# ROBUST PROCEDURES FOR GIS DATA HOMOGENIZATION

D. Fritsch, F. Crosilla

Chair for Photogrammetry and Remote Sensing/Istituto di Urbanistica et Pianificatione
Technical University Munich/Universita degli Studi di Udine
Arcisstr. 21, D-8000 Munich 2, Germany/Via Larga, 42, I-33100 Udine, Italy
Tel.:+49-89-2105 2671; Fax +49-89-2809573; Telex: 522854 tumue d
email:dieter@photo.verm.tu-muenchen.de

Commission III

## ABSTRACT

Vectorial data acquisition for Geographic Information Systems (GIS) is a real bottleneck which is to be overcomed by a combination of results of surveying, photogrammetry, and map data digitization. The homogenization problem consists of consistency checks in first part with the more accurate data set, therefore math models must be developed to decide on data acceptance and rejection respectively.

The paper introduces with overall accuracy measures for the three data acquisition methods. Its main part solves the mathematical problem when all three data sources are joined together. The corresponding linear models and hypothesis tests are shown. It concludes with pros and cons if different objective functions ($L_1, L_2, L_\infty$) are used for parameter estimation.

**Key words:** GIS data acquisition, data homogenization, math models, hypothesis tests, objective functions.

# 1 Introduction

Geometric data acquisition for Geographic Information Systems (GIS) can be done by different methods of surveying, photogrammetry and cartography. This process is driven by two main parameters: costs and accuracy which are depending on each other. In order to fill the databases of a GIS very fast maps are digitized and preprocessed to fit into a reference frame of control points, to overcome isolated mapping regions, and to realize constraints such as straight lines, perpendicularity and others. Map digitization is cheap in terms of acquisition time but bad in accuracy. It can considerably be improved when photogrammetry and surveying deliver a set of control points by means of photogrammetric restitution, tacheometry and GPS, as it is well-known.

In this context the homogenization process consists of similarity transforms between mass points obtained during map digitization and control and additional check points obtained by photogrammetry and surveying. Moreover, also photogrammetric models can be transformed to fit into the frame given by more precise reference points, e.g. GPS points.

The math model dealing with such transforms can be block adjustment with independent models (K. Schwidefsky/F. Ackermann, 1976). This model is capable for a rigorous handling of observations not only of photogrammetry but also of digitized maps and surveying. A realization of this approach can be found in H. Wiens (1986) and W. Benning/Th. Scholz (1990) – in the following not only the transform itself will be treated but also a comprehensive hypothesis test procedure. The combination of parameter estimation and hypothesis testing leads to a chaining procedure which is a feedback loop: after testing the residuals on Gaussian distribution the data snooping starts to detect points which do not fit into the given reference frame. After some iterations the overall accuracy of digital cartography is estimated which should be improved considerably compared with a priori values given in table 1. In this table accuracy measures are given according to different map scales.

Table 1: The ground tracking speed and accuracy of manual digitizing.

| Scale | Ground Speed (km/hr) | Ground Accuracy (m) |
|---|---|---|
| 1:10 000 | 54 | 2 |
| 1:20 000 | 108 | 4 |
| 1:25 000 | 135 | 5 |
| 1:50 000 | 270 | 10 |
| 1:100 000 | 540 | 20 |

Regarding the tracking speed during map digitizing a good operator captures data in a rate of about 1.5 mm per second. This is to maintain a tracking accuracy of about 0.2 mm. These figures indicate that there is no room left for more accurate data acquisition but the final data processing should result into much more accurate values in particular if large scale maps are digitized.

In order to complete the overall measures of accuracy for photogrammetry $\sigma_p$ and surveying $\sigma_s$ we can state the following values:

$$0.01m < \sigma_p < 1m$$
$$0.005m < \sigma_s < 0.1$$

A classification of accuracy leads to the relation $\sigma_c > \sigma_p > \sigma_s$; therefore these figures will be improved if data of cartography, photogrammetry, and surveying is merged with each other.

## 2 Block adjustment with independent models

In order to apply the chaining procedure for checking the metric quality of map digitizing the underlying adjustment model is reviewed. The important criteria of block adjustment with independent models (K. Schwidesfky/F. Ackermann, 1976) are extented to:

- the computing units can be isolated map regions, whole maps and image pairs
- the functional relation between the model/object space is a spatial similarity transform
- the block unit is constrained by means of control points, additional check points, and in case of photogrammetric images the perspective centres

Fig. 1 gives the well-known individual position of the different independent models.

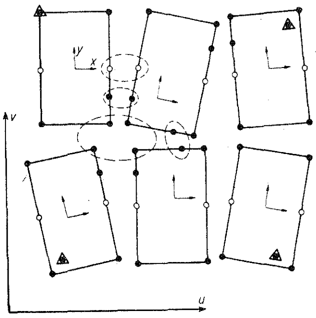

Fig. 1: Connection of independent models to a block unit (from K. Schwidesfky/F. Ackermann, p. 206)

The spatial similarity transform can be derived by differential or purely geometric considerations (K.R. Koch, 1987). Let be B the matrix of coefficients providing for three translations, three rotations and a change in scale (K.R. Koch/D. Fritsch, 1981)

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & \dots \\ 0 & 1 & 0 & 0 & 1 & \dots \\ 0 & 0 & 1 & 0 & 0 & \dots \\ 0 & -Z_1 & Y_1 & 0 & -Z_2 & \dots \\ Z_1 & 0 & -X_1 & Z_2 & 0 & \dots \\ -Y_1 & X_1 & 0 & -Y_2 & X_2 & \dots \\ X_1 & Y_1 & Z_1 & X_2 & Y_2 & \dots \end{pmatrix} \qquad (1)$$

In this matrix the coordinates $X_i, Y_i, Z_i$ can be approximate values of the object coordinates of point $P_i$; in case of a two-dimensional transform this matrix shrinks to two translations, one rotation and the scale change simply by deletion of the third row and the third column, and the forth and fifth row.

Thus, for every model $j$ the following observation equations are valid

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}_{ij} + \begin{pmatrix} v_X \\ v_Y \\ v_Z \end{pmatrix}_{ij} = \begin{pmatrix} U \\ V \\ W \end{pmatrix}_{i} - \begin{pmatrix} U_o \\ V_o \\ W_o \end{pmatrix}_{j} - \begin{pmatrix} 0 & -Z & Y & X \\ Z & 0 & -X & Y \\ -Y & X & 0 & Z \end{pmatrix}_{ij} \begin{pmatrix} d\Omega \\ d\Phi \\ dA \\ d\Lambda \end{pmatrix} \qquad (2)$$

In short we can write (2) as

$$E(l_{ij}) = l_{ij} + v_{ij} = x_i - B'_{ij}p_j, \qquad D(l_{ij}) = \sigma^2 P_i^{-1} \qquad (3)$$

in which $l_{ij}$ is the observation vector for point $P_i$ in model $j$ and $v_{ij}$ its corresponding residual vector, $x_i$ is the vector of unknown object coordinates of point $P_i$, $B'_{ij}$ contains the coefficients of the similarity transform und $p_j$ is the vector of the seven unknown datum parameters of model $j$. The operators $E$ and $D$ characterise expectation and dispersion respectively.

Control points can be considered twice: on the one hand non-random coordinates constrain (2) in form of the linear equation system

$$Hx = 0 \qquad (4)$$

and on the other hand random coordinates deliver additional observation equations

$$\begin{pmatrix} U \\ V \\ W \end{pmatrix}_{i} + \begin{pmatrix} v_U \\ v_V \\ v_W \end{pmatrix}_{i} = \begin{pmatrix} U \\ V \\ W \end{pmatrix}_{i}, \qquad D(\begin{pmatrix} U \\ V \\ W \end{pmatrix}_{i}) = \sigma^2 \begin{pmatrix} \sigma_{UU} & 0 & 0 \\ 0 & \sigma_{VV} & 0 \\ 0 & 0 & \sigma_{WW} \end{pmatrix} \qquad (5)$$

The parameter estimation by means of least-squares is done using (3) – known as *Gauss Markov model* – in which the residuals are minimized according to

$$\begin{matrix} \|v\|_P^2 = min \\ \text{subject to} Hx = 0 \end{matrix} \qquad (6)$$

if non-random control points exist.

## 3 Hypothesis testing

Testing the parameters and residuals of the parameter estimation we have to decide on the distribution of the residuals. This means the chain of hypothesis tests we propose is highly dependend on these results. For that reason two main approaches have to be outlined which may depend on the Normal Distribution and symmetric distributions respectively.

### 3.1 Hypothesis tests based on Normal Distribution

In order to verify the results of the parameter estimation let us first start with checking the residuals. Therefore the following initial test has to be solved:

Verify the normal distribution of residuals $v_j$ of check and control point coordinates for each model $j \quad \forall j = 1, 2, \ldots$ obtained by block adjustment. The condition is satisfied if the null hypothesis is not rejected:

$$H_o : v_j \sim N(0, \sigma^2) \tag{7}$$

To verify the null hypothesis apply the $\chi^2$ test against the theoretical normal distribution

$$H_o : \sum_{j=1}^{k} (f_j - np_j)^2 / np_j \leq \chi^2_{\alpha, k-r} \tag{8}$$

where *$k$ is the number of subsets in which the sampled residuals are subdivided*

*$f_j$ is the number of sampled residuals in the subset $j$*

*$n$ is the total number of sampled residuals*

*$p_j$ is the theoretical probability (according to a Binomial distribution for an outcome inside the subset $j$*

*$r$ are the degrees of freedom*

*$\alpha$ is the probability for an error of first kind*

In case of a normal distribution, $f_j$ will have a Binomial distribution with the theoretical mean $np_j$ and the variance $np_j(1 - p_j)$.

If the null hypothesis is not rejected apply the *data snooping test* already proposed by F.Crosilla/G. Garlatti (1991) for digital cartography. Let $x_{ip}$ and $x_{ic}$ be the $x$-coordinates of point $P_i$ under control coming from photogrammetric ($p$) and cartographic ($c$) procedures respectively:

$$\begin{array}{rcl} x_{ip} & \sim & N(\mu_{ixp}, \sigma^2_{xp}) \\ x_{ic} & \sim & N(\mu_{ixc}, \sigma^2_{xc}) \end{array} \tag{9}$$

Define the random variable

$$y_i = x_{ip} - x_{ic} \tag{10}$$

following a Normal Distribution

$$y_i = N(\mu_{ixp} - \mu_{ixc}, \sigma^2_{xp} + \sigma^2_{xc} - 2\sigma_{xpxc}) \tag{11}$$

Once the probability $\alpha$ for a first kind error is accepted and the parametric space $S$ is partitioned in the subspace of acceptance ($A$) and rejection ($R$)

$$\begin{array}{rcl} A & = & \{x_{ip}, x_{ic} \in S : -z\alpha/2 \leq y_i/\sigma_y < z\alpha/2\} \\ R & = & \{x_{ip}, x_{ic} \in S : y_i/\sigma_{iy} \leq -z\alpha/2 \quad \text{or} \quad y_i/\sigma_{iy} > z\alpha/2\} \end{array} \tag{12}$$

where

$$z\alpha/2 : P[z < -z\alpha/2] = p[z > z\alpha/2] = \alpha/2 \quad \forall z \in S \tag{13}$$

than holds

$$\begin{array}{rclcrcl} H_o & = & 0 & \text{if} & x_{ip}, x_{ic} & \in & A \\ H_o & = & \emptyset & \text{if} & x_{ip}, x_{ic} & \in & R \end{array} \tag{14}$$

## 3.2 Non parametric hypothesis tests

In case the normal distribution of $f_i$ is not accepted non parametric tests should be applied. The first question which arises is the symmetry behaviour therefore we have to verify the symmetry of the distribution of the sampled $y_i \quad \forall i = 1, 2, \ldots, m$ ($m$ number of points). The null hypothesis is formulated such that the mean value of distribution corresponds with the median value.

Let be the mean value of $y_i \Rightarrow \hat{y} \quad \forall i = 1, 2, \ldots, m$ and the median value of $y_i \Rightarrow y_{.50} \quad \forall i = 1, 2, \ldots, m$ Verify that $\hat{y} = y_{.50}$. This can be done by the two-tailed *Quantile Test* for which the following null hypothesis is introduced:

$$H_o : \text{the} \quad .50 \quad \text{population quantile is} \quad \hat{y} \tag{15}$$

or equivalently

$$H_o : P(Y \leq \hat{Y}) \geq .50\text{quantile and} \quad P(y < \hat{y}) \leq .50\text{quantile} \tag{16}$$

in which $Y$ has the same distribution as the sampled $y_i$.

For a decision rule let us introduce the quantities:

$T_2$ *number of observations ¡* $\hat{y}$

$T_1$ *number of observations* $\leq \hat{y}$ and

$T_1 = T_2$ *if none of the observations* $= \hat{y}$

The critical region corresponds to values of $T_2$ which are too large, and to values of $T_1$ which are too small. This region is found by entering a table of the *Binomial* distribution with the sample size $m$ and the hypothesized probability .50. We now have to solve the following problem: find the number $t_1$ such that

$$P(z \leq t_1) = \alpha_1 \tag{17}$$

where z has the binomial distribution with parameters $m$ and .50, and where $\alpha_1$ is about half of the desired level of significance. Then find the number $t_2$ such that

$$\begin{array}{rrcl} & P(z > t_2) & = & \alpha_2 \\ \text{or} & P(z \leq t_2) & = & 1 - \alpha_2 \end{array} \tag{18}$$

where $\alpha_2$ is chosen such that $\alpha_1 + \alpha_2$ is about equal the *desired level of significance* (Note: $\alpha_1$ and $\alpha_2$ are not integer numbers). The final decision rule is given by

$$H_o = \emptyset \quad \text{if} \quad T_1 \leq t_1 \quad \text{or} \quad T_2 > t_2 \tag{19}$$

otherwise accept $H_o$ with a significance level equals $\alpha_1 + \alpha_2$. The symmetry of the distribution can also be verified by the *Wilcoxon Signed Rank Test* which is reported in the following. In this context we will differentiate in two cases:

### Case 1: the condition of symmetry is accepted

Verify that the median $y_{.50} = 0$. Then apply the two-tailed Wilcoxon signed rank test

$$\begin{array}{rcl} H_o : y_{.50} & = & 0 \\ H_1 : y_{.50} & = & \emptyset \end{array} \tag{20}$$

If $H_o = 0$ then it follows for the symmetry condition

$$\hat{y} = 0 \tag{21}$$

and therefore systematic and gross errors **are not present** within the population with significance level $\alpha$. The test statistic $T$ equals the sum of the ranks assigned to those values of $y_i > 0 \quad \forall i = 1, 2, \ldots, m$.

$$T = \sum_{i=1}^{m} R_i \tag{22}$$

where $R_i = 0$ if $y_i < 0$ and $R_i$ equals the rank asigned to positive values of $y_i$.

As a decision rule we have to reject $H_o$ at a level of significance $\alpha$ if $T$ exceeds $w_{1-\alpha/2}$ or if $T$ is less than $w_{\alpha/2}$. If $T$ is between $w_{\alpha/2}$ and $w_{1-\alpha/2}$ or equal to either quantile, accept $H_o$. The quantile values of the Wilcoxon signed rank test statistics are usually reported in appropriate tables.

**Case 2: the condition of symmetry is not accepted**

This leads to the alternative hypothesis

$$H_1 : \text{the} \quad .50 \quad \text{population quantile is not equal to} \quad \hat{y} \tag{23}$$

In this case the *Sign Test* has to be applied. Classify as (+) those observations $y_i > 0$ and as (-) those observations $y_i < 0$ and as (0) the values $y_i = 0$. Having in mind that $y_i = x_{ip} - x_{ic}$ the null hypothesis for a two-tailed test can be stated as follows

$$H_o : P(x_{ip} < x_{ic}) = P(x_{ip} > x_{ic}) \ \forall i \tag{24}$$

Now the conclusion can be drawn: if $H_o = 0$ the presence of systematic or gross errors can be rejected wit the level of significance $\alpha$.

The test statistic $T$ equals the number of (+) pairs; that is $T$ equals the number of pairs $(x_{ip}, x_{ic})$ in which $x_{ic}$ is less than $x_{ip}$. As decision rule we obtain: disregard all tied points (if any), and let $n$ be the number of pairs that are not ties $\Rightarrow n =$ total number of (+) and (-).

If $n = 20$ use the table of Binomial distribution with the proper value of n and with $P = 0.5$. Select the value of about $\alpha/2$ and call it $\alpha_1$ (not integer). The corresponding value of $z$ is called $t$.

The critical region of size $2\alpha$ corresponds to values of $T$ less than or equal to $t$ or greater than or equal to $n - t$. Furthermore we have

$$H_o = \not{0} \quad \text{if} \quad T \leq t \quad \text{or} \quad T \geq n - t \tag{25}$$

at a level of significance 2 $\alpha_1$. For $n$ larger than 20 use the formula

$$t = 0.5(n + w_{\alpha/2}\sqrt{n}\,) \tag{26}$$

where $\alpha/2$ is the quantile of the standard normal distribution. For $\alpha = 0.05$ we have $w_{\alpha/2} =$ -1.96 and therefore

$$t = 0.5n - \sqrt{n} \tag{27}$$

**If $H_o = 0$ the metric quality is accepted.**
**If $H_o = \not{0}$ look for a systematic trend.**

This systematic trend can be detected by a *Cox and Stuart test* but this will not be treated in this paper.

# 4 Conclusions and outlook

The paper introduced with overall accuracy measures for digital cartography, photogrammetry and surveying. While map digitizing serves as data acquisition method for mass points photogrammetry and surveying deliver the reference frame to check and adjust cartographic data. The method for this necessary data processing consists of block adjustment with independent models and hypothesis testing. It was shown that the whole statistical inference process should care for the distribution of the residuals, the data snooping, the final choice of points to be controlled and checked, and, moreover, the estimation of an overall measure of the transformed digital cartographic data. All the formulas which are necessary for these chain of tests are given in the paper.

When using other objective functions then the method of least-squares it is known, that hypothesis testing becomes very difficult. For example, the $L_1$ norm is more sensitive against blunders, and the $L_\infty$ reduces the maximum error. But the costs in terms of computing time are much higher than using least-squares algorithms. Although linear and quadratic programming algorithms can be handled for large equation systems as well, the question on underlying distributions of the outcoming residuals is not yet solved. For that reason it becomes obvious to combine different error norms with each other: to start with blunder detection in $L_1$, to reduce the maximum errors by $L_\infty$, and to decide on the final data transform with $L_2$.

Further work should concentrate on the application of the tests proposed by this paper. The final aim is a quality check of map digitizing using all the data available from photogrammetry and surveying. Only in this way an objective overall measure for geometric data bases can be found.

# 5 References

Benning, W., Scholz, Th. (1990): Modell und Realisierung der Kartenhomogenisierung mit Hilfe strenger Ausgleichungstechniken. Zeitschr. Verm. Wesen (ZfV), Vol. 115, pp. 45 - 55.

Crosilla, F., Garlatti, G. (1991): Data snooping in digital cartography. Private communications, Udine.

Glaessel, H., Fritsch, D. (1991): Leistungsfaehigkeit der Homogenisierung von digitalisierten grossmassstaeblichen Karten. In: Geo-Informatik, Ed. M. Schilcher, Siemens Nixdorf Informationssystme AG, Berlin, pp. 301 - 310.

Koch, K.R., Fritsch, D. (1981): Multivariate hypothesis tests for detecting recent crustal movements. Tectonophysics, Vol. 71, pp. 301 - 313.

Koch, K.R. (1988): Parameter estimation and hypothesis testing in linear models. Springer, Heidelberg.

Schwidesfky, K., Ackermann, F. (1976): Photogrammetrie, Teubner, Stuttgart, 384 p.

Wiens, H. (1986): Flurkartenerneuerung mittels Digitalisierung und numerischer Bearbeitung unter besonderer Beruecksichtigung des homogenen Zusammenschlusses von Inselkarten zu einem homogenen Rahmen-Kartenwerk. Schrift. Reihe Inst. Kartogr., Topogr., No. 17, Univ. Bonn, Bonn.