Error Analysis in Cartographic Data with Application to the Geographic Information Systems

Fabián D.Barbato

Universidad de la Republica, Facultad de Ingenieria, Dept. of Geomatics, Calle Julio Herrera y Reissig 565, Montevideo, Uruguay, fbarbato@fing.edu.uy

Abstract

A map, as a product destined to bring information for any purpose, must be reliable, not only from the point of view of the associated thematic information that it contains and expresses, but also from the geometric precision with which every geographic object is identified and linked with the rest of the objects.

This work develops theoretical concepts and a specific software to control the quality of a map, led to the determination of geometric and topologic accuracies , based on statistical testing concepts, and determination of algorithms that produce parameters of acceptance or rejection of the tested cartography.

This investigation covers an application work of these algorithms to the study of the cartography 1:50000, basis of the National Geographic System of Uruguay, that includes the points taking with G.P.S. on field, and four sources of comparative digital cartography for the extraction of coordinates of punctual entities.

The observations involved in this study, include a refinement treatment through statistical tests and robust estimators.

Keywords: cartography, GIS, errors, accuracy, testing

1 Introduction

The study of the performance of errors in geomatics defines the observational magnitudes to be treated as mathematic variables, that is to say that are able to take on numeric value in the real field. If we include to these variables, performances based on the probabilistic mathematic, that is to say a representative sample elements of a bigger universe, those variables become “random variables”.

Indeed, this work studies the performance of determined variables, such as, it coordinates (and its differences) that define objects on a map, through sample sets, as in practice becomes impossible to range the infinite points contained on a map.

2 Estimators Used in Geomatics

For the determination of accuracy parameters for GIS basis cartography, it is necessary to set out what kind of estimators and functions associated to them we will use.

We must remember that in every case, the determination of these parameters is carried out over a rigid probabilistic mathematic base, for what it becomes necessary to be completely sure of the use of certain estimators, stating and controlling their consistence and reliability.

We will state as a statistical hypothesis, supported by work experiences and studies, that the distribution that best fits to the verification of geometric discrepancies between the "real world" and diverse cartographic supports, is the (N) Normal Distribution.

Being X a continuous random variable, whose probability or density function is (f), and whose domain is a given sample. A sample of (n) size means a set x_1, x_2, x_3,....x_n of X values. To make that sample "useful" or representative of the universe, we must have taken it by chance with the density or probability function (f).

That means:

1. each one of the xi values can be considered as a value of a new random variable Xi that has the same function (f) as X.

$$\Rightarrow E(Xi) = E(X_2) = ... = E(X_n) \quad \lozenge \quad \sigma^2(Xi) = \sigma^2(X_2) .... \forall\ i=1.........n.$$

2. the random variables Xi are independent or are not correlated.

The problem of making a correct choice of samples, fulfilling the conditions above, is not an easy task to carry out. From that n size sample, we get a "sample mean"($\overline{X}$) and a "sample variance"(S^2), that will represent the rest of the total universe of the infinite values that could take X. The random variables X_1,X_2,X_3,....Xn come as a result of the samples, and their value are the same values of the sample.Each one of these variables functions, that will not depend on unknown parameters, is defined as "Statistical". When one of these statisticals is used to "estimate" the value of any characteristic of population, it is called "Estimator" of that characteristic. Since the estimator is a statistical, it is a random variable as well.

Suppose a probability or density function that depends on a parameter $\theta \Rightarrow f(x,\theta)$. The problem of the parameters estimation is to deduce with certain probability the value of θ from samples taken in the population. The method of the “point estimation” consists of choosing a function from the sample data, whose value, with certain probability, can be taken as the θ value. The chosen function is a statistical, certainly called as θ estimator.

The method of “interval estimation” consists however on deducing, from the sample, an interval in which must be found with certain probability, the θ value.

2.1 Detection of “Outlier” or “Blunder” Errors and Systematic Errors

We classify the errors in random errors, systematic errors and outlier or blunder errors. Optimum estimators and diverse techniques of observation and adjustment, are thought and designed under the assumption that the observations have exclusively random or stochastic errors.

Mathematically, the systematic errors, are seen as different types of divert, making that “expecting mathematic” of the error size be different from zero.

In the geodesic and topographic measuring fields, the systematic errors normally come from bad calibration of the instruments to be used, environmental effects, personal equations of the observer, etc.The most reliable way to detect the existence of systematic errors, is through statistical tests and the application of robust estimators .

We mathematically understand by blunder errors, those random errors that result from 3 to 20 times bigger than expected standard errors, according to pre-established precision and tolerance.

Any form of preventing the errors occurrence, results insufficient to dismiss the existence of blunder errors specially in samples of big amount of observations.One of the most important problems in the probabilistic and statistic theory, is the later detection in the observational work, of blunder errors through the analysis of residuary.

The modelling knowledge of the observational errors performance help us to set up the procedure rules for the detection of not-random errors.

1. Starting from the premise that the observational errors are random errors with a specific associated distribution; then this is the hypothesis $H_{0.}$ The alternative hypothesis H_1 is just that the error will not be random errors according to an associated distribution.
2. We can build derivative variables that have the same distribution, according to the previous condition.
3. Finally, we must test the values taken from the previous statistical (samples), against the critical theoretic values for a certain level of significance or “risk” α.

If the hypothesis testing is accepted, we can definitely take in consideration the (1.) point. Otherwise, we must suspect the existence of blunders errors in the observational data set.

This testing can be directly applied on the observation or after a preliminary adjustment by Least Squares.

2.2 Tests for the Detection of Blunder Errors

Before stating when the errors coming from measurements have random performances or not, we must know the properties of random errors.

- The arithmetic mean ε_i must approximate to zero when the number of observations (n) is big enough.
- Positive or negative sign errors have the same chance of occurrence.
- Short magnitude errors have a bigger probability of occurrence than the absolute great magnitude errors.
- Under certain measurement conditions, the absolute magnitude of the errors must be within certain limits.

Taking the properties mentioned above in consideration, we can build many statisticals to test whether they are random errors or not.

Here we state in a summarised way, some five different kinds of tests, being the last one among the most important ones, as it allows to find a mistake or blunder error: Test of number of positive errors against negative ones, Test of the order of negative errors against positive ones, Test sum of squares of the negative and positive errors, Test sum of errors, Test maximum absolute value of the error.

2.3 Robust Estimators

Robust estimators are those ones that become insensitive to the limited variations of the distribution functions, for example, in case of blunders errors occurrence.

These types of estimators are based on other models or techniques different from the concept of the Least Squares.This topic has become crucial in the whole area of studies in the quality control of geographic data, and this is showed by a variety and increase in the articles published nowadays about the topic.

2.4 Regression Diagnostics

One of the robust estimators applied in this study, is the "Regression Diagnostics".

The Regression Diagnostic supposes a preliminary adjustment through Least Squares, and process in the following:

- It is initially determined an adjustment of the whole data through minimum squares.
- The residual is computed for each observation.
- All the observations that do not fulfil the minor conditions will be defeated.
- A new adjustment is determined with the reminding observations.

The success of this process depends on the quality of the initial adjustment, and does not guarantee a correct result, but as well as the GRIT, this process works very well with a moderate percentage of blunder errors.

2.4 Great Residuals Iterative Test (G.R.I.T.)

The robust estimator “Regression Diagnostics”, and a variation in the previous technique, is the G.R.I.T. F.Barbato (1998), highly efficient in data sets of n $\approx$ 30 as a first order approximation.

The main idea of this test, from the residuals calculation [$V_i = L_i - \overline{X}$], is to arrange them from greater to minor through absolute magnitude order.

This model presupposes the existence of fewer blunder errors.($\cong<$ 3%).

Before determining the residuals, it is necessary the identification and inmediate elimination of mistakes that can involve, for example, the variation of a major or minor order of the power in 10^i of the measures. This will help that the initial mean calculation will not look seriously distorted.

If it is taken the greater (V_f), (could be more than one), and it is important to control that its residual does not exceeds a certain tolerance value (ψ_f) according to the pre- established conditions, methodology, instruments, etc.

In our case, where the test is carried out taking as random variable the “differences” between the coordinates determined on field with the ones taken through varied procedures from cartographic supports, ψ_f will be defined as a function of the scale *(e)* of graphical representation of the map $\psi_f = F(e)$ in case of hard supports, and resolution function *(r)* or pixel size, in case of digital support $\psi_f = F(r)$.

The corresponding measurement to that residual is eliminated from the data set, and the residuals with the new measures are re-calculated, keeping the statement of tolerance, but for the consecutive cases, a smoothing rate between 10% to 20% is established.This is pointing out we must make the best use of the limited quality of observations, leaving for a next stage the determination of the set consistence with the normal distribution.

We will Classify as a “new suspect” of blunder error, that residual which most strays from [$1.20*\psi_f$]. The (1.20) factor has the aim to create a smoothing interval to make possible the adoption to samples of n<30, where the elimination of observations can degrade the density function [f(x)] and associated distribution properties.

It has been experimentally proved that eliminating the biggest blunder error, the model gets “extremely severe” with the reminding “outliers”.

This iteration is carried out until the quantity of blunder errors will not exceed the pre-established limit, which means that the blunder errors would be related to “abnormal” variations or disorder in the measuring process, being necessary to check and start again with the measuring process.

Summarising, the GRIT estimator suggests:

1. [$V_i = L_i - \overline{X}$] ordered from major to minor
2. discarding of measures different in $>10^i$ ($i >= 1$)
3. determination of ψ_f
4. rejection of V_f measures that do not fulfil the condition
5. re-calculation of the values

To complete the procedure, after refinements accepted by the GRIT process, we are ready to go on with the "verification of systematisms" test and the distribution control , associated to the "edge" blunder errors whose determination is not clear or definitive.

From these five tests, a sample rejection by only one of the techniques will be enough to make us check the measurement values.

3 Cartographic Quality Control

One of the main aspects in the study of a GIS implementation, is to know the quality and accuracy of the information that will supply the system. One of the principal data-bases is the cartography.

The most frequent case, is to pre-dispose an "existent" cartography created by our institution, or a cartographic base externally obtained.

Due to the high costs of generation of new cartographic bases, and the easy access to other pre-existent bases, it will be necessary to justify correctly, from the technical point of view, whether it is convenient or not to implement a GIS from this cartography, taking in consideration towards which applications or objectives will be oriented our GIS.

Leaving some thematic aspects aside, we will concentrate our analysis on the accuracy and geometric exactness as well as topologic.

We define geometric accuracy (PG) as the parameter associated to a (N) distribution, that expresses in a consistent way the discrepancies, with a certain grade of reliability among the object positions (points, arcs, polygons and areas) obtained from the cartography in hard support as well as digital, and the entities positions in the "real world", determined from field geodesic measurements, specially using the GPS technology.

3.1 Statistical Testing for (N)

In terms of statistical hypotheses it is possible to state the following conclusions:

I. H0 is accepted when H0 is certain
II. H0 is rejected when H0 is certain
III. H0 is accepted when H0 is false
IV. H0 is rejected when H0 is false

I and IV are "correct" options
II is called "Type I Error"
III is called "Type II Error"

$\{\mu_{.0} = data\}$

α = P[reject H_0 when H_0 is certain] = Type I Error

$$\begin{cases} H_0 : \mu = \mu_{.0} \\ H_1 : \mu \neq \mu_{.0} \end{cases} \qquad \text{(1-α=reliability degree)}$$

To be able to control a map, we define our random variable (X) as the difference between coordinates captured on field (GPS) and coordinates taken from a determined map, so that:

$[\,X\,] \equiv \Delta X$

$$[\,\overline{X}\,] \equiv \Delta\overline{X} = \frac{1}{n}\sum_{i=1}^{i=n} \Delta X_i$$

These are our new random variables ⇒ ($\Delta\overline{X}$ and ΔX)

$$\left\{\begin{matrix} X \to N\left(\mu, \sigma^2\right) \\ \overline{X} \to N\left(\mu, \sigma^2/n\right) \end{matrix}\right\}$$

$$\to z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to$$

$$\boxed{\to z_0 = \frac{\overline{X} \;\; - \mu_0}{\sigma_0/\sqrt{n}} \to}$$

$$\text{P}\left[-z < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z\right] = 2.\phi(z) - 1 \Rightarrow$$

$$\text{P}\left[\overline{X} - \frac{z.\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{z.\sigma}{\sqrt{n}}\right] = 2.\phi(z) - 1 \to$$

“ Mean Probability“

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z \rightarrow \overline{X}-\mu < \frac{z.\sigma}{\sqrt{n}} = C \rightarrow$$

$$\mathrm{P}\left[\left(\mu_{.0}-C\right) < \overline{X} < \left(\mu_{.0}+C\right)\right] = 2.\phi(z)-1 \quad \Rightarrow$$

- $|z_0| \leq z(2.\phi(z)-1)$ *"accepted"*
- $|z_0| > z(2.\phi(z)-1)$ *"rejected"*

3.2 Variance Testing (interval)

Determining the sample variance, we state the statistical hypothesis in a certain interval.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{i=n}\left(x_i-\overline{X}\right)^2 \quad \begin{cases} \mathrm{H}_0 : \sigma^2 = \sigma_0^2 \\ \mathrm{H}_1 : \sigma^2 \neq \sigma_0^2 \end{cases}$$

$[\,\mathrm{X}\,] \equiv \Delta\mathrm{X}$

$$[\,\overline{X}\,] \equiv \Delta\overline{X} = \frac{1}{n}\sum_{i=1}^{i=n}\Delta X_i$$

It's possible to define an estimator $\mathrm{Y} = \frac{n-1}{\sigma^2}.S^2$ with *n*-1 degrees of freedom

⇒ we build a confidence interval:

$$\mathrm{P}\left[\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,1-\frac{\alpha}{2}}\right] =$$

$$\mathrm{P}\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] = 1-\alpha$$

$$\left\{ \chi_0^2 = \frac{(n-1)S^2}{\sigma_{.0}^2} \right\} \text{ if } \left\{ \begin{matrix} \chi_0^2 < \chi^2_{n-1,\frac{\alpha}{2}} \\ \chi_0^2 > \chi^2_{n-1,1-\frac{\alpha}{2}} \end{matrix} \right\}$$

is *rejected.*

$$\left. \mathrm{P}\left[\frac{\chi^2_{n-1,\frac{\alpha}{2}}.\sigma_{.0}^2}{n-1} < S^2 < \frac{\chi^2_{n-1,1-\frac{\alpha}{2}}.\sigma_{.0}^2}{n-1} \right] = 1-\alpha \right\} \text{ is } \textit{accepted}.$$

3.3 Variance Testing (semi-interval)

In other cases we need to know if a certain variance fulfils the condition for a ≤ value.

$$\begin{cases} \mathrm{H}_0 : \sigma^2 \geq \sigma_{.0}^2 \\ \mathrm{H}_1 : \sigma^2 < \sigma_{.0}^2 \end{cases} \qquad \chi_0^2 = \frac{(n-1)S^2}{\sigma_{.0}^2}$$

but we must use:

$$\begin{cases} \mathrm{H}_0 : \sigma^2 \leq \sigma_{.0}^2 \\ \mathrm{H}_1 : \sigma^2 > \sigma_{.0}^2 \end{cases}$$

Given the σ_0^2 tolerance, we must determine whether our map complies

or not with the pre-established accuracy.

$\Rightarrow \chi_0^2 = \frac{(n-1)S^2}{\sigma_{.0}^2} \leq \chi^2_{n-1,1-\alpha}$ is *accepted* → the map complies with the "pre-established accuracy".

4 Algorithm of Map Quality Control

This algorithm of quality control is thought for the geometric control. The methodology stated here is the following:

- "Bias" control.
- If there are no systematisms or being them corrected, acceptation or rejection of the calculated parameters of accuracy against the theoretical ones derived from the distributions.
- In case of being approved, determination of a mean accuracy for the universe of points.

The base of the algorithm is to design a GPS network of control points, perfectly individualised on the field as well as on the map (geodetic points, route crosses, building, vertex).In the design of the control points selection, it must be taken in consideration the distribution and density degree of the points, as uniform as possible, and quantity of points should not be under 20.According to the specifications of the FGDC-STD-007.3-1998, the points must be separated in a 10% of the diagonal length of a rectangle that covers the area of study, and at least the 20% of the points must contemplate all the quadrants. The observations either of maps or of field are submitted to strict controls of quality and refinements, with the purpose of assuring a data set that expresses the differences between the sample values without any kind of foreign interference to the process.

Calculating the differences between "field" coordinates and "map" coordinates, we define a new random variable that will be submitted to all the mentioned process in this work. It will be determined the existence or not of bias, and the acceptation or rejection of the general accuracy of the map with regard to required accuracy by the user.

The algorithm gives a similar treatment to the X and Y coordinates, so the application is exactly the same.

5 Algorithm Application in Cartography 1:50000 of Uruguay

One of the main objectives of this work is the application and evolution of this algorithm and all the considerations in concrete examples of the GIS Base Cartography operations in Uruguay.

This project, takes as a base the scale cartography 1:50000 designed in hard support by the Military Geographic Service of Uruguay, covering the totality of the national territory.

5.1 Result of the QCGIS application.

The application of the algorithm in the case of the K27 map, from the Military Geographic Service, results in:

- The map in paper support does not comply with the testing which corresponds to the systematism detection , nor with the required accuracy.
- If we do the same operations with the rest of the supports IMG, DWG, SHP, we will obviously get similar results.
- The mean accuracy of this map is found about 100 meters, which would mean a "virtual" scale of::
  1:250,000 for the 95% of the events
  1:100,000 for the 68% of the events

The importance of the systematisms detection, implies that it would not be correct to work with the parameters and estimators of the (N), for what the algorithm must be cut short on detecting a systematism in the control

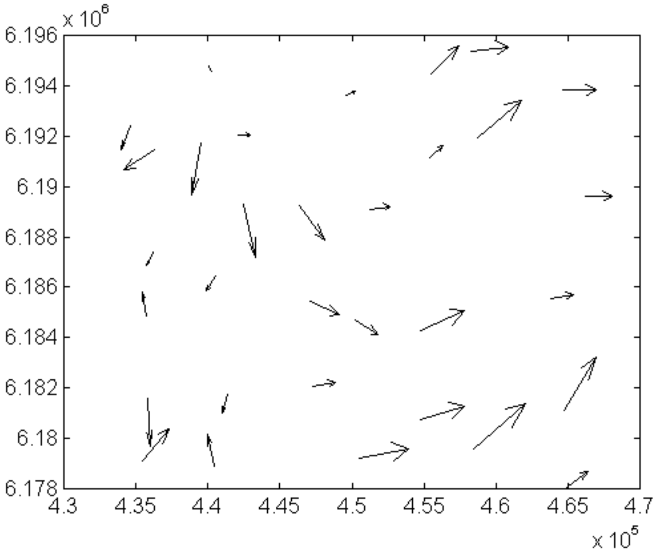

Fig. 1. Vector of Differences

6 Conclusions

The aim of this work has been the determination of robust and consistent algorithms to control the accuracy in cartography base of GIS.After their design, these algorithms have been proved in groups of real data.

The algorithm, including the treatment of observation, detection and refining of bundle errors, statistic tests and specific software support, has been tested, getting excellent results, being each one verified in the “real world”.It is important to highlight as well, that this algorithm can be applied to any cartographic support including measurement maps, with which it becomes a very useful implement, particularly to the Cadastre and Council Offices.The data processed shows that the accuracy parameter clearly falls within the range of insufficiency” strongly influenced by the poor cartographic quality used as a base of the system.

References

Barbato F, (2001). *First International Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS.* IAG, SSG 4.190, ISBN-3-906467-29-25, Institute of Geodesy and Photogrammetry of Zurich ETH.

Deakin RE ,(1996). *The Australian Surveyor*, Dept. of Land Information, Royal Melbourne Institute of Technology.

Fan Huaan, (1997). Theory of Errors, *K.T.H.*-ISBN 91- 7170-200-8.

Featherstone,WE, (1998). *Geomatics Research Australasia, N.68*, ISSN 1324-9983.

Hofmann-Wellenhof and J. Collins (1992). *GPS Theory and Practic*e. Springer Verlag, Wien.

Lembo J and Hopkins P, (1998). The use of Adjustment Computations in Geographic Information Systems for Improving the Positional Accuracy for Vector Data, *S& LIS, Vol 58 No.4*.

Mikhail E.M Gracie G, (1981), Analysis and Adjustment of Survey Measurements. Van Nostrand Reinhold, New York.

Sevilla M.J.,(1996), Criterios de Precisión Cartográfica, Monografía,Topografía y Geodesia,España.

Sjöberg Lars, (1983). Unbiased Estimation of Variance-Covariance Components in Condition Adjustment with Unknows-A MINQUE Approach *,ZFV ,108*.

Stokes G,(1949). *On the Variation of the Gravity on the Surface of the Earth*.Trans.Cambridge Phil.