Atti Accademia Nazionale dei Lincei

Classe Scienze Fisiche Matematiche Naturali

Rendiconti

Arnaldo Chiarini, Lamberto Pieri

Statistical analysis of models for testing discrepancies in high precision levelling

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti, Serie 8, Vol. **43** (1967), n.6, p. 502–508.

Accademia Nazionale dei Lincei

<http://www.bdim.eu/item?id=RLINA_1967_8_43_6_502_0>

Geodesia. — *Statistical analysis of models for testing discrepancies in high precision levelling.* Nota di Arnaldo Chiarini e Lamberto Pieri, presentata [*] dal Socio P. Dore.

RIASSUNTO. — In un precedente lavoro [1] degli Autori è stata compiuta un'analisi degli errori della livellazione di precisione e precisamente della variabile aleatoria

$$x_{ij} = \frac{\rho_{ij}}{\sqrt{R_{ij}}}$$

(ove ρ_{ij} è la discrepanza fra le misure in andata e ritorno della differenza di quota di due caposaldi consecutivi della i–esima linea ed R_{ij} è la loro distanza), come nuovo contributo all'annoso problema, ancora aperto, degli errori della livellazione di precisione. Il campione esaminato consta di una rete parziale della livellazione italiana.

La presente Nota approfondisce criticamente tale metodologia dimostrando, sulla base del campione esaminato (di circa Km. 900 della livellazione italiana), che la variabile aleatoria studiata è quella che rende minima la dipendenza delle discrepanze pesate dalla distanza.

In una seconda parte della Nota si studia un modello di regressione che considera la dipendenza delle discrepanze dalla distanza e dalla differenza di quota tra due caposaldi consecutivi. I risultati di questo studio mostrano che il modello:

$$\rho_{ij} = \mu_i \sqrt{R_{ij}} + \varepsilon_{ij} \sqrt{R_{ij}}$$

in cui μ_i è una costante per ogni linea i ed ε_{ij} è una variabile generalmente normale con valore medio zero, sembra essere il più adatto.

1. Introduction.

In a previous work [1] we have shown some aspects of high precision levelling errors by means of non-parametric statistical methods, choosing a sample of 15 lines of the Italian high precision levelling net and studying, according to classical theory, the behaviour of the random variable

$$x_{ij} = \frac{\rho_{ij}}{\sqrt{R_{ij}}} \tag{1}$$

where ρ_{ij} : discrepancy between the direct and reverse measurements of the relative height of consecutive bench marks in the i–th line;
R_{ij}: distance between consecutive bench marks in the i–th line.

The purpose of this choice was the study of a variable which is independent of the distance and hence the study of homogeneous sets of data. In this work we examine the consistency of such a hypothesis, also testing the discrepancies by means of a regression analysis.

[*] Nella seduta del 9 dicembre 1967.

2. STUDY OF CONSISTENCY.

The choice of model (1) was made by taking into account well known classical considerations; however, this "a priori" model needs an adequate experimental check. The check will be carried out using a model for a random variable x_{ij} depending on a parameter α the value of which can be obtained with a maximum likelihood method: the comparison of this new model with (1) will give us the measure of the consistency of our assumption.

2.1. *Choice of $x_{ij}(\alpha)$.*

The problem we wish to study consists in finding a random variable, function of discrepancy and distance, which minimizes the dependence on the distance itself, according to the above-mentioned requirements. Of course there is something arbitrary in the choice of the form of this function, but the one which follows (2) seems the most practical for ascertaining whether the model (1) is to be preferred in comparison with other similar hypotheses.

We assume therefore:

$$x_{ij}(\alpha) = \frac{\rho_{ij}}{R_{ij}^{\alpha}} \tag{2}$$

where α is a parameter to be determined together with a convenient interval.

2.2. *Procedure.*

The procedure we have devised for our study consists of the following steps. First we calculate a non-parametric correlation coefficient of an appropriate function of $x_{ij}(\alpha)$ and R_{ij}. To be precise, the Spearman [3] correlation coefficient of the two variates

$$|x_{ij}(\alpha) - \bar{x}_i(\alpha)| \quad \text{and} \quad R_{ij}$$

in each line for different values of α around $\bar{\alpha} = 0.5$.

The reason for the choice of $|x_{ij}(\alpha) - \bar{x}_i(\alpha)|$, where $\bar{x}_i(\alpha)$ is the average value of $x_{ij}(\alpha)$ in the i-th line, depends upon two facts:

1) $x_{ij}(\alpha)$ might have a mean value different from zero and the use of $x_{ij}(\alpha)$ only, could mask effects of dependence;

2) since we are interested in studying the correlation between the magnitude of $x_{ij}(\alpha) - \bar{x}_i(\alpha)$ and R_{ij} irrespective of the sign, an appropriate variable seems to be $|x_{ij}(\alpha) - \bar{x}_i(\alpha)|$ [1].

The reason for the choice of the Spearman coefficient is that the variates taken into consideration cannot be assumed as normal and therefore the usual Pearson estimate of the correlation coefficient cannot be properly used.

[1] Of course other functions could have been chosen, e.g.

$$[x_{ij}(\alpha) - \bar{x}_i(\alpha)]^2.$$

The Spearman coefficient [2] is computed as follows:

given a set of n observations $\{\xi_i, \eta_i\}$ $(i = 1, 2, \cdots, n)$ of two variables ξ, η of which only one must be necessarily a random variable, and defining

$$\hat{\xi}_i = \text{the rank of } \xi_i$$

$$\hat{\eta}_i = \text{the rank of } \eta_i$$

$$d_i = \hat{\xi}_i - \hat{\eta}_i$$

$t_{\hat{\xi}}$, $t_{\hat{\eta}}$ the number of tied values of $\hat{\xi}$ and $\hat{\eta}$ respectively

$$T_{\hat{\xi}} = \frac{t_{\hat{\xi}}^3 - t_{\hat{\xi}}}{12}$$

$$T_{\hat{\eta}} = \frac{t_{\hat{\eta}}^3 - t_{\hat{\eta}}}{12},$$

the Spearman coefficient is

$$r = \frac{\dfrac{n^3 - n}{6} - \Sigma T_{\hat{\xi}} - \Sigma T_{\hat{\eta}} - \sum_1^n {}_i d_i^2}{2\sqrt{\left(\dfrac{n^3 - n}{12} - \Sigma T_{\hat{\xi}}\right)\left(\dfrac{n^3 - n}{12} - \Sigma T_{\hat{\eta}}\right)}}$$

where the sums $\Sigma T_{\hat{\xi}}$ and $\Sigma T_{\hat{\eta}}$ are extended to all the ties.

For a sufficiently large n $(n > 10)$ the random variable

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

has a Student distribution with $n - 2$ degrees of freedom.

We calculate for each line the value of $t_i(\alpha)$ for various values of α. A likelihood function is then computed using for each line the probability:

$$F(t_i(\alpha)) = P(t > |t_i(\alpha)|) = 1 - 2\left|\Phi(t_i(\alpha), n_i - 2) - \frac{1}{2}\right|$$

where $\Phi(t, n)$ is the Student distribution function with n degrees of freedom. Finally the logarithm of the product of these functions, i.e.

$$L(\alpha) = \log \prod_{i=1}^{k} F(t_i(\alpha)),$$

where k is the number of independent samples, is calculated.

The value of α that maximizes $L(\alpha)$, minimizes the dependence of x_{ij} on R_{ij}, since the values of $t_i(\alpha)$, and therefore $r_i(\alpha)$, are the minimum ones.

2.3. *Results of calculations.*

The calculation was carried out taking into consideration the only 11 lines for which we previously obtained results of randomness using the model (1). The exclusion of the remaining 4 lines seems necessary in a search for consistency, like the one undertaken here.

The calculation of the values of L (α) was carried out and the results appear in Table I.

As is well-known [3] the estimator α of a parameter α_0 obtained with a maximum likelihood procedure is asymptotically normally distributed; therefore L (α) is asymptotically equal to $-\frac{1}{2}\left(\frac{\alpha-\alpha_0}{\sigma_\alpha}\right)^2 + k$, where σ_α^2 is the variance of the estimator and k is a normalisation constant.

This procedure, as a first approximation, can generally be used even when the sample is rather small; therefore we have fitted, using the least square method, the L (α) values with a parabola:

$$L'(\alpha) = a\alpha^2 + b\alpha + c,$$

obtaining:

$$a = -42.55$$

$$b = +53.74$$

$$c = -27.00$$

with a very good approximation, as can be seen from Table I.

The estimators of α_0 and σ_α are obtained from the following relations [3]:

$$\sigma_\alpha = \left(-\frac{1}{\frac{d^2 L(\alpha)}{d\alpha^2}}\right)^{1/2} \quad , \quad \frac{dL(\alpha)}{d\alpha} = 0$$

from which, substituting for L (α) its approximate value L$'$ (α), we obtain

$$\hat{\sigma}_\alpha = \frac{1}{\sqrt{-2a}} = 0.11$$

$$\hat{\alpha} = -\frac{b}{2a} = 0.63 .$$

We may now test the null hypothesis $\alpha_0 = 0.5$ corresponding to the classical theory against the alternative one: $\alpha_0 = 1$, as proposed by other authors [5]. According to the above-mentioned results the acceptance interval of the null hypothesis with a 5 % significance level is given, with a good approximation, using the most powerful test [3], by:

$$-\infty < \alpha \leq 0.68$$

in which falls the value $\hat{\alpha} = 0.63$.

On the other hand the power of the criterion relative to the comparison of both hypotheses is 0.998, corresponding to a 2 ‰ probability of an error of the second kind.

Therefore the assumption of the value $\alpha = 0.5$ seems confirmed by our analysis.

TABLE I.

α	L (α)	L′ (α)	L′ (α) — L (α)	$\dfrac{\text{L}'(\alpha) - \text{L}(\alpha)}{\text{L}'(\alpha)}$
0,00	— 26,825	— 26,999	— 0,174	— 0,006
0,25	— 16,535	— 16,224	+ 0,311	— 0,019
0,50	— 10,820	— 10,767	+ 0,053	— 0,005
0,75	— 10,350	— 10,631	— 0,281	+ 0,026
1,00	— 15,845	— 15,813	+ 0,032	— 0,002
1,25	— 16,375	— 26,315	— 0,060	— 0,002

L′ (α) = values computed by the regression model.

3. REGRESSION ANALYSIS OF DISCREPANCIES.

The indications we obtain from the previous results are that model (1) seems adequate for describing a random variable independent of the distance but, at this stage of our study, we extend our model introducing other kinds of dependence, limiting our choice to linear dependence only.

3.1. *Method of study.*

Systematic effects, pointed out in [1] using model (1), are represented by the mean values of x_{ij} in each line, that are not equal to zero. Therefore we can transform the model in this way

$$x_{ij} = \mu_i + \varepsilon_{ij} \tag{3}$$

or

$$\rho_{ij} = \mu_i \sqrt{\text{R}_{ij}} + \varepsilon_{ij} \sqrt{\text{R}_{ij}} \tag{3'}$$

where μ_i is a constant for each line (mean value) and ε_{ij} is generally a normal variate with mean value zero and variance σ_i^2.

It seems quite natural to try to reduce the variance of ρ_{ij} using a more complicated model in which terms depending on R_{ij} and $|\Delta\text{H}_{ij}|$ (difference in elevation of two consecutive bench marks) are taken into account.

For this purpose we have chosen the model

$$\begin{aligned} \rho_{ij} = a_i + b_i \sqrt{\text{R}_{ij}} + c_i \text{R}_{ij} + d_i \text{R}_{ij}^2 + e_i \sqrt{|\Delta\text{H}_{ij}|} + f_i |\Delta\text{H}_{ij}| + \\ + g_i |\Delta\text{H}_{ij}|^2 + \varepsilon_{ij} \sqrt{\text{R}_{ij}} \end{aligned} \tag{4}$$

in which

a_i through g_i are constants in each line and ε_{ij}, as above, is a random variate with mean value zero.

A stepwise linear regression calculation [4], which enables us to take into account only the terms whose coefficients are significantly different from zero, was in effect carried out using the variate

$$x_{ij} = \frac{\rho_{ij}}{\sqrt{R_{ij}}} = \frac{a_i}{\sqrt{R_{ij}}} + b_i + c_i \sqrt{R_{ij}} + d_i \sqrt{R^3_{ij}} + e_i \frac{\sqrt{|\Delta H_{ij}|}}{\sqrt{R_{ij}}} + f_i \frac{|\Delta H_{ij}|}{\sqrt{R_{ij}}} + g_i \frac{|\Delta H_{ij}|^2}{\sqrt{R_{ij}}} + \varepsilon_{ij}$$

by reducing all terms to the same weight.

The computation was carried out for all the 15 lines irrespective of the lack of normality and of randomness in some of them, in order to bring out possible anomalous effects. The results show that, except for lines 5, 7 and 15, model (3′) is sufficient to explain the variance of ρ_{ij}. The coefficient of model (4) for lines 5, 7 and 15 and relative F values obtained from the analysis of variance of multiple regression, are shown in Table II.

Table II.

Line	Line number	a_i	b_i	c_i	d_i	e_i	f_i	g_i	F obtained from the analysis of variance of the regression	n_1	n_2	$F_{n_1 n_2 ; 0.05}$
Firenze–Bologna	5		0.533			—0.283	0.035		3.25	2	186	3.05
Ferrara–Padova	7		0.533			—0.319			4.68	1	90	3.95
Rimini–Bologna	15	0.253	0.351					—0.01	3.37	2	109	3.08

We must notice however that line 7 showed non random behaviour and lines 5 and 15 non normal behaviour. A more detailed analysis of the abovementioned lines should be carried out, but in agreement with the opinions of other authors [5], there seems no point, on the basis of these results, in introducing a more complicated model for ρ_{ij} and (3′) seems to be the most appropriate.

An extension of our study to the whole Italian high precision levelling net and eventually to other nets is planned to give further information on this topic. A possible extension of such an analysis could be the creation of a model in which other parameters such as the temperature or refraction coefficient of the air can be taken into account.

4. Conclusions.

The results of our studies seem to confirm the validity of model (3′) in the absence of non-randomness effects according to the classical theory from both points of view: consistency (§ 2) and regression analysis (§ 3). We emphasize however the necessity of carrying out similar analyses on broader samples, possibly covering nets of different countries, to confirm these results.

BIBLIOGRAPHY.

[1] A. CHIARINI and L. PIERI, *Studio statistico degli errori della livellazione geometrica di precisione*, « Atti dell'Accademia delle Scienze dell'Istituto di Bologna », serie XII, tomo IV (1967).

[2] C. SPEARMAN, *The proof and measurements of association between two things*, « American Journal of Psychology », n° 15, 72–101 (1904).

[3] M. G. KENDALL and A. STUART, *The advanced theory of statistics*, vol. 2, London (1961).

[4] M. A. EFROYMSON, *Multiple Regression Analysis* in « Mathematical Methods for Digital computers », edited by A. Ralston and H. S. Wilf – New York (1960).

[5] A. M. WASSEF, *Statistical analysis of levelling errors*, Progress Report 1960–1967 (To be published).