ones;

- then, we proceed as above.

3.6 Regression analysis in non-degenerate case: in brief

In general, if there is a dependence of y on $p, \ldots, q$, then, in the first approximation, it is natural to consider linear dependence $y = c_0 + c_p \cdot p + \ldots + c_q \cdot q$ with some (unknown) coefficients $c_0, c_p, \ldots, c_q$. These coefficients can be determined, e.g., by the Least Squares Method (see, e.g., (Devore, 1999), (Wadsworth, 1993)), in which we find the unknown coefficients from the condition that the sum of the squared residuals of all the measurements $e_1^2 + e_2^2 + \ldots$ is the smallest possible. The optimized function is quadratic, so if we differentiate this function with respect to the coefficients and equate these derivatives to 0, we get an (easy-to-solve) system of linear equations for determining the unknown values $c_0, c_p, \ldots, c_q$.

If we expect a significant portion of measurements to be erroneous, then it makes sense to use robust versions of the regression analysis ((Iglewicz and Hoaglin, 1993), (Scott, 1994), (S-Plus, 1989), (S-Plus, 1991)).

After we have determined the coefficients, we can compute the residuals, and use the above techniques to mark the measurements for which the residuals are unusually large.

3.7 Testing the traditional regression techniques on gravity database: these techniques eliminate some erroneous measurements but overall, are not perfect

We have already mentioned that the gravity database is an example where we do not expect any dependence. So, to this database, we apply the techniques corresponding to the degenerate case.

3.7.1 First test

First, we tested the above method on the gravity database for the El Paso region, with the three seeded erroneous measurements added. As measurements, we took the values of the Bouguer anomaly.

Since El Paso database contains about 500 points, we selected $K = 3$. As a result, we got an average -160.08 and the standard deviation 9.92. The only values outside the corresponding three sigma interval are the three seeded points. So, for the gravity database with the seeded points, this method works perfectly well.

To test this method further, we applied this same method to the gravity database with the three seeded measurements removed (i.e., to the original cleaned gravity database). Since all the values in this database are clean, we expected this method to not eliminate any of these measurements. Instead, we got a mistaken elimination. Namely, for the clean database, the average is -160.32, the standard deviation is 8.21, and there are two measurement outside the corresponding three sigma interval:

- latitude 32.931831, longitude -109.050827, Bouguer anomaly -185.99;
- latitude 32.938831, longitude -109.184158, Bouguer anomaly -189.44.

According to the experts, these two measurements were correct and low because they were located in a deep geologic basin, but they were mistakenly eliminated by the tested method.

Again, we expected less than 1 point to be mistakenly eliminated, but got 2 instead. This fact clearly shows that this method is not working well enough.

3.7.2 Second test

We also tested this method on the measurements from the Mojave desert region. For this region, the average is -115.96, the standard deviation is 52.89, and no suspicious measurements were reported at all, in spite of the fact that some measurements are erroneous. This result further confirms that the above method is not working well enough.

3.7.3 Maybe some dependence on the latitude and elevation is missed in the definition of Bouguer anomaly? just checking

To be on the safe side, we decided to check whether, by any chance, some residual dependence of gravity on latitude and elevation is missed in the definition of Bouguer anomaly. To check this, instead of taking the average of all the values of Bouguer anomaly y, we:

- first tried to approximate it by a linear formula $y \approx c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + c_3 \cdot x_3$, where x_1 is the latitude, x_2 is the longitude, and x_3 is the elevation; and
- then, applied to above outlier-detecting techniques to the residual errors $y - (c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + c_3 \cdot x_3)$.

For the El Paso region, the results are even worse than for the degenerate case: instead of 2, we now get 5 points that are mistakenly labeled as erroneous:

- latitude 32.581501, longitude -109.395332, Bouguer anomaly -155.73;
- latitude 32.887489, longitude -109.313492, Bouguer anomaly -170.77;
- latitude 32.931831, longitude -109.050827, Bouguer anomaly -185.99;
- latitude 32.938831, longitude -109.184158, Bouguer anomaly -189.44;
- latitude 32.995659, longitude -109.076828, Bouguer anomaly -173.67.

The fact that we did not get a better result shows that the formulas for Bouguer anomaly do not miss any dependence.

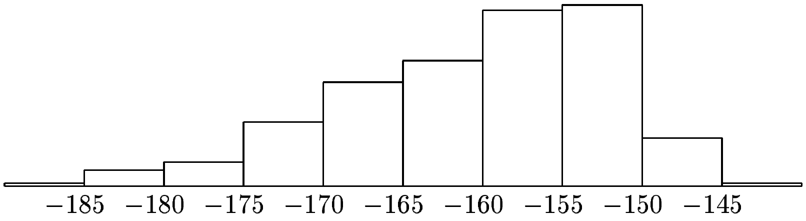
The fact that we got a worse result can be easily explained: the formulas for the gravity model used in computing Bouguer anomaly are based on fundamental physical principals and employ constants that are based on hundreds of thousands of measurements. What we did was, in effect, try to derive an empirical formula from only 550 measurements. Of course, based on 550 measurements, we got a much less accurate formulas for the gravity model, and this additional inaccuracy further decreased the accuracy of our statistical analysis.

4 Analysis of Why the Traditional Regression Techniques Do Not Work Well, And the Design of New – Localized – Regression Techniques

4.1 First conclusion: the overall distribution is not Gaussian

The traditional regression techniques are based on the assumptions that the measured values are normally distributed. In this case, we should indeed expect less than 1 out of 550 measurement to be outside the three sigma interval. The fact that we have two measurements outside the interval shows that the overall distribution is not normal.

The histogram of the values of the Bouguer anomaly confirms that the distribution is not Gaussian:



4.2 Why is the overall distribution not Gaussian?

One of the main reasons why the overall distribution is not Gaussian is that the region combines several zones with different geophysical structure. Within each zone, the distribution seems to be Gaussian, but when we put together measurements coming from different zones, we thus combine Gaussian distributions with different values of mean and standard deviation. The resulting combination is not Gaussian, it is closer to a bimodal distribution.

4.3 How to overcome this difficulty? Main idea of localization

To overcome this difficulty, we can, therefore, instead of considering the overall mean and standard deviation, *localize* the computations, i.e.:

- subdivide the original region into subregions, and then
- check whether a given measurement is erroneous or not, we compare it with the mean and standard deviation for the corresponding subregion.

To make sure that the data within our analyzed subregion is as homogeneous as possible, we must select these subregions to be as small as possible.

If a selected subregion contains too few points, we cannot get any meaningful statistical estimates of a and σ out of these points and thus, we cannot check whether a given measured value lies within the "k sigma" interval. (For geospatial data, the fact that we need a sufficient number of data points to make statistically meaningful conclusions is emphasized, e.g., in (Kennedy, 1989).) Thus, we must select these subregions in such a way that we will be able to apply statistical estimates to them.

From measurement theory, it is known that we need at least 40 points to make a valid statistical estimate of a and σ and make a justified

decision on the error (see, e.g., (Rabinovich, 1993)). So, we divide the geographical region into subregions each of which contains at most 40-50 points, and apply the above regression analysis techniques not to the entire region, but only to these subregions.

4.4 Subdividing into subregions: details

A natural way of subdividing the geographic region is to subdivide it into smaller zones. We would like each zone to contain approximately 40-50 points. Therefore, to estimate the number of such zones, we can simply divide the total number of measurements by 50. In the El Paso region, that would mean that we need to divide the original region into approximately 550/50=11 zones.

There are many different ways to subdivide the original region into a selected number N_Z of zones. Let us describe which subdivision is the most reasonable one.

Let $\mathcal{A}$ be the area of the original geographic region (measured in square kilometers or square miles or in any other appropriate units). When we divide the region into N_Z subregions, the average area of the resulting subregions is $\mathcal{A}/N_Z$.

Usually, in a geospatial database, we deal with a rectangular region of certain width W and certain height H. For such a region, a natural way to divide it into subregions is to do the following:

- subdivide its width W into a certain number n_1 of equal parts; and
- subdivide its height H into a certain number n_2 of equal parts.

As a result, we divide the original rectangle into $N_Z = n_1 \cdot n_2$ smaller rectangles of size $w \times h$, where $w = W/n_1$ and $h = H/n_2$.

For the same number N_Z, we can have different possible subdivisions. For example, there are four different ways to subdivide a rectangle into $N_Z = 6$ subregions:

- we can subdivide the width into 6 parts ($n_1 = 6$) and leave height unchanged ($n_2 = 1$);
- we can subdivide the width into 3 parts ($n_1 = 3$) and height into 2 parts ($n_1 = 2$);

- we can subdivide the width into 2 parts ($n_1 = 2$) and height into 3 parts ($n_1 = 3$);

- finally, we can leave the width unchanged ($n_1 = 1$) and subdivide the height into 6 parts ($n_2 = 6$).

Which of the subdivisions is the most appropriate?

The main idea behind selecting subregions is that the data within each subregion should be homogeneous. In general, the larger the distance between the two points, the more different the geophysical conditions at these points. Thus, as an estimate of the quality of a subdivision, we can take the largest possible distance between the two points in a subregion. For a rectangular subregion, the largest distance is between the opposite edges, so this largest distance is equal to the diagonal of the rectangle. Thus, among all possible subdivisions, we must pick up a subdivision in which the diagonals are the shortest possible.

The area of each rectangular subregion is known – as we have shown, it is determined by the number of subregions, which, in its turn, is determined by the total number of measurements in the given geospatial area. Thus, in order to find an appropriate subdivision, we must select, among all possible rectangles of a given area $h \times w$, the rectangle with the smallest possible diagonal.

The length of the diagonal equals $\sqrt{h^2 + w^2}$. Thus, we must find, among all possible values h and w with a given product $w \cdot h$, the values for which the sum $w^2 + h^2$ is the smallest possible. One can easily see that the optimal values correspond to $w = h$ – a square. Thus, we should have subregions as close to squares as possible.

On average, the resulting squares contain around 50 measurements. The number of measurements may differ from zone to zone, so some zones may contain more than 50 measurements. In view of what we said before, in Section 4, we need to subdivide them further.

Subdividing all such zones would mean a lot of computations. Luckily, as we have mentioned in Section 4.3, if a zone is not too large (e.g., if this size zone contains, on average, about 50 points), then we can get all erroneous points marked as suspicious without further subdividing this zone. In view of this observation, if in one of the zones with more than 50 measurements in it, we do not find any suspicious points (i.e., if all measurements are within a reasonable multiple of standard deviation from the average over this zone), then it is reasonable to assume that this zone is clean.

However, as we mentioned in Section 4.3, if in a zone with more than 50 measurements, one of these measurements does deviate too much from the average over this zone, this does not necessarily indicate that this particular measurement is erroneous. So, to decrease the number of suspicious measurements, it is desirable to:

- further subdivide this zone into subzones until we get approximately 50 points in the corresponding subzone, and

- check whether the originally suspicious points indeed remain suspicious when we compute average and standard deviation not over the entire zone, but only over the corresponding subzone.

4.5 Second idea: detecting erroneous measurements by tracing problematic data sources

In the above text, we discussed the possibility of detecting individual erroneous measurements. However, as have mentioned earlier, one of the main reasons why measurements occur is that some data sources are problematic. It therefore makes sense not only to look at the measured values, but also to look at the data source. If it turns out that a certain data source indeed contains erroneous measurements, then it makes sense to mark all measurements coming from this particular data source as suspicious. This natural idea was successfully used in (Scott, 1994) to detect erroneous measurements in groundwater database.

A typical way in which a problematic data source errs is by having a consistent bias. Thus, to detect such a data source, we can do the following:

- For each zone (subregion) z which contains the values from this particular data source, we compute the average $m(z)$ of all the measurements from this zone z coming from this data source, and compute the bias $b(z)$ as the difference between the average $m(z)$ and the average $a(z)$ of all the other measurements performed in this zone: $b(z) = m(z) - a(z)$.

- If the absolute value of the resulting bias $b(z)$ exceed some threshold (30 in the case of a gravity database), then we declare all the measurements in this zone which come from this particular data source to be suspicious.

We have already mentioned that a statistical estimation only make sense if we have at least 40-50 different values. Thus, to have a meaningful estimate of a bias corresponding to a given data source, we must only consider data sources who have made at least 50 measurements in a given geographical region.

4.6 Third idea: suspicious zones

In the original idea, we assumed that there are very few "dirty" points, and most measurements are correct. In this case, the estimated mean and standard deviation mainly reflect the correct points.

In some cases, however, in some geographical zones, the relative number of dirty points is higher than usual. As a result, the estimated standard deviation reflects not only the correct points, but the erroneous points as well. Since the resulting estimate for standard deviation comes from two different populations with drastically different measurements, the resulting estimate becomes large, and so, we may not be able to use a "k sigma" criterion to detect the dirty points in the analyzed zone.

For example, if, in the degenerate case, we have an equal number of correct and erroneous measurements, all correct measurement are approximately 0 and all erroneous measurement are approximately 1, then the estimated average is 0.5, and the estimate standard deviation is 0.5. Hence, in this situation, all the points (both correct and the erroneous ones) lie within the three sigma interval.

How can we detect such zones? As we have mentioned, in such zones, the estimated standard deviations are unusually large. So, to detect such zones, we can use the same idea as we used to detect the erroneous measurements in the first place: we compute the average A and standard deviation D of the estimated standard deviations $\sigma(z)$ corresponding to different zones. If for some zone z, $\sigma(z)$ is outside the corresponding "k sigma" interval $[A - k \cdot D, A + k \cdot D]$, we declare this zone to be suspicious.

4.7 Last idea: suspicious neighboring zones

We have mentioned that one reason for erroneous measurements is the fact that we have problematic data sources.

If several zones are neighbors to each other, this probably indicates that in the entire connected area, there are problematic data sources affecting these zones. Since the borders between zones are rather arbitrary, it is quite possible that the same problematic data source contributed to the zones, which are close to this area. Therefore, if we have detected a connected block of zones with erroneous measurements, it makes sense to also check zones which are direct neighbors to these ones.

To avoid accidental groupings of two zones, it makes sense to consider only connected areas that consist of at least three zones. These neighbors should be classified as "somewhat suspicious", to indicate that they are not as highly suspicious as the original zones. If a zone is a "diagonal" neighbor to the connected area (i.e., if it only has a common edge with one of the zones from the area), then we still consider it suspicious, but we consider it even less suspicious - "mildly suspicious". As a result, we arrive at the following method.

5 Localized Regression Analysis as a Method for Detecting Erroneous Measurements in Geospatial Databases

5.1 General description of a method

We propose the following method for automatic detection of erroneous measurements in geospatial databases:

1) First, we use the total number N of measurements to select the parameter k in the "k sigma" interval. Specifically, we select k in such a way that the probability of getting a normally distributed variable outside the "k sigma" interval should be smaller than $1/N$ and thus, the expected number of values outside this interval is less than 1. For example:

- for $N \approx 500$, we use $k = 3$;
- for $N \approx 10,000$, we get $k = 4$;
- for $N \approx 50,000$, we get $k = 5$.

2) Second, we subdivide the geographic region into approximately square subregions (zones) each of which contains, on average, approximately 50 measurements.

3) For each zone z, we apply the standard regression analysis techniques to detect the outliers. In particular, in the degenerate the following:

   1. We estimate the average $a(z)$ and standard deviation $\sigma(z)$ of all the measured values from this zone.
   2. measurements outside the interval $[a(z) - k \cdot \sigma(z), a(z) + k \cdot \sigma(z)]$ are considered to be outliers.

   If the standard regression analysis did not detect any outliers, we proclaim the zone to be (so far) clean, and move to other zones.

   On the other hand, if some measurements from the zone z were detected as outliers, we count the overall number of measurements $N(z)$ in the zone z, and compare it with 50. Then:

   Y. If $N(z)$ is 50 or less, we consider all outliers to be suspicious.

   N. If $N(z)$ is larger than 50, then we further subdivide the zone z into approximately square subsubregions (subzones) each of which contains, on average, approximately 50 measurements. Since we start with an almost square zone, in order to make sure that the resulting subzones are square, we must make sure that the number of latitudinal subdivision is approximately equal to the number of longitudinal subdivisions.

   Specifically:

   - If $N(z)$ is between 50 and 100, we divide into 2 subzones. This subdivision can be either latitudinal or longitudinal, the choice depends on whether the geophysical structure of the region changes more from North to South or from East to West: If the structures mainly change from East to West, then we should divide longitudes by half. If the structures mainly change from North to South, then we should divide latitudes by half.
   - If $N(z)$ is between 100 and 150, we divide into 3 subzones. This subdivision can be either latitudinal or longitudinal, the choice depends on whether the geophysical structure of the region changes more from North to South or from East to West.
   - If $N(z)$ is between 150 and 200, we divide into 2 x 2 subzones.
   - If $N(z)$ is between 200 and 300, we divide into 2 x 3 or 3 x 2 subzones, depending on the prevailing geophysical structure.
   - If $N(z)$ is between 300 and 450, we divide into 3 x 3 subzones.

   (More than 450 measurements in a single zone is highly unlikely, because with an average of 50 measurements per zone, this would mean an extremely uneven distribution of measurements.)

   For each of the resulting subzones Z, we again apply the standard regression analysis estimate the, and check whether the measurements originally marked as outliers are marked as outliers by the new analysis. If they are so marked, we declare them suspicious.

4) Select all data sources that contributed at least 50 measurements in the given geographical region. For each of these data sources, we do the following:

   1. We mark all the zones that contain measurements from this data source. For each such zone (subregion) z, we:

      i) compute the average $m(z)$ of all the measurements from the zone z which comes from this data source;

      ii) compute the bias $b(z) = m(z) - a(z)$ (where $a(z)$ is the average of all the other measurements performed in this zone).

2. If the absolute value $|b(z)|$ of the resulting bias $b(z)$ exceed a pre-defined threshold, then we declare all the measurements in this zone which come from this particular data source to be suspicious.

5) We compute the average A and standard deviation D of the estimated standard deviations $\sigma(z)$ corresponding to different zones.

6) We check, for each zone z, whether $\sigma(z)$ is outside the corresponding "k sigma" interval $[A - k \cdot D, A + k \cdot D]$. If it is, we declare this zone to be suspicious.

7) If the list of suspicious zones contains a connected block of three or more zones, we should also mark:

   - as somewhat suspicious, all the zones which are direct neighbors to these ones;
   - as mildly suspicious, all the zones which are diagonal neighbors to these ones.

As a result of this algorithm, we have three group of zones:

- zones marked as suspicious;
- zones marked as somewhat suspicious;
- zones marked as mildly suspicious.

6 Testing the Proposed Method on the Actual Gravity Database

6.1 Selecting the value of k

We have $N = 550$ measurements in the El Paso region, so, in accordance with the above algorithm, we selected $k = 3$.

For the Mojave desert region, we have $N \approx 60,000$ measurements, so, in accordance with the above algorithm, we selected $k = 5$.

6.2 Subdividing the original region into subregions

In accordance with the above algorithm, we need to divide the El Paso region into approximately $550/50 = 11$ subregions. The El Paso region is almost square, so, to get almost square subregions, we must subdivide it into the same number of subdivisions both in latitude and in longitude. If we divide both width and height of this rectangle by n, we get n^2 subregions. The desired number 11 is not a square; so, we can either pick $n = 3$ and get 9 zones, or pick $n = 4$ and get 16 zones. Since 11 is closer to 9 than to 16, we selected $n = 3$. This selection means that we divide the $1/2 \times 1/2$ degree geographical region into $1/6 \times 1/6$ degree zones (approximately 30×30 km).

It is easy to check that a similar subdivision into $1/6 \times 1/6$ degree zones works well for the Mojave desert region as well: indeed, we get a subdivision into $30 \times 45 = 1,350$ zones, an average of ≈ 47 measurements per zone.

6.3 Erroneous measurements indicated by experts

As we have mentioned, according to the experts, the gravity data set corresponding to El Paso region does not contain any erroneous measurements.

We asked experts to point out the erroneous measurements in the Mojave desert region data set. According to the experts, this data set contains several "dirty" zones, i.e., zones, in which, according to the experts, some measurements are clearly "dirty" (erroneous). There are 22 such zones, described by latitude×longitude:

$$\left[33\frac{2}{3}, 33\frac{5}{6}\right] \times \left[-117\frac{1}{2}, -117\frac{1}{3}\right]; \quad \left[34, 34\frac{1}{6}\right] \times \left[-113\frac{1}{2}, -113\frac{1}{3}\right]; \quad \left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right];$$

$$\left[34\frac{1}{6},34\frac{1}{3}\right]\times\left[-115\frac{2}{3},-115\frac{1}{2}\right];\ \left[34\frac{1}{6},34\frac{1}{3}\right]\times\left[-115\frac{1}{2},-115\frac{1}{3}\right];\ \left[34\frac{1}{6},34\frac{1}{3}\right]\times\left[-115\frac{1}{3},-115\frac{1}{6}\right];$$

$$\left[34\frac{1}{3},34\frac{1}{2}\right]\times\left[-115\frac{5}{6},-115\frac{2}{3}\right];\ \left[34\frac{1}{3},34\frac{1}{2}\right]\times\left[-115\frac{2}{3},-115\frac{1}{2}\right];\ \left[34\frac{1}{3},34\frac{1}{2}\right]\times\left[-115\frac{1}{2},-115\frac{1}{3}\right];$$

$$\left[34\frac{1}{3},34\frac{1}{2}\right]\times\left[-115\frac{1}{3},-115\frac{1}{6}\right];\ \left[34\frac{1}{2},34\frac{2}{3}\right]\times\left[-115\frac{1}{6},-115\right];\ \left[34\frac{1}{2},34\frac{2}{3}\right]\times\left[-115,-114\frac{5}{6}\right];$$

$$\left[34\frac{1}{2},34\frac{2}{3}\right]\times\left[-115\frac{5}{6},-115\frac{2}{3}\right];\ \left[34\frac{1}{2},34\frac{2}{3}\right]\times\left[-115\frac{2}{3},-115\frac{1}{2}\right];\ \left[34\frac{1}{2},34\frac{2}{3}\right]\times\left[-115\frac{1}{2},-115\frac{1}{3}\right];$$

$$\left[34\frac{1}{2},34\frac{2}{3}\right]\times\left[-115\frac{1}{3},-115\frac{1}{6}\right];\ \left[34\frac{2}{3},34\frac{5}{6}\right]\times\left[-115\frac{1}{6},-115\right];\ \left[34\frac{2}{3},34\frac{5}{6}\right]\times\left[-115,-114\frac{5}{6}\right];$$

$$\left[34\frac{2}{3},34\frac{5}{6}\right]\times\left[-114\frac{5}{6},-114\frac{2}{3}\right];\ \left[35\frac{1}{6},35\frac{1}{3}\right]\times\left[-115\frac{1}{3},-115\frac{1}{6}\right];\ \left[35\frac{1}{2},35\frac{2}{3}\right]\times\left[-111\frac{5}{6},-111\frac{2}{3}\right];$$

$$\left[36\frac{5}{6},37\right]\times\left[-113\frac{1}{2},-113\frac{1}{3}\right].$$

6.4 Testing the method on the El Paso region data set

We started by testing our new method on the El Paso region data set. In accordance with our algorithm, for each of the 9 zones z, we estimated the average $a(z)$ and the standard deviation $\sigma(z)$ and checked whether each measurement belongs to the corresponding interval $[a(z)-k\cdot\sigma(z), a(z)+k\cdot\sigma(z)]$, where, for this data set, $k=3$. All measurements except for one turned out to be within the corresponding interval.

This seemingly suspicious measurement is within the zone $\left[32\frac{2}{3},32\frac{5}{6}\right]\times\left[-109\frac{1}{2},-109\frac{1}{3}\right]$. For this zone, the average is -152.62, and the standard deviation is 2.72. The measurement outside the corresponding interval $[-146.46,-162.72]$ is:

- latitude 32.805328, longitude −109.492828, Bouguer anomaly −143.50.

To check whether this measurement is truly erroneous, we first found out the number $N(z)$ of measurements in this zone, it was 86. So, in accordance with our algorithm, we subdivided the original zone into two subzones.

Since the geologic structures trend N-S in the El Paso region, we divided the original zone into two subzones by latitude: $\left[32\frac{2}{3},32\frac{3}{4}\right]\times\left[-109\frac{1}{2},-109\frac{1}{3}\right]$ and $\left[32\frac{3}{4},32\frac{5}{6}\right]\times\left[-109\frac{1}{2},-109\frac{1}{3}\right]$. The suspicious point belongs to the second subzone. For this subzone Z, we have $a(Z)=-151.92$, $\sigma(Z)=3.34$, and so the resulting "three sigma" interval $[141.90,161.94]$ does contain the seemingly suspicious value. Thus, we can conclude that this seemingly suspicious value is not really suspicious.

So, according to our algorithm, all the measurements from the El Paso region data set seem to be correct – exactly as the experts suggested.

6.5 Testing the method on the Mojave desert region data set: Part I

For the Mojave desert data set, stages 1)-3) of the above algorithm detected the following 26 suspicious measurements:

latitude	longitude	BA	latitude	longitude	BA	latitude	longitude	BA
33.28350	−114.794	19.79	33.31383	−116.494	−98.55	33.35367	−111.661	−142.58
33.40517	−111.939	−132.72	33.438	−112.815	−368.65	33.67433	−114.062	−97.92
33.68234	−117.304	−20.68	33.70233	−117.474	−141.88	33.86984	−117.167	−116.03
34.00467	−117.688	−179.37	34.047	−115.219	147.93	34.12983	−113.475	−146.41
34.14017	−116.631	−150.90	34.205	−114.291	55.21	34.2315	−114.490	114.07
34.34167	−116.697	−161.46	34.98383	−114.409	−128.31	35.02633	−115.357	−322.02
35.10933	−114.025	−144.45	35.523	−111.725	−90.69	35.57683	−118.113	−250.06
35.641	−117.875	−36.56	36.77033	−118.437	−243.85	36.78249	−118.006	−262.97
36.91283	−116.288	−273.98	37.0075	−113.550	−124.85			

Three of these dirty points confirmed that the three originally given zones indeed contain erroneous measurements. These three points are:

latitude	longitude	BA	latitude	longitude	BA	latitude	longitude	BA
33.70233	−117.474	−141.88	34.12983	−113.475	−146.41	35.523	−111.725	−90.69

These points belong to the following three zones:

$$\left[33\frac{2}{3}, 33\frac{5}{6}\right] \times \left[-117\frac{1}{2}, -117\frac{1}{3}\right] \quad \left[34, 34\frac{1}{6}\right] \times \left[-113\frac{1}{2}, -113\frac{1}{6}\right] \quad \left[35\frac{1}{2}, 35\frac{2}{3}\right] \times \left[-111\frac{5}{6}, -111\frac{2}{3}\right].$$

The remaining 23 points were not in the original list of dirty zones, but when we showed these points to the experts, they agreed that these measurements – which deviate from the average of the corresponding zones by more than 5 sigma – must be erroneous.

Out of the original 22 dirty zones, the stage 2)-3) of our algorithm picked 3. To pick the rest, we continued the application of our algorithm.

6.6 Testing the method on the Mojave desert region data set: Part II

In accordance with our algorithm, we did the following:

- First, we selected all the data sources which contributed at least 50 measurements to the database.
- For each such data source, we computed the bias $b(z)$ for every zone z which contains data from this particular data source. All the zones for which, for some data source, the absolute value of the bias exceeds 30, were marked as suspicious.

Here is the resulting list of the zones that were thus marked as suspicious, with the indication of the exact data source which caused this suspicion:

- $\left[34, 34\frac{1}{6}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]$; suspicious data sources:
    - c2, $a(z) = -76.53$, $m(z) = -107.94$, $|b(z)| = 31.42$;
- $\left[34, 34\frac{1}{6}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]$; suspicious data sources:
    - ARIZ, $a(z) = -71.45$, $m(z) = -33.75$, $|b(z)| = 37.70$;
- $\left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right]$; suspicious data sources:
    - c1, $a(z) = -84.31$, $m(z) = -149.41$, $|b(z)| = 65.10$;
- $\left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{2}{3}, -115\frac{1}{2}\right]$; suspicious data sources:
    - c1, $a(z) = -78.22$, $m(z) = -126.66$, $|b(z)| = 48.44$;
- $\left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{2}{3}, -115\frac{1}{2}\right]$; suspicious data sources:
    - .c0m=, $a(z) = -97.30$, $m(z) = -67.19$, $|b(z)| = 30.11$;
- $\left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]$; suspicious data sources:
    - NDL, $a(z) = -96.16$, $m(z) = -63.49$, $|b(z)| = 32.67$;
- $\left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]$; suspicious data sources:
    - ARIZ, $a(z) = -101.97$, $m(z) = -70.45$, $|b(z)| = 31.52$;

  - 86RCJ, $a(z) = -101.61$, $m(z) = -69.02$, $|b(z)| = 32.59$;
  - c1, $a(z) = -89.67$, $m(z) = -147.55$, $|b(z)| = 57.88$;
  - NDL, $a(z) = -101.25$, $m(z) = -65.83$, $|b(z)| = 35.42$;
  - ow, $a(z) = -107.17$, $m(z) = -76.22$, $|b(z)| = 30.94$;

- $\left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115, -114\frac{5}{6}\right]$; suspicious measurements:
  - w, $a(z) = -85.65$, $m(z) = -171.03$, $|b(z)| = 85.38$;
- $\left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-115\frac{1}{6}, -115\right]$; suspicious measurements:
  - ARIZ, $a(z) = -156.11$, $m(z) = -88.60$, $|b(z)| = 67.52$;
  - w, $a(z) = -92.04$, $m(z) = -187.66$, $|b(z)| = 95.62$;
- $\left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-115, -114\frac{5}{6}\right]$; suspicious measurements:
  - ARIZ, $a(z) = -171.15$, $m(z) = -83.42$, $|b(z)| = 87.73$;
  - w, $a(z) = -84.67$, $m(z) = -174.53$, $|b(z)| = 89.86$;
- $\left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-114\frac{5}{6}, -114\frac{2}{3}\right]$; suspicious measurements:
  - ARIZ, $a(z) = -98.01$, $m(z) = -63.58$, $|b(z)| = 34.42$;
  - w, $a(z) = -62.17$, $m(z) = -165.11$, $|b(z)| = 102.93$;
  - NDL, $a(z) = -91.66$, $m(z) = -61.35$, $|b(z)| = 30.31$;
- $\left[35, 35\frac{1}{6}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]$; suspicious measurements:
  - HW, $a(z) = -158.35$, $m(z) = -123.83$, $|b(z)| = 34.52$;
  - kh, $a(z) = -165.09$, $m(z) = -120.27$, $|b(z)| = 44.83$;
- $\left[34\frac{5}{6}, 35\right] \times \left[-115, -114\frac{5}{6}\right]$; suspicious measurements:
  - w, $a(z) = -90.08$, $m(z) = -188.87$, $|b(z)| = 98.79$;

Out of these 13 zones, 10 were originally picked up by an expert as suspicious:

$$\left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right]; \quad \left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{2}{3}, -115\frac{1}{2}\right]; \quad \left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{2}{3}, -115\frac{1}{2}\right];$$

$$\left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]; \quad \left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115, -114\frac{5}{6}\right];$$

$$\left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-115\frac{1}{6}, -115\right]; \quad \left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-115, -114\frac{5}{6}\right]; \quad \left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-114\frac{5}{6}, -114\frac{2}{3}\right];$$

$$\left[35, 35\frac{1}{6}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right].$$

Out of the remaining three zones, two were detected on the first stage of the algorithm and confirmed by the expert to be dirty: $\left[34, 34\frac{1}{6}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]$ and $\left[34\frac{5}{6}, 35\right] \times \left[-115, -114\frac{5}{6}\right]$. The remaining zone $\left[34, 34\frac{1}{6}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]$ does not seem to be suspicious.

Before we proceed with the third stage of the algorithm, let us summarize the results of applying the first two stages.

- We started with 22 dirty zones as indicated by an expert.
- After the first stage of the algorithm,
  - we have detected 3 out of these 22 zones, and,
  - in addition, we detected 23 more dirty zones.

  19 out of the original 22 zones were undetected.
- After the second stage of the algorithm:
  - 10 out of the remaining 19 dirty zones were detected, and
  - in addition, 1 zone was marked as suspicious.

So, after the first two stages of the algorithm,

- 13 out of the original 22 zones were detected;
- 9 out of the original 22 dirty zones still remain undetected.

6.7 Testing the method on the Mojave desert region data set: Part III

On the third stage of our algorithm, we do the following:

- We compute the average A and standard deviation D of the estimated standard deviations $\sigma(z)$ corresponding to different zones.
- We check, for each zone z, whether $\sigma(z)$ is outside the corresponding "k sigma" interval $[A-k\cdot D, A+k\cdot D]$. If it is, we declare this zone to be suspicious.

The average and standard deviation of 1,350 values of $\sigma(z)$ are $A=7.37$ and $D=4.74$. For the Mojave desert zone, $k=5$. Thus, the "k sigma" interval is $[-16.33, 31.07]$. Since the standard deviation $\sigma(z)$ is always non-negative, the only way that a standard deviation can be outside this interval is when $\sigma(z)>31.07$.

Out of the 1,350 zones, the following 8 satisfy this inequality:

$$\left[34, 34\frac{1}{6}\right]\times\left[-115\frac{1}{3}, -115\frac{1}{6}\right];\quad \left[34\frac{1}{2}, 34\frac{2}{3}\right]\times\left[-115, -114\frac{5}{6}\right];\quad \left[34\frac{2}{3}, 34\frac{5}{6}\right]\times\left[-115\frac{1}{6}, -115\right];$$

$$\left[34\frac{2}{3}, 34\frac{5}{6}\right]\times\left[-115, -114\frac{5}{6}\right];\quad \left[34\frac{2}{3}, 34\frac{5}{6}\right]\times\left[-114\frac{5}{6}, -114\frac{2}{3}\right];\quad \left[35, 35\frac{1}{6}\right]\times\left[-115\frac{1}{2}, -115\frac{1}{3}\right];$$

$$\left[35, 35\frac{1}{6}\right]\times\left[-115\frac{1}{3}, -115\frac{1}{6}\right];\quad \left[36\frac{5}{6}, 37\right]\times\left[-113\frac{1}{2}, -113\frac{1}{3}\right].$$

Out of these 8 zones, the following 6 zones were already marked as suspicious on Stage 2 of our algorithm:

$$\left[34, 34\frac{1}{6}\right]\times\left[-115\frac{1}{3}, -115\frac{1}{6}\right];\quad \left[34\frac{1}{2}, 34\frac{2}{3}\right]\times\left[-115, -114\frac{5}{6}\right];\quad \left[34\frac{2}{3}, 34\frac{5}{6}\right]\times\left[-115\frac{1}{6}, -115\right];$$

$$\left[34\frac{2}{3}, 34\frac{5}{6}\right]\times\left[-115, -114\frac{5}{6}\right];\quad \left[34\frac{2}{3}, 34\frac{5}{6}\right]\times\left[-114\frac{5}{6}, -114\frac{2}{3}\right];\quad \left[35, 35\frac{1}{6}\right]\times\left[-115\frac{1}{3}, -115\frac{1}{6}\right].$$

Out of the remaining 2 zones, one was originally marked by an expert as dirty: $\left[36\frac{5}{6}, 37\right]\times\left[-113\frac{1}{2}, -113\frac{1}{3}\right]$, and one zone suspicious seems not to be dirty at all: $\left[35, 35\frac{1}{6}\right]\times\left[-115\frac{1}{2}, -115\frac{1}{3}\right]$.

We also tested the ability of our algorithm to pick up hard-to-detect erroneous measurements. Specifically, we tested our algorithm on a region in which, according to the original expert estimates, out of 1,350 zones, 22 contain hard-to-detect erroneous measurements. Our algorithm detected 39 suspicious zones. Of these 39 zones:

- 14 zones were marked originally by an expert as containing erroneous measurements;
- 23 zones were not originally marked by an expert, but, on close analysis, turned out to contain erroneous measurements;
- 2 suspicious zones turned out, on expert analysis, to be OK.

Out of the original 22 zones, 8 remained undetected.

6.8 Testing the method on the Mojave desert region data set: Part IV

According to the final stage of the algorithm, we take all the zones marked as suspicious on the first three stages, find the connected areas consisting of at least three zones, and declare:

- all the zones which are direct neighbors to these ones as "somewhat suspicious", and
- all the zones which are diagonal neighbors to these ones as "mildly suspicious".

As a result, the following 18 zones are declared "somewhat suspicious":

$$\left[34, 34\frac{1}{6}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right]; \quad \left[34, 34\frac{1}{6}\right] \times \left[-115\frac{2}{3}, -115\frac{1}{2}\right]; \quad \left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-116, -115\frac{5}{6}\right];$$

$$\left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]; \quad \left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]; \quad \left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right];$$

$$\left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{1}{6}, -115\right]; \quad \left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115, -114\frac{5}{6}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{2}{3}, -115\frac{1}{2}\right];$$

$$\left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{1}{6}, -115\right];$$

$$\left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-114\frac{5}{6}, -114\frac{2}{3}\right]; \quad \left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]; \quad \left[34\frac{2}{3}, 34\frac{5}{6}\right] \times \left[-114\frac{2}{3}, -115\frac{1}{2}\right];$$

$$\left[34\frac{5}{6}, 35\right] \times \left[-115\frac{1}{6}, -115\right]; \quad \left[34\frac{5}{6}, 35\right] \times \left[-114\frac{5}{6}, -114\frac{2}{3}\right]; \quad \left[35, 35\frac{1}{6}\right] \times \left[-115, -114\frac{5}{6}\right].$$

Out of these 18 zones, 7 were originally marked by an expert as containing erroneous measurements:

$$\left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]; \quad \left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]; \quad \left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right];$$

$$\left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{2}{3}, -115\frac{1}{2}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{1}{2}, -115\frac{1}{3}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right];$$

$$\left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{1}{6}, -115\right].$$

The remaining 11 zones turned out to be OK.

The following 10 zones were declared to be "mildly suspicious":

$$\left[34, 34\frac{1}{6}\right] \times \left[-116, -115\frac{5}{6}\right]; \quad \left[34\frac{1}{6}, 34\frac{1}{3}\right] \times \left[-115\frac{1}{6}, -115\right]; \quad \left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-116, -115\frac{5}{6}\right];$$

$$\left[34\frac{1}{3}, 34\frac{1}{2}\right] \times \left[-114\frac{5}{6}, -114\frac{2}{3}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right]; \quad \left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-114\frac{2}{3}, -114\frac{1}{2}\right];$$

$$\left[34\frac{5}{6}, 35\right] \times \left[-115\frac{1}{3}, -115\frac{1}{6}\right]; \quad \left[34\frac{5}{6}, 35\right] \times \left[-114\frac{2}{3}, -114\frac{1}{2}\right]; \quad \left[35, 35\frac{1}{6}\right] \times \left[-115\frac{1}{6}, -115\right];$$

$$\left[35, 35\frac{1}{6}\right] \times \left[-114\frac{5}{6}, -114\frac{2}{3}\right].$$

Out of these 10 zones, 1 was originally marked by an expert as containing erroneous measurements: $\left[34\frac{1}{2}, 34\frac{2}{3}\right] \times \left[-115\frac{5}{6}, -115\frac{2}{3}\right]$. The remaining 9 zones turned out to be OK.

6.9 General conclusions

First, we tested the ability of our algorithm to pick up crude measurement errors. Specifically, we tested our algorithm on a region with simulated (seeded) crude erroneous measurements. As a result of our test:

- all the seeded erroneous measurements were successfully detected, and
- none of the non-erroneous measurements were marked as suspicious.

We also tested the ability of our algorithm to pick up hard-to-detect erroneous measurements. Specifically, we tested our algorithm on a region in which, according to the original expert estimates, out of 1,350 zones, 22 contain hard-to-detect erroneous measurements. Our algorithm detected 39 suspicious zones, 18 somewhat suspicious zones, and 10 mildly suspicious zones.

Of the 39 suspicious zones:

- 14 zones were marked originally by an expert as containing erroneous measurements;
- 23 zones were not originally marked by an expert, but, on close analysis, turned out to contain erroneous measurements;
- 2 suspicious zones turned out, on expert analysis, to be OK.

Out of the 18 somewhat suspicious zones:

- 7 zones were marked originally by an expert as containing erroneous measurements;
- 11 somewhat suspicious zones turned out, on expert analysis, to be OK.

Out of the 10 mildly suspicious zones:

- 1 zone were marked originally by an expert as containing erroneous measurements;
- 9 mildly suspicious zones turned out, on expert analysis, to be OK.

Overall:

- Out of 22 originally marked zones with erroneous measurements, all were successfully detected.
- In addition to these 22 zones, 23 new zones with erroneous measurements were detected.

Overall, out of 67 zones marked as suspicious, somewhat suspicious, or mildly suspicious, 45 (more than two third) turned out to be actually dirty.

7 Future Work

7.1 Detecting large differences between neighboring data points

In addition to using statistical techniques for eliminating erroneous measurements, we can also detect suspicious data by simply comparing the measurements at two neighboring points (this idea was, in effect, described in (Scott, 1994)). If the ratio between this difference and the distance exceeds a certain threshold (decided by experts), then it is highly probable that one of these measurements is erroneous. This approach capitalizes on the fact that there are physical limits on the horizontal gradients in some types of data. Preliminary results of using this idea are presented in (Coblentz et al., 2000).

7.2 From eliminating to correcting erroneous measurements

In this paper, we mainly addressed the issue of eliminating erroneous measurements in a set of point data. One of the reasons why a measurement would be declared erroneous is because we have detected a significant bias in the measurements coming from the corresponding data source. As we have mentioned, in Section 1.1, this bias may come from the bias in the instrument calibration, or, for gravity measurements, from the erroneous base station values. In this case, instead of simply *eliminating* the erroneous measurements, it may be more advantageous to *correct* them by correcting for this bias.

This can be done in a manner similar to how bias is corrected in astronomy when we combine several catalogs into a single one (e.g., Podobed, 1965). Specifically, to correct the bias, we can do the following:

Select all data sources which contain at least 50 measurements in the given geographical region. For each of these data sources, we do the following:

1. We mark all the zones that contain measurements coming from this particular data source. For each such zone (subregion) z, we:

    i) compute the average $m(z)$ of all the measurements in the zone z which come from this data source;

    ii) compute the bias $b(z) = m(z) - a(z)$ (where $a(z)$ is the average of all the measurements performed in this zone).

2. Then, we compute the bias b of this data source as an average of all the biases $b(z)$ for all the zones z which contain measurements from this data source.

3. Subtract this average bias b from all the measurements coming from this particular data source.

This correction may change the averages, so it makes sense to repeat this procedure several times until the corrected values stop changing.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209 and grant NCC 2-1232, by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant number F49620-00-1-0365, by Grant No. W-00016 from the U.S.-Czech Science and Technology Joint Fund, and by Grant NSF 9710940 Mexico/Conacyt.

The authors are thankful to Ra'ed Aldouri, Mark R. Baker, Alan Mabry, Carlos Montana, Joan Staniswalis, Hongjie Xie (University of Texas at El Paso), and to Lauren M. Scott (ESRI, California) for their help and advice.

References

Adams, D.C., Keller, G.R., 1996. Precambrian basement geology of the Permian Basin region of West Texas and eastern New Mexico: A geophysical perspective, American Association of Petroleum Geologists Bulletin 80, 410-431.

Anselin, L., Getis, A., 1992, Spatial Statistical Analysis and Geographic Information Systems, The Annals of Regional Science 26, 19–33.

Birt, C.S., Maguire, P.K.H., Khan, M.A., Thybo, H., Keller, G.R., Patel, J., 1997. The influence of pre-existing structures on the evolution of the southern Kenya Rift Valley: Evidence from seismic and gravity studies, Tectonophysics 278, 211–242.

Braile, L.W., Hinze, W.J., Keller, G.R., 1997. New Madrid seismicity, gravity anomalies, and interpreted ancient rift structures: Seismol. Research Letters 67, 599–610.

Coblentz, D.D., Keller, G.R., Kreinovich, V., Beck, J., Starks, S.A., 2000. Interval Methods in Remote Sensing: Reliable Sub-Division of Geological Areas, Abstracts of the 9th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN'2000, held jointly with the International Conference on Interval Methods in Science and Engineering Interval'2000, Karlsruhe, Germany, September 19–22, 2000, 41.

Cordell, L., Keller, G.R., 1982. Bouguer Gravity Map of the Rio Grande Rift, Colorado, New Mexico, and Texas Geophysical investigations series, U.S. Geological Survey.

Cordell, L., Zorin, Y.A., Keller, G.R., 1991. The decompensative anomaly and deep structure of the Rio Grande rift, Journal of Geophysical Research 96, 6557–6558.

Devore, J., Peck, R., 1999. Statistics: the Exploration and Analysis of Data, Duxbury, Pacific Grove, California.

FGDC Federal Geographic Data Committee, 1998. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998), Federal Geographic Data Committee, Washington, D.C., http://www.fgdc.gov/metadata/contstan.html

Fliedner, M.M., Ruppert, S.D., Malin, P.E., Park, S.K, Keller, G.R., Miller, K.C., 1996. Three-dimensional crustal structure of the southern Sierra Nevada from seismic fan profiles and gravity modeling, Geology 24, 367–370.

Goodchild, M., Gopal, S. (Eds.), 1989. Accuracy of Spatial Databases, Taylor & Francis, London.

Grauch, V.J.S., Gillespie, C.L., Keller, G.R., 1999. Discussion of new gravity maps of the Albuquerque basin, New Mexico Geol. Soc. Guidebook 50, 119–124.

Heiskanen, W.A., Meinesz, F.A., 1958. The Earth and its gravity field, McGraw-Hill, New York.

Heiskanen, W.A., Moritz, H., 1967. Physical Geodesy, W.H. Freeman and Company, San Francisco, California.

Iglewicz, B., Hoaglin, D.C., 1993. How to Detect and Handle Outliers, American Society for Quality Control, Statistics Division, Milwaukee, Wisconsin.

Keller, G.R., 2001. Gravitational Imaging, In: The Encyclopedia of Imaging Science and Technology, John Wiley, New York (to appear).

Kennedy, S., 1989. The small number problem and the accuracy of spatial databases, In: Goodchild, M., and Gopal, S. (Eds.), Accuracy of Spatial Databases, Taylor & Francis, London, 187–196.

McCain, M., William C., 1998. Integrating Quality Assurance into the GIS Project Life Cycle, Proceedings of the 1998 ESRI Users Conference. http://www.dogcreek.com/html/documents.html

Podobed, V.V., 1965. Fundamental astrometry: determination of stellar coordinates, University of Chicago Press, Chicago.

Rabinovich, S., 1993. Measurement Errors: Theory and Practice, American Institute of Physics, New York.

Rodriguez-Pineda, J.A., Pingitore, N.E., Keller, G.R., Perez, A., 1999. An integrated gravity and remote sensing assessment of basin structure and hydrologic resources in the Chihuahua City region, Mexico, Engineering and Environ. Geoscience 5, 73–85.

Scott, L., 1994. Identification of GIS Attribute Error Using Exploratory Data Analysis, Professional Geographer 46(3), 378–386.

Sharma, P., 1997. Environmental and Engineering Geophysics, Cambridge University Press, Cambridge, U.K.

Simiyu, S.M. Keller, G.R., 1997. An integrated analysis of lithospheric structure across the East African Plateau based on gravity anomalies and recent seismic studies, Tectonophysics 278, 291–313.

S-Plus, 1989. User's Manual, Seattle, Washington.

S-Plus, 1991. Guide to Statistical and Mathematical Analysis, Seattle, Washington.

Tesha, A.L., Nyblade, A.A., Keller, G.R., Doser, D.I., 1997. Rift localization in suture-thickened crust: Evidence from Bouguer gravity anomalies in northeastern Tanzania, East Africa, Tectonophysics, 278, 315–328.

Wadsworth, H.M. Jr., ed., 1990. Handbook of Statistical Methods for Engineers and Scientists, McGraw-Hill Publishing Co., New York.