FUZZY LOGIC – AN ALTERNATIVE TO CONVENTIONAL METHODS FOR IDENTIFYING OUTLIERS FOR PROCESSING GEODETIC NETWORKS

Alexandru ILIES, Lecturer, eng., U.T.C.B., Romania, alexandru_ilies@yahoo.com
Doina VASILCA, Lecturer, PhD eng., U.T.C.B., Romania, doinavasilca@yahoo.com

Abstract: *It is well known that a least square adjustment is very sensitive to large errors in observations. Therefore, any estimated parameters will be affected by these errors. In this paper, we present methods based on the principle of iterative weight, by emphasizing the disadvantage of the use of these methods when there is more than one large error present. We also analyze the possibility of using the Fuzzy Logic as an alternative to methods developed on the principle of iterative detection of large errors.*

Keywords: *adjustment geodetic network, fuzzy logic, measurement, outlier detection*

1. Introduction

Precision in determining geodetic networks mainly depends on the measurement instruments, the techniques of such measurement and the mathematical model of adjustment.

The mathematical model used for adjustment influences the precision of determining networks through conditions imposed by the model, as well as the possibilities of the model to identify the measurements which have a negative influence over the unknown parametric precision of the model.

In this way, identifying all measurements affected by large errors and removing them, or reducing their influence over the unknown parameters increases the precision of determining the geodetic networks.

It is important, therefore to establish the minimal value of the errors' influence may exert on unknown parameters. This value must be the criterion underlying among measurements with acceptable errors, as well as measurements with large errors (outliers).

The procedures for identifying outliers are based on statistical tests, which are applied to the squares residuals and the correlation among residuals. Because of this, the sensitivity of these procedures is limited.

For more than 45 years, the geodetic community has been conducting research in the field of large error identification.

In 1965, Prof. Dr. Willem Baarda started research in this field and published the results in: "*A test procedure for use in geodetic networks*".

Other researchers have also presented different techniques of identifying large errors.

In the present paper, we shall introduce a procedure - less known in our country - which uses statistical tests and fuzzy logic.

As far as this procedure is concerned, firstly, we shall define large errors and then briefly present a few of the classical methods of error identification. We also present basic concepts of fuzzy logic, as well as fuzzy sets as well as their properties. Following this, we shall describe the algorithm of identifying outliers by using the fuzzy technique. The last part of the paper presents an application of the fuzzy logic in a GPS network.

Defining "*outliers*"

It is generally assumed that geodetic measurements have random errors with normal distribution. The localised, larger disturbances are considered outliers, whereas the smaller ones are constant and are labelled as systematic errors.

Normally, we consider outliers to be random errors within the deviation of a normal distribution; such errors do not belong to the population featuring a normal distribution and, therefore, they are difficult to identify.

The study of large errors has mainly interested statisticians who have developed different methods for discovering them. Unfortunately, their methods are not universally applicable.

In conventional methods, such as "*Data Snooping (DS)* ", "*Tau*" and "*t-test*", the outliers are determined through an iterative process, by applying statistical tests and, then, they are eliminated from the observations set.

Classical methods

Data Snooping (DS), has been suggested by Prof. Baarda and can by applied only if the theoretical value of unit weight $\left(\sigma_0^2\right)$ is known. If this value is unknown, the *a priori* variance $\left(s_0^2\right)$ can be used instead. In triangulation, levelling and GPS networks, the Ferro equation can be used:

$$s_0 = \sqrt{\frac{w^T w}{n_m \cdot n_p}} \tag{1}$$

where:

w -is the vector of discrepancies in triangles (polygons);

n_m -is the number of measurements in each triangle (polygon);

n_t -is the number of triangles (polygons) from network;

n_m = 3 in triangulation networks, 9 in GPS networks, minimum 3 in levelling networks;

DS is realized using normalized residuals. The statistic is compared with critical value obtained from normal distribution:

$$T_i = \frac{(Pv)_i}{s_0 \sqrt{PQ_{vv}P)_{ii}}} \tag{2}$$

where:

v -is the vector of residuals;

P -is the observations weight matrix

Q_{vv} -is the cofactor matrix of residuals

The critical value is:

$$q = N_{1-\alpha_0/2} = \sqrt{F_{1,\infty,1-\alpha_0}} = \sqrt{\chi^2_{1,\infty,1-\alpha_0}} \tag{3}$$

where:

α_0 -is the significance level

N, F, χ^2 - are values from tables of: normal distribution, Fisher's distribution and χ^2 distribution.

The significance level for one observation α_0 is calculated by using the relation:

$$\alpha_0 = 1-(1-\alpha)^{1/n} \equiv \alpha / n \tag{4}$$

where:
α -is the total significance level, usually is considered 5%.
n -is the number of observations.

Tau test: If the *a priori* variance is not known, or a certain value cannot be established prior to adjustment, then *a posteriori* variance (m_0^2) is used for indentifying outliers.

The residuals normalized with *a posteriori* variance are not normally distributed. These are distributed in Tau (τ):

$$m_0^2 = \frac{v^T P v}{f} = \frac{v^T P v}{n-u+d} \tag{5}$$

where:
f = degrees of freedom
n = is the number of measurements
u – is the number of unknown parameters
d – is the rank defect

The relation for calculating the statistic is:

$$T_i = \frac{(Pv)_i}{m_0 \sqrt{PQ_{vv}P)_{ii}}} \tag{6}$$

The relation for determination of the critical value τ :

$$q = \tau_{f,1-\alpha/2} = \sqrt{\frac{f \times t^2_{f-1,1-\alpha_0/2_0}}{f-1+t^2_{f-1,1-\alpha_0/2}}} \tag{7}$$

Where:
t –is the theoretical value of Student test;
τ –is the value of Tau given in table

T-test: If the observation (l_i) has an error (Δ_i), it is not recommended to identify the outlier by using *a posteriori* variance from an erroneous adjustment. In such a situation, it is better to use the variance from which the influence of the suspected error has been eliminated:

$$\overline{m}_0^2 = \frac{1}{f-1}\left(f m_0^2 - \frac{v_i^2}{(Q_{vv})_{ii}} \right) \tag{8}$$

Where:
$\overline{m}_0^2$ -is *a posteriori* variance from which the influence of the suspected error has been removed

The relation for a statistical calculation is:

$$T_i = \frac{(Pv)_i}{\overline{m}_0 \sqrt{PQ_{vv}P)_{ii}}} \tag{9}$$

The relation for calculating the critical value is:

$$q = t_{f-1,1-\alpha_0/2} \tag{10}$$

The Fuzzy Logic Method

All procedures based on probabilistic calculation can only test one large error. If the observations have more large errors, then an iterative process is applied. In doing so, for each step, the error exhibiting the largest absolute value is tested.

According to the model of adjustment measurements based on the Gauss-Marcov least squares method, the determination of parameters $\boldsymbol{x}$, having minimum dispersions, from observations $\boldsymbol{l}$ (having covariance matrix $\boldsymbol{Q}_l$) is based on the relation:

$$\begin{aligned} \boldsymbol{l} &= \boldsymbol{f}(\boldsymbol{x}^0) + \boldsymbol{A}\boldsymbol{dx} + \Delta \\ \boldsymbol{D}(\boldsymbol{l}) &= \sigma_0^2 \boldsymbol{Q}_l \end{aligned} \tag{11}$$

where:

$\boldsymbol{A}$ -is the design matrix;

$\boldsymbol{dx}$ - is the vector of corrections applied to the approximate parameters $\boldsymbol{x}^0$;

Δ - is the vector of observations errors;

σ_0^2 -is the variance factor;

$\boldsymbol{rang}(\boldsymbol{A}) = \boldsymbol{u}$

$\boldsymbol{u}$ - is the number of parameters;

The mathematical relation existing among residuals and observations errors is:

$$\boldsymbol{v} = (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{Q}_x\boldsymbol{A}^T\boldsymbol{Q}_l^{-1})\Delta \tag{12}$$

$$\boldsymbol{v} = \boldsymbol{Q}_v\boldsymbol{Q}_l^{-1}\Delta \tag{13}$$

$$\boldsymbol{v} = \boldsymbol{R}\Delta \tag{14}$$

where:

$\boldsymbol{Q}_x = \boldsymbol{A}^T\boldsymbol{P}\boldsymbol{A}$ with the weight matrix $\boldsymbol{P} = \boldsymbol{Q}_l^{-1}$;

$Q_v = (P^{-1} - AQ_xA^TP)$ - the variance-covariance matrix of corrections $\boldsymbol{v}$;

$\boldsymbol{R} = \boldsymbol{Q}_v\boldsymbol{P}$ - is denominated the redundant matrix;

The matrix $\boldsymbol{R}$ is:

$$\begin{bmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \boldsymbol{v}_3 \\ \vdots \\ \boldsymbol{v}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_{11} & \boldsymbol{r}_{12} & \boldsymbol{r}_{13} & \cdots & \boldsymbol{r}_{1n} \\ \boldsymbol{r}_{21} & \boldsymbol{r}_{22} & \boldsymbol{r}_{23} & \cdots & \boldsymbol{r}_{2n} \\ \boldsymbol{r}_{31} & \boldsymbol{r}32 & \boldsymbol{r}_{33} & \cdots & \boldsymbol{r}_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{r}_{n1} & \boldsymbol{r}_{n2} & \boldsymbol{r}_{n31} & \cdots & \boldsymbol{r}_{nn} \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \\ \vdots \\ \Delta_n \end{bmatrix} \tag{15}$$

where:

$\boldsymbol{r}_{ii}$ with $i = 1, 2, \ldots, n$, are redundancy components.

According to relation (15) the residuals $\boldsymbol{v}$ depend not only on the observation errors Δ, but also on the components of the redundant matrix $\boldsymbol{R}$.

Applying the least squares adjustment, the trace of matrix $\boldsymbol{R}$ gives us the number of degrees of freedom of the network, regardless of the fact that the observations are correlated or not.

$$\boldsymbol{tr}(\boldsymbol{R}) = \boldsymbol{d} \tag{16}$$

Equation (15) demonstrates that each value v_i is obtained from a complicated relation including all observations errors resulting from adjustment. This relation does not only depend on the accuracy of the observations (expressed by Q_l) of geometrical and physical constraints between measurements (expressed by $f(x)$), but also on the adjustment principle.

Strictly speaking, if a value v_i does not pass the statistical test, this merely indicates the fact that some observations associated with this value are strongly affected by large errors. To identify these errors, all the coefficients r_{ij} of the R matrix must be taken into account.

The identification of outliers by using the Fuzzy Logic presented by W. Sun is based on Professor Lotif Zadeh's idea to extent the clear limits of the variables under discussion exhibiting a specific degree of uncertainty, to less clear limits, by using membership functions.

If classical logic can suggest that if an object belongs to a population (set) or not, the Fuzzy Logic allows for a more flexible interpretation of the membership notion. Therefore, in various degrees, many objects can belong to a set. Mathematically this may be expressed as follows:

Let X be a set of the elements x. A fuzzy set A defined on set X is characterized by the membership function $\mu_A(x)$. This function associates a degree of membership to each element x in set A.

$$\mu_A(x): X \to [0,1] \tag{17}$$

In order to represent a fuzzy set, the membership function must be defined.

Thus, a fuzzy set A is completely defined by the set of ordered pairs:

$$A = \{(x, \mu_A(x)) \mid x \in X\} \tag{18}$$

For detecting outliers by using the Fuzzy Logic some ***steps*** are required:

The first one consists of testing the existence of large errors in measurements. This can be done by testing the hypothesis of equality between the theoretical variance and the estimated one (*a priori* variance and *a posteriori* variance).

$$H_0 : \hat{\sigma}_0^2 = \sigma_0^2 \text{ and } H_a : \hat{\sigma}_0^2 > \sigma_0^2 \tag{19}$$

The null hypothesis is accepted if the test is verified:

$$\frac{t \cdot \hat{\sigma}_0^2}{\sigma_0^2} < \chi^2_{t,1-\alpha} \tag{20}$$

Where:

t – the redundancy of the concerned problem;

α – significance level for statistical test.

Because:

$$\frac{\hat{\sigma}_0^2}{\sigma_0^2} = \frac{r^T P r}{t\sigma_0^2} \tag{21}$$

The equation (20) can be written:

$$\boldsymbol{r}^T \boldsymbol{P}\boldsymbol{r} < \chi^2_{t,1-\alpha} \tag{22}$$

If this test is not passed, then the hipothesis $\boldsymbol{H}_0$ is rejected with a significance level $\boldsymbol{\alpha}$. This is indicative of the fact that there is something wrong with the observations.

We assume that the observations with large errors and which are most probably abnormals, are those with the largest contribution in residuals, whereas the observations with the smallest contribution are normals.

In order to make a fuzzy set for localizing large errors among the set of observation errors, we form two subsets: subset A, which is defined as the set of observation errors with the largest contribution in residuals, and subset B, which is defined as the set of observations errors with the smallest contribution in residuals.

In fuzzy set terms, the set of large errors $\boldsymbol{H}$ is defined as the intersection of sets $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{H} = \boldsymbol{A} \cap \boldsymbol{B} \tag{23}$$

If the membership functions $\mu_A(\Delta_i)$ and $\mu_B(\Delta_i)$ of sets A and B are known, then the membership function of set $\boldsymbol{H}$ is:

$$\mu_H(\Delta_i) = \min(\mu_A(\Delta_i), \mu_B(\Delta_i)) \qquad i=(1,2,...,n) \tag{24}$$

According to the relation (24), $\mu_H(\Delta_i)$ is equal to 1.0 only if $\mu_A(\Delta_i)$ and $\mu_B(\Delta_i)$ are 1.0 simultaneously. In other words, the observation errors are suspected to be outliers only when they have the largest contribution to abnormal residuals and the smallest contribution to normal residuals at the same time.

We argue that the error with the largest value of membership function is likely to be an outlier.

If a critical value $\boldsymbol{C}_{\mu_H}$ is established for the values of membership function $\mu_H(\Delta_i)$ we can differenciate errors as following:

$$\Delta_i \text{ belongs to } \begin{cases} \text{large errors} & \text{if } \mu_A(\Delta_i) > \boldsymbol{C}_{\mu_H} \\ \text{normal errors} & \text{if } \mu_A(\Delta_i) < \boldsymbol{C}_{\mu_H} \end{cases} \tag{25}$$

To make this differentiation, we argue that a method for evaluating membership function as well as a method for calculating the critical value must be found.

Acording to the Data Snooping theory (DS), if the observations are uncorelated, then normalized residuals have a normal distribution.

$$w_i = \frac{|v_i|}{\sigma_0 \sqrt{q_{v_i}}} = \frac{|v_i|}{\sigma_{v_i}} \in N(0,1) \tag{26}$$

Where:

$\boldsymbol{w}_i$ - normalized residuals;

$\boldsymbol{q}_{v_i}$ - elements of diagonal of cofactor matix of residuals;

$$\boldsymbol{Q}_v = (\boldsymbol{P}^{-1} - \boldsymbol{A}\boldsymbol{Q}_x\boldsymbol{A}^T\boldsymbol{P}) \tag{27}$$

For significance level $\boldsymbol{\alpha}$ when testing each observation we realize this comparison:

$$\boldsymbol{w}_i \le \boldsymbol{N}_{1-\alpha/2}(0,1) \tag{28}$$

If a standardised residual is greater than the critical value, i this is considered as affected by large errors. Otherwise, the residual is not influenced by large errors.

In this way, two fuzzy subsets were formed: subset $N(v_i)$, consisting of normal residuals (the tested values are smaller than the critical value), and subset $M(v_i)$, consisting of abnormal residuals (the tested values are larger than the critical value).

For these subsets, the membership function is calculated as follows: for subset $N(v_i)$, the membership function is zero and the values of the membership function of the subset $M(v_i)$ belong to the interval (0,1).

According to this rule, the membership function is:

$$\mu_M(v_i) = \begin{cases} 0; & w_i \le N_{1-\alpha/2} \\ \dfrac{1}{1 + r_{ii}\left(\dfrac{\alpha}{w_i - N_{1-\alpha/2}}\right)^2}; & w_i > N_{1-\alpha/2} \end{cases} \tag{29}$$

where: r_{ii} is the redundance and α is the significance level.

With relation (29), the membership function values are determined for the residuals from subset $M(v_i)$, that are most probably affected by large errors. For subset $N(v_i)$, the values of membership function are calculated using the complementarity property of the Fuzzy Theory:

$$\mu_N(v_i) = 1 - \mu_M(v_i) \tag{30}$$

For determining the membership function of observations errors is used the redundancy matrix normalizing all elements thus:

$$\tilde{r}_{ij} = \frac{|r_{ij}|}{\max|r_{ij}|} \quad \text{i,j = 1, 2, 3,...,}n \tag{31}$$

In this way we obtain the relative redundancy matrix $\tilde{R}$ with elements belonging to the interval (0,1).

$\tilde{r}_{ij}$ represents the relative contribution of the j^{th} observation error on the i^{th} residual.

Therefore the rows of matrix $\tilde{R}$ represent the relative contribution of all observation errors to an individual residual and columns of the same matrix represents the contribution of an observation error to all residuals.

Furthermore, the membership functions $\mu_A(\Delta_i)$ and $\mu_B(\Delta_i)$ asociated to subsets A and B are calculated, using the membership functions asociated to residuals μ_N and μ_M and matrix $\tilde{R}$, as follows:

Using α_cut in the subset $M(v_i)$, the maximum effect of i^{th} observation error in the residuals with membership function values $\mu_M(v_i) \ge 0.5$ is calculated using the relation:

$$\tilde{r}_{mi} = \max(|\tilde{r}_{ki}|) \; ; \; v_k \in M_{0.5} \tag{32}$$

Then, the membership function values of observation errors from subset A are calculated using the following relation:

$$\mu_A(\Delta_i) = \tilde{r}_{mi} \cdot \mu_M(v_i) \tag{33}$$

In the same way, the membership function values of observation errors from subset B are determined as follows:

$$\mu_B(\Delta_i) = 1 - \tilde{r}_{ni} \cdot \mu_N(v_i) \tag{33}$$

where: $\tilde{r}_{ni}$ - the maximum effect of i^{th} observation error in the residuals with membership function values $\mu_N(v_i) \geq 0.5$ calculated with the relation:

$$\tilde{r}_{ni} = \max(|\tilde{r}_{ki}|) \; ; \; v_k \in N_{0.5} \tag{34}$$

The observations possibly affected by large errors are those which have the maximum effect in abnormal residuals, or have the minimum effect in normal residuals.

The maximum membership function value, obtained from equations (33) and (34), indicate the degree of deviation from the normal of the L_i observation.

According to the theory of fuzzy sets, after reunion the two fuzzy subsets ***A*** and ***B***, having membership functions $\mu_A(\Delta_i)$ and $\mu_B(\Delta_i)$, the ***H*** subset is obtained with the membership function given by:

$$\mu_H(\Delta_i) = \max(\mu_A(\Delta_i), \mu_B(\Delta_i)) \tag{35}$$

From this relation, we can conclude that the membership function value from the ***H*** subset indicates the observation outlying degree from normal. Thus, analyzing the membership function value we can decide if an observation is affected by large errors or not.

In order to determine the critical value we use the weighted average defuzzyfication method as follows:

$$C_H = \frac{P_i \cdot \mu_H(\Delta_i)}{\sum P_i} \tag{36}$$

where:

$$P_i = \begin{cases} \tilde{r}_{m_i} & ; \; \eta_H(\Delta_i) = \eta_A(\Delta_i) \\ \dfrac{1}{\mu_N(v_i)} - \tilde{r}_{n_i} & ; \; \eta_H(\Delta_i) = \eta_B(\Delta_i) \end{cases} \tag{37}$$

At the end, the test $\mu_H(\Delta_i) \geq C_H$ is applied to detect if subsets A or B contain the large errors.

Case study

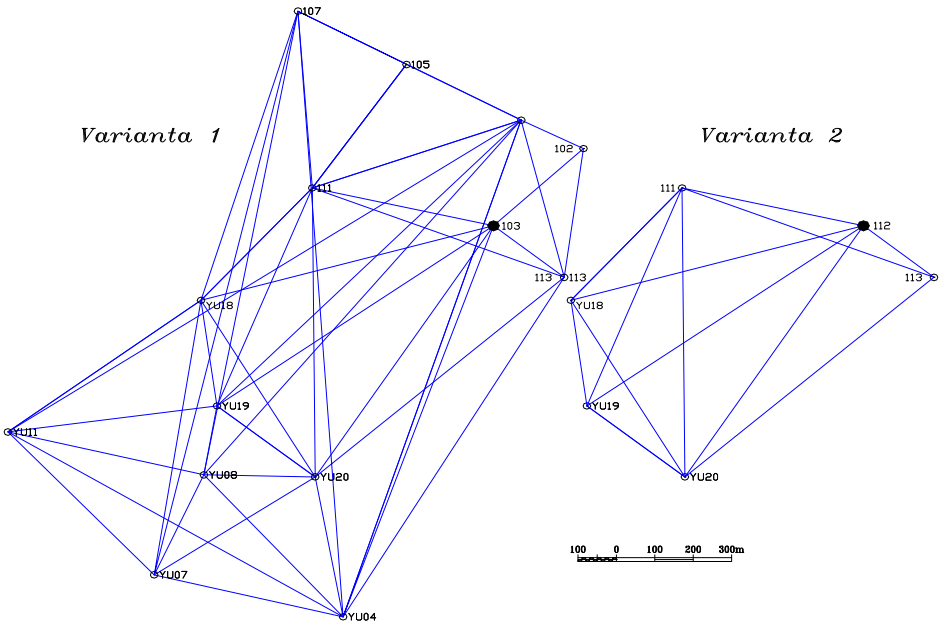

Fig. 1

In this study the GPS network observations (fig. 1) have been analized in two cases: variant 1 and variant 2. The properties of both variants are presented in table below:

Tabel 1.

Informations about network		Rețea1	Rețea 2
Number of points		13	6
Number of fixed points		1	1
Number of basis		61	15
Number of observations	*n*	183	45
Number of unknowns	*u*	39	15
Rank deficiency	*d*	0	0
Redundance	*r*	144	30

Conclusions

Both networks have been adjusted as unconstrained networks, by using point 112 as a fixed point. Since, after the adjustment the normality test has been passed, observations with large errors have been simulated.

Vectors 112-YU19 and 111-ZU18 have been modified on purpose. Following the adjustment, the normality test has not been passed. To determine large errors we used the ***t***-test.

Because in 3D networks the vectors are determined, if a component of a vector is determined as being a large error, all the other components of that vector are eliminated. The same network was tested by using the fuzzy logic. The same components affected by large errors were also determined by using the Fuzzy Logic.

This method of approach is not entirely correct, because we assume that observations are independent. Considering onlythe correlation among the three components of a measured base will be subject of a different research.

References

1. *Boz, Y., and Gokalp, E., 2006, Robust Estimation of the Outliers in GPS Baseline Components, XXIII FIG Congress, Munich, Germany, October 8-13.*
2. *Chuang, C.C., Su, S. F., and Chen, S.S., 2001, Robust TSK Fuzzy Modeling for Function Approximation UIT Outliers, IEEE Transactions on Fuzzy Systems, Vol. 9, No. 6, December.*
3. *Gokalp, E., and Boz, Y., 2005, Outlier Detection in GPS Networks with Fuzzy Logic and Conventional Methods, From Pharaohs to Geoinformatics, FIG Working Week 2005 and GSDI-8, Cairo, Egypt april 16-21.*
4. *Gullu, M. and Yilmaz, I., 2010, Outlier Detection for Geodetic Nets Using ADALINE Learning Algorithm, Scientific research and Essay Vol. 5(5), pp. 440-447.*
5. *Koch, K. R., 2007, Outlier Detection in Observations Including Leverage Points by Monte Carlo Simulation, AVN 10/2007, pp. 330-336.*
6. *Neumann, I., Kutterer, H., Schon, S., 2006, Outlier Detection in Geodetic Applications with respect to Observation Imprecision, "REC 2006 – The NSF Workshop on Reliable Wngineering Computing"< Savannah, Georgia, USA, pp. 75-90.*
7. *Pope, A. J., 1976, The Statistics of Residuals and The Detection of Outliers, NOAA Technical Report NOS65 NGS 1*
8. *Sun, W., 1994, A New Method for Localisation of Gross Errors, Survey Review, 32, 252, pp. 344-358.*