TOWARDS AUTOMATIC DETECTION OF ERRONEOUS MEASUREMENT RESULTS IN A GRAVITY DATABASE

QIAN WEN, ANN Q. GATES, JAN BECK,
VLADIK KREINOVICH, and G. RANDY KELLER

NASA Pan-American Center for Earth and Environmental Studies
University of Texas, El Paso, TX 79968, USA, vladik@cs.utep.edu

Abstract

Geospatial databases often contain erroneous measurements. For some such databases such as gravity databases, the known methods of detecting erroneous measurements – based on regression analysis – do not work well. As a result, to clean such databases, experts use manual methods which are very time-consuming. In this paper, we propose a (natural) "localized" version of regression analysis as a technique for automatic cleaning. We illustrate the efficiency of this technique on the example of the gravity database.

Keywords

Gravity database, Erroneous measurement, Automatic detection

1 Introduction

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., [7]. Each type of data is usually stored in a large geospatial database. Based on these measurements, geophysicists generate maps and images and adjust geophysical models which fit these measurements. For example, the geophysical use of gravity databases is described, e.g., in [3].

The main problem with the existing geospatial databases is that they are known to contain many erroneous points; see, e.g., [2], [4], [6]. For example, there are several reasons why gravity measurements can be erroneous:

- there can be measurement errors in gravity and in elevation;
- there can be transcription errors;
- there may be an error in the instrument calibration;
- finally, there may be base station problems (gravity measurements are always relative to some "known" value).

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

The main goal of this research was to "clean" a gravity database. At present, the cleaning of gravity databases is done mainly "by hand", by a professional geophysicist looking both at the raw measurement data, at the preliminary results of processing these raw data, and at other types of information such as geological maps. There are many useful map overlay and statistical techniques to help with this manual analysis (see, e.g., [6]), 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.

2 Case Study: Gravity Database

Gravity measurements are one of the most important sources of geophysical and geological information. There are two reasons

Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference

for this importance. First, in contrast to more widely used geophysical data (like ultrasound waves) which mainly reflect the conditions on the Earth's surface, gravitation comes from the whole Earth (see, e.g., (Heiskanen and Meinesz, 1958), (Heiskanen and Moritz, 1967)). and thus, contains, in particular, 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, reflects also the areas which are not well covered by more traditional geophysical methods.

Since the gravity value is determined by the integral of the Earth's mass distribution over a large area, the measured value is almost the same in all the places. If we take into consideration natural physical differences caused by the difference in latitude and elevation, we get an almost perfect description 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 *Bouguer anomaly* (BA).

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).

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

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. 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 the simplest possible case, y does not depend on any of the known physical quantities $p, \ldots, q$, i.e., the function $f(p, \ldots, q)$ is simply a constant).

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., [1]. 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.

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 $K = 5$.

4 Testing Traditional Regression Techniques on Gravity Database

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 of two good measurements.

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 confirms that the above method is not working well.

5 Towards Localized Regression Techniques: Main Idea

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.

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.

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.

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., [5]. 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.

6 Second Idea: 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 [6] 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.

7 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 $\sigma(z)$ orresponding 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.

8 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 suspicious 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 which 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.

9 Localized Regression Analysis as a Method for Detecting 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

$$[a(z) - k \cdot \sigma(z), a(z) + k \cdot \sigma(z)]$$

5 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. If $N(z) \geq 50$, then we further subdivide the zone z into approximately square sub-subregions (subzones) each of which contains, on average, approximately 50 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 which 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, and zones marked as mildly suspicious.

10 Testing the Proposed Method on the Actual Gravity Database

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

For Mojave desert region, 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 were marked originally by an expert as containing erroneous measurements, 23 were not originally marked by an expert, but, on close analysis, turned out to contain erroneous measurements, and 2 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.

11 Future Work

1) In addition to using statistical techniques for eliminating erroneous measurements, we can also some preliminary detection by simply comparing the measurements in two neighboring points (this idea was, in effect, described in [6]. 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. Preliminary results show the prospectiveness of this idea.

2) Instead of simply *eliminating* biased 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. Specifically, to correct the bias, we can select all data sources which contain at least 50 measurements in the given geographical region. For each of these data sources, for each zone z which contain measurements coming from this particular data source, we compute the average $m(z)$ of all the measurements in the zone z which come from this data source, and then compute the bias $b(z) = m(z) - a(z)$ (where $a(z)$ is the average of all the measurements performed in this zone). 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.

To correct the measurements, we subtract this average bias b from all the measurements coming from this particular data source. This correction may change the averages $a(z)$, 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 grants NCC5-209 and NCC 2-1232, by the Air Force Office of Scientific Research 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 advise.

References

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

[2] M. Goodchild and S. Gopal, Accuracy of Spatial Databases, Taylor & Francis, London, 1989.

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

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

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

[6] L. Scott, Identification of GIS Attribute Error Using Exploratory Data Analysis, Professional Geographer, 1994, Vol. 46, No. 3, pp. 378–386.

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