ARPHA Proceedings

Proceedings CAUSummit 2024, 0065-0071
doi: 10.3897/ap.7.e0065

World Summit: Civil Engineering-Architecture-Urban Planning Congress

Multiple Regression in Help of the Precise Levelling

Christina Mickrenska[1, a)], Vasil Cvetkov [1]

[1] *University of Architecture, Civil Engineering and Geodesy, Faculty of Geodesy, 1 Hristo Smirnenski Blvd., 1164 Sofia, Bulgaria*

[a)] Corresponding author: kmikrenska_fgs@uacg.bg

Abstract. The precise geometric levelling is the main method for solving many scientific and engineering tasks related to the recent vertical movements of the Earth's crust, the changes in slopes and levels of the oceans and seas, defining of continental and / or state reference height systems, validation of global geopotential models, monitoring of engineering construction, etc. Despite being in use more than a century, there are some incorrect beliefs about accumulation of uncertainties in the results obtained by this geodetic method. These mistaken theories cast their damage over the methodology of execution of measurements, the assessment of the accuracy, processing of the observation data and so on. The purpose of the current research is to throw light on one of the popular beliefs regarding accumulation of differences between both measurements of heights between terminal benchmarks in the precise levelling lines. In the research, the accumulation of the absolute values of the differences between both height measurements in the lines $|D|$ is analyzed by multiple regression. As independent variables we use the square root of the length of leveling lines $\sqrt{L}$, the length of leveling lines L, the sum of the absolute heights in levelling sections in the lines $\sum|h|$, and the absolute value of the height difference between terminal benchmarks |H|. In the interest of plausibility, we analyzed the levelling data from different campaigns in two countries with contrasting climate and geological formations, those of the Third precise levelling of Bulgaria /1975-1984/, the Second precise levelling of Finland /1935-1955/ and the Third precise levelling of Finland /1984-2006/. The results from our analyses show that the multiple coefficients of determination of the differences $|D|$ in respect of independent variables $\sqrt{L}$, L, $\sum|h|$ *and* $|H|$ are 0.29, 0.36 and 0.28 for the Third precise levelling of Bulgaria, the Second precise levelling of Finland and Third precise levelling of Finland, respectively. The most important independent variable for explaining the differences $|D|$ in the analyzed levelling networks, which is the only one statistically significant at 99% confidence level, is the sum of the absolute heights in levelling sections in the lines $\sum|h|$. The traditionally supposed variable, the square root of the length of leveling lines $\sqrt{L}$, is not statistically significant even at 90% confidence level in the case of each mentioned above network. The major conclusion, which we can make on the basis of the research results, is that under 40% of accumulation of the differences $|D|$ between both measurements of heights in the precise levelling lines, we can explain by $\sqrt{L}$, L, $\sum|h|$ and $|H|$. Therefore, we need some new approaches in order to define the maximum accepted values of the differences $|D|$ and levelling accuracy estimators.

Keywords: Multiple regression, precise levelling, accuracy control

Published by CAUSummit and peer-reviewed under the responsibility of "World Summit: Civil Engineering-Architecture-Urban Planning Congress - CAUSummit-2024".

INTRODUCTION

The precise geometric levelling is the main method for solving many scientific and engineering tasks related to the recent vertical movements of the Earth's crust [5], defining of continental and / or state reference height systems [6, 10, 12], validation of global geopotential models [9], monitoring of engineering construction [1], etc. Despite being in use more than a century, there are some incorrect beliefs about accumulation of uncertainties in the results obtained by this geodetic method.

In fact, the precise geometric levelling is a process of n repeated measurements of station elevations along a levelling line, where the number of stations n is usually different from line to line. Figure 1 gives a simple illustration of the measurement process of the elevation h_{AB} between the terminal benchmarks *A* and *B* in a levelling line AB.

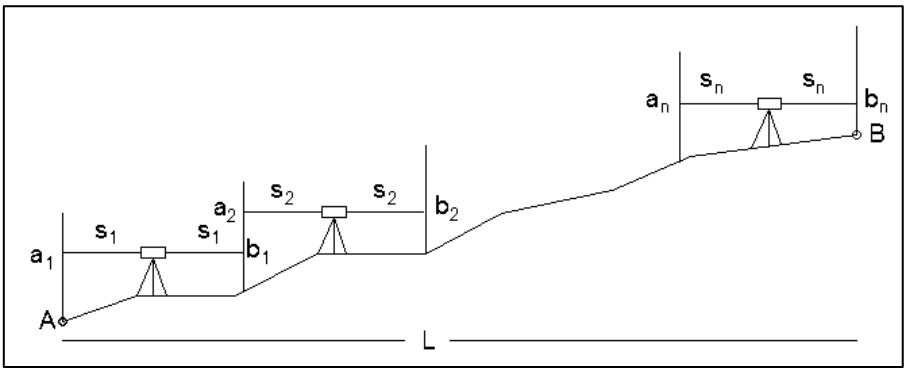

FIGURE 1. The process of measuring the elevation between points A and B in a levelling line [2, Figure 25.1].

According to Figure 1, the elevation h_{AB} can be expressed by equation (1),

$$h_{AB} = a_1 - b_1 + a_2 - b_2 + \cdots + a_n - b_n \tag{1}$$

where a_i and b_i are the backward and the forward staff readings at station *i*. Suppose that expression (2) is valid,

$$S_1 = S_2 = \cdots = S_n = S \tag{2}$$

than we can write the standard errors of the staff readings a_i and b_i by (3).

$$m_{a_1} = m_{b_1} = m_{a_2} = m_{b_2} = \cdots = m_{a_n} = m_{b_n} = m \tag{3}$$

Based on the Gauss's law, the standard error $m_{h_{AB}}$ of the height difference h_{AB} can be given by (4).

$$m_{h_{AB}} = \sqrt{m^2{}_{a_1} + m^2{}_{b_1} + \cdots + m^2{}_{a_n} + m^2{}_{b_n}} = m\sqrt{2n} \tag{4}$$

According to Figure 1, the relationships between the length of levelling line *L*, the number of set up n in the line and the station sights lengths *S*, can be expressed by (5) are (6).

$$L = 2nS \tag{5}$$

$$2n = L/S \tag{6}$$

Putting (6) into (5) and assuming that the accuracy of 1 km levelling m_0 follows equation (7), we can present the standard error $m_{h_{AB}}$ of the height difference h_{AB} by equation (8).

$$m_0 = m/\sqrt{S} \tag{7}$$

$$m_{h_{AB}} = m\sqrt{L/S} = m_0\sqrt{L} \quad (8)$$

Equation (8) expresses the classical view about the accumulation of uncertainties and / or double-run differences in the precise levelling. Based on equation (8), the limits for acceptance of both the double-run differences $|D|$ and the closing errors in levelling loops are defined by formulas (9) and (10), respectively [1, 2, 7, 8, 11, 12].

$$|D| = \pm m_l\sqrt{L} \quad (9)$$

$$W = \pm m_L\sqrt{L} \quad (10)$$

The main objective of the current study is to demonstrate that the real data from three state levelling campaigns do not supports the above presented logic about accumulation of double-run differences $|D|$. In addition, some facts about the accumulation of the random values will be presented in order to explain the accumulation of $|D|$.

MATERIAL AND METHODS

In the current research, we use the data of the Third Levelling of Bulgaria /1975-1984/, the data of the Second Levelling of Finland /1935-1955/ [7], and the data of the Third Levelling of Finland [11]. Contrary to the Finnish data [7, 11], the Bulgarian is not published. This data is available on demand from Agency of Geodesy, Cartography and Cadastre in Sofia, Bulgaria.

In order to explain the accumulation of double-run differences $|D|$, in analysed networks, we performed multiple regression in the form of Stepwise Backward Regression [5, 9]. Our initial equation was given in the form (11).

$$|D| = a + b.\sqrt{L} + c.L + d.\sum|h| + e.|H| \quad (11)$$

As independent variables in equation (11) we use the square root of the length of leveling lines $\sqrt{L}$, the length of leveling lines L, the sum of the absolute heights in the levelling sections in the lines $\sum|h|$, and the absolute value of the height difference between terminal benchmarks |H|. The criteria for excluding an independent variable was its p-value. If some variable had p-values greater than 0.05, we excluded the variable with the greatest p-value. In order to produce unbiased estimations about the squared coefficients of correlation, we intentionally kept the free term of equation (11).

On the next stage of our research, we did some simulations in order to investigate the nature of accumulation of differences between random pair, derived from pre-defined distribution. In our simulations we used the Standard Normal Distribution with an expectation μ=0 and standard deviation σ=1. Our choice of distribution and its parameters was motivated by the classical assumptions about the errors in the precise levelling [7]. According to [7], the errors in the precise levelling "are divided into two classes, accidental and systematic, which are assumed to be independent of each other. The accidental errors are caused by sources that are independent in all successive observations and obey the laws of Gauss. The systematic errors are due to causes acting in a similar manner on successive or adjacent levelling observations; these do not obey the laws of Gauss. They become accidental only for distances exceeding a certain limit Z of order of several tens of kilometers. The total error is the combined influence of both types of errors and is calculated for distances longer than the above limit as the quadratic sum of the accidental and systematic errors". Thus, the Kääriäinen's thought implies that after several tens of kilometers we have only accidental error. If we suppose that the accuracy of the precise levelling is 1mm/$\sqrt{km}$, than the Standard Normal Distribution N(0,1) seems appropriate distribution for our simulations.

The procedure of simulations is a simple one, and the reader can repeat it by Excel, following next steps:

1. Generate one sample random numbers of size 1000 from Standard Normal Distribution. Let us denote this sample by S_1.
2. Generate a second sample random numbers of size 1000 from Standard Normal Distribution. Let us name this sample as S_2.
3. Generate a new sample that contains the differences of the observation from both the first and the second samples that have equal order of appearance, i.e., $P_i = S_1^i - S_2^i$, where i={1, 2, 3, …, 1000}.
4. Generate a final sample of accumulation of the differences $D_i = P_i + D_{i-1}$. The first value in the sample D_i will be equal to P_1. Thus, $D_1 = P_1$.

5. Plot the sample D_i in respect to the order of its members and will result a graph similar to those in Figure 2 below.

RESULTS

This section contains the results from the multiple regression analysis of the data of the Third Levelling of Bulgaria – Table 1, the Second Levelling of Finland – Table 2, and the Third Levelling of Finland – Table 3. Summarized regression analysis results, in order to be easy compared, are given in Table 4.

TABLE 1. Coefficient and their standard errors, t-Stats and p-Values of the independent variables in equation (11) concerning the Third Levelling of Bulgaria /1975-1984/ data.

Step		Coefficients	Standard Error	t Stat	P-value
Initial	a	4.7903	4.4074	1.087	0.278
	b	-1.7437	1.7593	-0.991	0.323
	c	0.3070	0.1773	1.732	0.085
	d	0.0030	0.0016	1.901	0.059
	e	0.0044	0.0021	2.120	0.035*
Final	a	0.8514	0.7390	1.152	0.251
	c	0.1174	0.0366	3.207	0.002*
	d	0.0050	0.0013	3.898	1.0E-4*

* Significance at 95% level.

TABLE 2. Coefficient and their standard errors, t-Stats and p-Values of the independent variables in equation (11) concerning the Second Levelling of Finland /1935-1955/ data.

Step		Coefficients	Standard Error	t Stat	P-value
Initial	a	3.0425	2.6203	1.161	0.248
	b	-0.9815	0.9294	-1.056	0.293
	c	0.1248	0.0829	1.506	0.135
	d	0.0191	0.0072	2.651	0.009*
	e	0.0007	0.0202	0.323	0.747
Final	a	1.0919	0.7131	1.531	0.128
	d	0.0288	0.0036	8.058	4.9E-13*

* Significance at 95% level.

TABLE 3. Coefficient and their standard errors, t-Stats and p-Values of the independent variables in equation (11) concerning the Third Levelling of Finland /1978-2006/ data.

Step		Coefficients	Standard Error	t Stat	P-value
Initial	a	-0.6057	3.7097	-0.163	0.870
	b	0.1951	1.4408	0.135	0.892
	c	0.0461	0.1332	0.348	0.730
	d	0.0251	0.0080	3.127	0.002*
	e	0.0879	0.0271	3.239	0.001*
Final	a	0.5903	1.0412	0.567	0.571
	d	0.0337	0.0050	6.796	9.3E-11*
	e	0.0927	0.0266	3.485	6.0E-4*

* Significance at 95% level.

According to Tables 1-4, the p-values of the independent variables $\sqrt{L}$ and L in equation (11) are greater than 0.05, which means, that these variables have insignificant contribution for forming of the absolute values of the differences between the elevations in the leveling lines $|D|$. This conclusion is valid at 95% significance level. The coefficients of determination of $|D|$ in respect to the predictors $\sqrt{L}$, L, $\sum|h|$ and |H| are 0.29, 0.35 and 0.28 in the case of the Third Bulgarian Levelling data, the Second Finnish Levelling data, and the Third Finnish Levelling data. This simply means that 70% of accumulation on the differences $|D|$ should be sought elsewhere.

TABLE 4. Summary regression statistics.

Step	Regression Statistics	BG III	FIN II	FIN III
Initial	Multiple R	0.539	0.597	0.526
	R Square	0.290	0.357	0.277
	Adjusted R Square	0.277	0.336	0.264
	Standard Error	4.687	5.102	9.848
	Observations	225	129	231
Final	Multiple R	0.524	0.581	0.520
	R Square	0.274	0.338	0.270
	Adjusted R Square	0.268	0.333	0.264
	Standard Error	4.716	5.113	9.848
	Observations	225	129	231

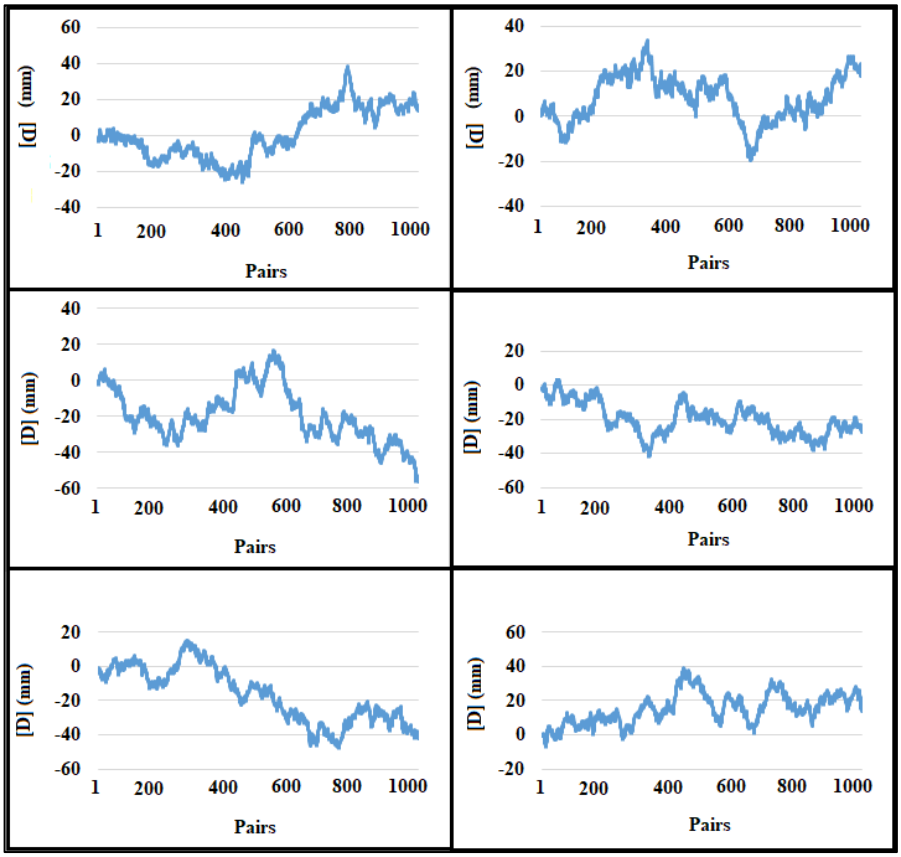

FIGURE 2. Accumulation of differences between 1000 random number pairs, derived from the Standard Normal Distribution N(μ=0, σ=1).

As can be expected, the accumulation of differences between random pairs is a random /stochastic/ process. Regardless this fact, the lines in the charts in Figure 2 do not oscillate around zero axis. Actually, they form some trends.

DISCUSSION

The main objective of the current study was to demonstrate that the real data from three state levelling campaigns do not supports the popular belief about accumulation of double-run differences $|D|$ as a function of a square root of the lengths of levelling lines $\sqrt{L}$. Despite differences in climate and geological conditions in Bulgaria and Finland, the modernization of levelling equipment and the methodology of the levelling, Tables 1-4 paint a common picture.

According to Tables 1-3 it is obvious that the sum of measured elevations along levelling lines $\sum|h|$ is the most important variable for accumulation of the differences $|D|$. Is it a surprise? In fact, it is not, and explanation is simple. Let us for reasons of comfort rewrite equation (1) as (12),

$$h_{AB} = a_1 - b_1 + a_2 - b_2 + \cdots + a_n - b_n \tag{12}$$

where a_i and b_i are the backward and the forward staff readings at station *i*. Based on (12), the variance of the elevation h_{AB} can be determined by (13).

$$\begin{gathered} Var(h_{AB}) = Var(a_1 - b_1 + a_2 - b_2 + \cdots + a_n - b_n) = \\ = Var(a_1) + Var(b_1) + Var(a_2) + Var(b_2) + \cdots + Var(a_n) + Var(b_n) - \\ -2\rho_{a_1 b_1} + 2\rho_{a_1 a_2} - 2\rho_{a_1 b_2} + \cdots - 2\rho_{a_n b_n} \end{gathered} \tag{13}$$

In equation (13) the symbol ρ denotes the coefficient of correlation between two independent staff reading. We would like to know where in equation (13) we have measured length in the plane. The answer is nowhere. We only have measurements of height differences in the form of the staff reading. Therefore, the accuracy of the quantity h_{AB} is function of the accuracy of the staff readings and the correlation among them. As a result, the differences $|D|$ in some lines with a similar length are different. Of course, there is a correlation between the differences $|D|$ and the square root of the lengths of levelling lines $\sqrt{L}$, but this is a spurious correlation because of the heteroscedasticity of the relationship between $|D|$ and $\sqrt{L}$ [3, 4]. Spurious relationships will initially appear to show that one variable directly affects another, but that is not the case. This misleading correlation is often caused by a third factor that is not apparent at the time of examination. This is sometimes called a confounding factor. Our confounding factor is actually the sum of measured elevations along levelling lines $\sum|h|$.

According Table 4, excluding the independent variables $\sqrt{L}$, L and $|H|$ from (11) do not affect the accuracy of the model (11). Regardless of this fact, the accumulation of the differences $|D|$ in the analysed three precise levelling networks with total length approximately 20 000 /or a half of Earth's circumference along the Equator/ can be explained by all above mentioned variables up to 40%. Thus, more than 60% of the variance of the differences stays unexplained by supposed factors. The reason is the stochastic nature of the accumulation of differences between two random observations. Looking at Figure 2, one can find parts in each chart with systematic upward or downward trends, which recalls those several tens of kilometres for which told Kääriäinen [7] and estimated by [12, Appendix 27].

CONCLUSION

The major factor of accumulation of the differences $|D|$ between both measurements of heights in the precise levelling lines is the sum of measured elevations along levelling lines $\sum|h|$. The contribution of such predictors in equation (11) as a square root of the length of levelling lines $\sqrt{L}$, the length of levelling lines *L,* and the absolute values of the height differences between the levels of the terminal benchmarks in levelling lines $|H|$, is insignificant. The full model (11) cannot explain more than 40% of accumulation of the differences $|D|$ between both measurements of heights in the precise levelling lines because there is a significant part with stochastic nature in the differences $|D|$.

Consequently, we need some new approaches, distinctly different from the prescribed [7, 8, 11, 12], in order to define the maximum accepted values of the differences $|D|$ and the levelling accuracy estimators.

REFERENCES

1. A. Angelov, "Geodetic methods in the study for deformation process of high buildings and engineering facilities", Monographic, Second Edition, 2022, https://uacg.bg/filebank/att_22864.pdf , (in Bulgarian).
2. S. Atanasov, "Theory of mathematical processing of geodetic measurements", First Edition, "Technica", Sofia, 1978
3. V. Cvetkov, "Two adjustments of the second levelling of Finland by using nonconventional weights", Journal of Geodetic Science, vol. 13, no. 1, 2023, https://doi.org/10.1515/jogs-2022-0148.
4. V. Cvetkov, "An attempt at explaining of the absolute discrepancies |D| accumulated between the fundamental bench marks of the first and the second order in the third levelling of Bulgaria by some regression models", Annual of the University of Architecture, Civil Engineering and Geodesy, Sofia, Vol. 55, Issue 2, 275-282, 2022
5. N. Draper and H. Smith, "Applied Regression Analysis", Second Edition, John Wiley & Sons, 1981.
6. S. Gospodinov, E. Peneva and P. Penev, "A specific approach to least squares adjustment of the state levelling network", 22nd International Multidisciplinary Scientific GeoConference SGEM 2022, Albena, Bulgaria, 2020, https://doi.org/10.5593/sgem2022/2.1/s09.20.
7. E. Kääriäinen, "The Second Levelling of Finland in 1935–1955", Publications of the Finnish Geodetic Institute No. 61, 1966, Helsinki, http://hdl.handle.net/10138/347072.
8. I. Kunchev, "On standards and specifications for vertical control surveys", Annual of the University of Architecture, Civil Engineering and Geodesy – Sofia, Vol. 54, Issue 2, pp. 253 – 262, 2021, ISSN 1310-814X, ISSN 2534-9759 (on-line), in Bulgarian.
9. D. Montgomery and G. Runger, "Applied Statistics and Probability for Engineers (6th ed.)", Wiley, 2014, ISBN-13 9781118539712.
10. E. Peneva and I. Georgiev, "Evaluation of the Quasigeoid Models EGG97 and EGG07 with GPS/levelling Data for the Territory of Bulgaria", In IAG Book series "Gravity, Geoid and Earth Observation", International Association of Geodesy Symposia 135, Mertikas (ed.), Springer-Verlag Berlin Heidelberg, pp. 303-307, 2010, ISBN 978-3-642-10633-0, https://doi.org/10.1007/978-3-642-10634-7_39 .
11. V. Saaranen, P. Lehmuskoski, M. Takalo and P. Rouhiainen, "The Third Precise Levelling of Finland". FGI Publications No. 161, Kirkkonummi, 2021, ISBN 978-951-48-0266-9 (print), ISSN 2342-7353 (online)
12. Specifications for precise levelling I and II order, GUGK, 1980, Sofia, in Bulgarian.
13. Y. Tsanovski, "GNSS Measurements for the Needs of Precise Height Determinations in Bulgaria", Annual of the University of Architecture, Civil Engineering and Geodesy –Sofia, Vol. 54, Issue 2, pp. 211 –218, 2021, ISSN 1310-814X, ISSN 2534-9759 (on-line), in Bulgarian.