Outlier Detection by Fuzzy-Statistical Procedure

S.Jamali
Department of Surveying, Faculty of Engineering, Znajan University , P.O.Box 45195-313, Zanjan, IRAN
Email: jamali@mail.znu.ac.ir Tel: +98-241-515-26-11 Fax: +98-241-528-32-04

Abstract
It is known that the reliability of a geodetic network depends on magnitude of minimum undetectable error and its effect on the unknown parameters. Thus detection and remove of all outliers from geodetic observations improves the reliability of that geodetic network.
Since base of statistical procedure for detection of outliers is statistical tests on the least square residuals and generally there are correlations between least square residuals so sensitivity of this procedure is limited.
This paper presents fuzzy-statistical procedure for determination of outliers set using fuzzy techniques. This procedure is tested on two simulated geodetic networks and the results have been compared with the results of the conventional outlier testing (COT) method.

Keywords: Outlier, Statistics, Fuzzy Technique

Introduction

Detection of outlier has been an important research field for the past 40 years in the geomatics community. Since Baarda started his research in this field in 1965, a lot of papers including several techniques for outlier detection have been published.
Most of above techniques based on statistical tests on least squares residuals.The statistical correlation between the residuals and those between the residuals and the observations often decrease the quality of the results.
In this paper a new procedure that detects outliers using statistical tests and fuzzy techniques is presented. At first, the definition of outlier and conventional outlier testing method are revisited. Then principles of the new procedure are presented. Two simulated examples have been given to compare the results of this new procedure and conventional outlier testing method.

Review definition of outlier

The local and large disturbances are considered as gross errors, blunders or outliers whereas smaller and global deviations are considered as systematic errors(Kavouras 1982). Usually outliers are considered as variables that have the same variance as the random errors in a particular group of observations but different expectations(Baarda 1968).

Review COT method

COT method is a post-adjustment outlier detection method. Based on the theory of data snooping (Baarda 1968), if the observations are uncorrelated the normalized residuals w_i follows the standard normal distribution

$$w_i = \frac{|v_i|}{\sigma_0\sqrt{q_{v_i}}} = \frac{|v_i|}{\sigma_{v_i}} \sim N(0,1) \qquad (1)$$

By using α as the level of significance for testing each observation, the critical value of the above statistic w_i then is

$$w_i \le N_{1-\alpha/2}(0,1) \qquad (2)$$

The data snooping procedure is repeated if there is more than one outlier in the observations. In this method, only the observation corresponding to the largest w_i is deleted in each iteration. This procedure is continued until no more outlier is detected(Cen,M.2003).

Fuzzy-Statistical procedure

Localisation of outliers by fuzzy logic have first been introduced by Aliosmanoglu,S. and Akyilmaz,O. in 2001, thus the following fuzzy-statistical procedure have the same theory with some improvement in its fuzzy operators that I have done to reach better results.

The main idea of the fuzzy logic that have first been introduced by Zadeh, is to extend the crisp boundaries by using proper membership functions for the variables in question(Zadeh,L.A.1965). In classical sets, the membership value of any element is one if it exist in the set and zero if it dose not. But in fuzzy sets, the membership value of any element that indicates the degree of belonging of that to the set, is between zero and one. The greater membership value represents the greater degree of belonging.

In the following, the fuzzy techniques has been used to compare the fuzzy and crisp set of outliers due to their membership values obtained by using proper membership functions and residuals properties. In outlier detection, since the real errors of observation are unknown, the residuals and the redundancies are used as tools for testing.

Let Δ be the vector of observation errors in the linear functional model. The mathematical relation between residual vector and vector of observation error is given by follows(Aliosmanoglu,S.2001):

$$V = -(I - A.Q_X.A^T.Q_L^{-1}).\Delta \quad (3)$$

$$= -Q_V.Q_L^{-1}.\Delta \quad (4)$$

$$= -R \cdot \Delta \quad (5)$$

The matrix R in Eq.(5) is called the redundancy matrix and indicates the relation between observation errors and residuals. In as open form, the transformation form to residuals is as follows:

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ . \\ . \\ . \\ v_n \end{bmatrix} = - \begin{bmatrix} r_{11} & r_{12} & r_{13} & . & . & . & r_{1n} \\ r_{21} & r_{22} & r_{23} & . & . & . & r_{2n} \\ r_{31} & r_{32} & r_{33} & . & . & . & r_{3n} \\ . & . & . & . & . & . & . \\ .. & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ r_{n1} & r_{n2} & r_{n3} & . & . & . & r_{nn} \end{bmatrix} \begin{bmatrix} \Delta_1 \\ \Delta_2 \\ \Delta_3 \\ . \\ . \\ . \\ \Delta_n \end{bmatrix} \quad (6)$$

The above expression explicitly shows that residual of any observation is affected by all the observation errors with respect to the corresponding elements of the redundancy matrix. Either in correlated or non-correlated observations adjustment, the trace of the matrix R is equal to the degree of freedom of the network. However, because the matrix R is a rang-deficient matrix, a significant inverse transformation of Eq.(6) is not available.

Considering the above relationship denoted by Eq.(6), after the first adjustment, the test values are calculated for each of the residuals. In classical outlier detection, test values are calculated iteratively and only the residual having the largest test value is taken out of the all observations at each step.

After comparing the test values of each residual with its statistical limit, the residuals are classified into two groups of fuzzy sets: the set of observation with normal residuals (the test values

are less than statistical limit) $N(v_i)$ and the set of observation with abnormal residuals(the set values are greater than statistical limit) $M(v_i)$ (Aliosmanoglu, S.2001).

In statistical inference, the observations of whose test values slightly exceeding the statistical limit are also contained in the set of abnormal observation. In order to overcome such an uncertainty the fuzzy membership relation constructed between observation errors and the elements of redundancy matrix can be used.

After the hypothesis testing, following membership function is initiated satisfying the residuals with the values under the statistical limit which are the elements of subset $N(v_i)$ have the membership value of zero, and the residuals with test values greater than the statistical limit which are the elements of subset $M(v_i)$ have the membership values between (0,1) due to their discrepancy from the statistical limit, their redundancy number (r_{ii}) and the confidence level of the test (α). By this manner, the membership function concerned with the residuals, which are most probably affected by outliers is composed.

$$m_{\tilde{M}}(v_i)=\begin{cases} 0 \;;\; w_i \le N_{1-\alpha/2}(0,1) \\ \dfrac{1}{1+r_{ii}\cdot\left(\dfrac{\alpha}{w_i - N_{1-\alpha/2}(0,1)}\right)^2} \;;\; w_i > N_{1-\alpha/2}(0,1) \end{cases} \quad (7)$$

Through the above function, the membership values of each residual are determined. Using the Eq.(7), the membership function values of the residuals, which are most probably not affected by outlier, is composed easily by using complementary property of set theory as follows:

$$m_{\tilde{N}}(v_i) = 1 - m_{\tilde{M}}(v_i) \quad (8)$$

In order to determine the fuzzy membership relation of the observation errors, the redundancy matrix is used by normalizing all the elements of it as follows:

$$\tilde{r}_{ij} = \frac{|r_{ij}|}{\max|r_{ij}|} \qquad i, j = 1,2,3,\ldots, n \quad (9)$$

By normalizing due to Eq(9), a relative redundancy matrix, which has elements between 0 and 1, is obtained. The rows and columns of this matrix indicate the relative effects of the observation errors on any residual and of the any observation error on the residuals respectively.

The observation errors can also be handled in two groups in the same sense with residuals. For example; the subset A consist of the observation errors which have the maximum effects on the abnormal residuals, and the subset of B consist of observation errors which have the minimum effects on the normal residuals (Aliosmanoglu ,S.2001).

In order to determine the membership values of the so-called fuzzy sets A and B, the relative redundancy matrix $\tilde{R}$ and the membership functions $m_{\tilde{N}}$ and $m_{\tilde{M}}$ of the residuals are used. Let the fuzzy membership values of the elements of the sets A and B be $m_{\tilde{A}}(\Delta_i)$ and $m_{\tilde{B}}(\Delta_i)$ respectively. So-called membership values are calculated as follows:

Using α_cut in fuzzy set $M(v_i)$, the maximum relative effect of the error of observation i on the residuals of which have membership values $m_{\tilde{M}}(v_i) \ge 0.5$ is determined by:

$$\tilde{r}_{mi} = \max\left(|\tilde{r}_{ki}|\right) \;;\; v_k \in M_{0.5} \quad (10)$$

Then, the membership value of that observation error is calculated as follows:

$$m_{\tilde{A}}(\Delta_i) = \tilde{r}_{mi}.m_{\tilde{M}}(v_i) \quad (11)$$

In the same sense, the membership

value of set B is calculated as follows:

$$m_{\tilde{B}}(\Delta_i) = 1 - (\tilde{r}_{ni} . m_{\tilde{N}}(v_i)) \quad (12)$$

where $\tilde{r}_{ni}$ is maximum relative effect error of the observation i on the residuals of which have the membership values $m_{\tilde{N}}(v_i) \geq 0.5$, and obtained as follows:

$$\tilde{r}_{ni} = \max (|\tilde{r}_{ki}|) \; ; \; v_k \in N_{0.5} \quad (13)$$

The observations mostly affected by outliers are those, which have maximum effects on the abnormal residuals or have the minimum effects on the normal residuals or have both of them. The maximum value of the membership values given by Eq.(11) and Eq.(12) indicates the degree of outlying of the observation L_i in question.

Due to the fuzzy set theory, the union of fuzzy sets $\tilde{A}$ and $\tilde{B}$ being the set $\tilde{H}$ is obtained by:

$$m_{\tilde{H}}(\Delta_i) = \max (m_{\tilde{A}}(\Delta_i), m_{\tilde{B}}(\Delta_i)) \quad (14)$$

Regarding the membership values of the elements of the H, it can be inferred that greater the value of the observation greater the degree of outlying of that observation. Therefore, an observation can be decided whether it is an outlier or not with respect to its membership value.

In order to determine a significant limiting value, the following weighted average defuzzification method is used:

$$C_H = \frac{\sum P_i . m_{\tilde{H}}(\Delta_i)}{\sum P_i} \quad (15)$$

where

$$P_i = \begin{cases} \tilde{r}_{m_i} \; ; \; m_{\tilde{H}}(\Delta_i) = m_{\tilde{A}}(\Delta_i) \\ \dfrac{1}{m_{\tilde{N}}(v_i)} - \tilde{r}_{n_i} \; ; \; m_{\tilde{H}}(\Delta_i) = m_{\tilde{B}}(\Delta_i) \end{cases} \quad (16)$$

As a result, the membership values $m_{\tilde{H}}(\Delta_i)$ are compared with the limiting value C_H and the observations with $m_{\tilde{H}}(\Delta_i) \geq C_H$ are decided as outliers and contained in a different set.

Experiment 1

The following easy experiment verifies the application of fuzzy-statistical procedure for the detection of outlier. Let L be the vector of direct uncorrelated observations with the 10 measurements.

$L^T = [14,19,20,20,20.5,20,19.5,19,17.5,21]$

With $\sigma_{L_i} = 1.27$, $\alpha = 0.0455$ and r =10-1=9

Where σ_{L_i} is standard deviation of any observation, α is the level of significance for the testing each observation and r is degree of the freedom.

After least square solution, L_1 is successfully flagged as outlier in both COT and fuzzy-statistical procedure.

Experiment 2

Figure 1 shows a leveling network in which the number of uncorrelated observations (observation elevation differences) is 19, the number of essential observations is 10 and the number of redundant observations is 9. The simulated random errors of the observations are listed in Table 1.The standard deviations of the observed elevation differences are $\alpha_0 = \pm 0.14$ (mm) per kilometer of leveling observations (Cen, M. 2003).

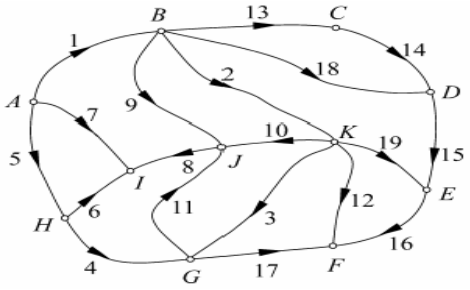

Figure 1. A leveling network

Table1. Simulated leveling measurements for experiment 2

L_i	Elevation Differences (mm)	Distance (km)	Error (mm)
L_1	171 302.9	68	+0.5
L_2	007 513.3	41	+0.9
L_3	038 637.6	35	-1.4
L_4	117 336.5	47	-0.5
L_5	100 117.3	20	-0.8
L_6	309 268.9	22	+1.1
L_7	409 386.2	18	-1.0
L_8	081 197.7	31	+0.2
L_9	156 885.6	23	-0.3
L_{10}	149 372.3	29	+0.8
L_{11}	110 734.7	26	+0.4
L_{12}	134 363.9	15	-0.7
L_{13}	056 263.5	42	+0.2
L_{14}	042 425.1	27	+0.1
L_{15}	025 345.6	17	-0.5
L_{16}	017 843.0	28	+1.2
L_{17}	095 726.3	36	+0.9
L_{18}	098 688.6	32	-0.2
L_{19}	116 520.9	43	-1.0

The following three cases are considered in the test.

1-An outlier is added to observations L_6, L_{13}, L_{14} and L_{18} in each simulation. Four simulations are therefore done for this case.

2-Outliers are added to two of the three observations L_6, L_7 and L_8 in each simulation. A total of four simulations are carried out for this case.

3-Outliers are added to three of the four observations L_3, L_4, L_{11} and L_{17} in each simulation. A total of four simulations are carried out for this case.

Table2. Simulated outliers for experiment 2 and results from the COT method and fuzzy-statistical procedure.(The level of significance $\alpha = 0.0455$)

Case	Simulated outliers (mm)	COT			Fuzzy-Statistical		
		Outliers detected	Error type I	Error type II	Outliers detected	Error type I	Error type II
1	$L_{6=3.2}$	L_6			L_5,L_6,L_7	L_5,L_7	
	$L_{18=3.5}$	L_{18}			L_{15},L_{18}	L_{15}	
	$L_{13=2.8}$	L_{13}			L_{13},L_{14}	L_{14}	
	$L_{14=2.8}$	L_{13}	L_{13}	L_{14}	L_{13},L_{14}	L_{13}	
2	$L_{6=-5.0},L_{7=-6.0}$	L_7,L_8	L_8	L_6	L_1,L_7,L_8	L_1,L_8	L_6
	$L_{7=-1.9},L_{8=4.4}$	L_7,L_8			L_1,L_7,L_8	L_1	
	$L_{6=2.0},L_{8=5.3}$	L_7,L_8	L_7	L_6	L_1,L_7,L_8	L_1,L_7	L_6
	$L_{6=-3.1},L_{7=-6.1}$	L_7,L_8	L_8	L_6	L_1,L_7,L_8	L_1,L_8	L_6
3	$L_{3=5.5},L_{11=4.6},L_{17=-3.5}$	L_{11},L_4	L_4	L_3,L_{17}	L_3,L_{11}		L_{17}
	$L_{3=2.1},L_{4=-6.1},L_{11=9.5}$	L_{11},L_4		L_3	L_3,L_9,L_{11}	L_9	L_4
	$L_{4=-7.4},L_{11=11.5},L_{17=3.4}$	L_{11},L_3,L_{17}	L_3	L_4	L_3,L_{11}	L_3	L_4,L_{17}
	$L_{3=9.7},L_{4=3.7},L_{17=-7.7}$	L_{11},L_4	L_{11}	L_3,L_{17}	L_3,L_{11}	L_{11}	L_4,L_{17}

The outliers added to the observations and results from the COT method and fuzzy-statistical procedure are given in Table 2.
The type I error refers to cases where a method incorrectly flags an observation (or observations) as having outlier although the observation is free from any outlier, while the type II error refers to case where a method is unable to detect an outlier or outliers.
Only while a test procedure can successfully identify all the outliers do we consider that it has no type I or type II error.
In case 1 and 2, the chance of COT method successfully detecting the outliers is less than the chance of fuzzy-statistical procedure. For case 3 the chance of COT successfully detecting all the outliers is equal to the chance of fuzzy-statistical procedure .
The results show that the fuzzy-statistical procedure is more successfully able to detect all the outliers than the COT method, but the COT method is more successfully able to identify the outliers.

Conclusions

In this paper fuzzy-statistical procedure has been introduced for detecting the outliers within the geodetic networks. Both results of using classical hypothesis testing on residuals and fuzzy techniques are used in this procedure.
In COT method that directly uses the residuals with the a-posteriori stochastic information, the residuals which have test values slightly greater or smaller than critical value are also assumed as outliers or not, respectively but fuzzy-statistical procedure uses the output of COT method as a former information and exposes the fuzzy relationship between residuals and observation errors which would be more realistic for detection of outliers, by using the relative redundancy of each observation.
Results of using the both procedures for detection of outliers in two experiments show that they are close to each other, there are some different decisions about some observations by each procedure. The fuzzy-statistical procedure has more ability to detect all of outliers but less ability to identify all of them successfully.
Improving the membership functions of fuzzy sets in fuzzy-statistical procedure by Artificial Neural Network (ANN) techniques, hope us to reach better results.

References

1-Aliosmanoglu,S.,Akyilmaz,O., (2001)."A comparison between statistical & Fuzzy techniques in outlier detection." IAG samposia, Vol.125, Springer, 382-387.

2-Baarda, W.(1968)."A test procedure for use in geodetic networks." Neth Geod Comm Publ Geod . New Ser 2(5):27-55.

3-Cen, M., Li, Z., Ding, X., Zhuo, J.(2003)."Gross error diagnostics before least square adjustment of observation." Journal of Geodesy 77:503-513.

4-Kavouras, M.(1982)."On the detection of outliers and the determination of reliability in the geodetics network." University of New Brunswick, Technical Report No, 87.

5-Zadeh, L.A. (1965): Fuzzy Sets. Information Control 8: 338-353.