KEY QUESTIONS OF SAMPLING FREQUENCY ESTIMATION DURING SYSTEM CALIBRATION, ON THE EXAMPLE OF THE KIS-BALATON WATER PROTECTION SYSTEM'S DATA SERIES

József Kovács [1], István Gábor Hatvani [1*], Ilona Székely Kovács [2], Pál Jakusch [3], Péter Tanos [1] János Korponai [4]

[1] Eötvös Loránd University, Department of Physical and Applied Geology, H-1117 Budapest, Hungary. E-mail: kevesolt@geology.elte.hu; *hatvaniig@gmail.com, tanospeter@gmail.com

[2] Budapest Business School, Institute of Methodology, H-1054 Budapest, Hungary. E-mail: iszekely@geology.elte.hu

[3] University of Pannonia, Georgikon Faculty of Agriculture, H-8360 Keszthely, Hungary. E-mail: korponai.janos@nyuduvizig.hu

[4] West Transdanubian Water Authority, Department Kis-Balaton, H-8360 Keszthely, Hungary. E-mail: jakusch.pal@gmail.com

Abstract

In this study the practice of sampling frequency estimation is described on data series from the Kis-Balaton Water Protection System.

The main aim was to point out the milestones, and common problems of sampling frequency estimation using variograms. Firstly, the importance of sampling frequency estimation during the calibration of the system is emphasized, embedding it in the idea of sustainable development. The applied method itself, the variogram is then discussed. This

is a function used in geostatistics; basically, it is the expected squared increment of the values between locations x and y (Wackernagel, 2003).

The results of the variogram analysis pointed to a series of problems and solutions which must be faced when an attempt is made to prepare the data sets for analysis. These problems mainly concern the removal of periodicity as a "special trend".

The final results explicitly show a 7 day (or less) sampling frequency is needed in the case of the total phosphorous parameter in order to permit long term assumptions and interventions to be based on it. The method described in the article can be used in the analysis of many other parameters complying with the needs of the variogram analysis, so it may prove to be highly useful to any scientist working with data received from sampling.

Keywords: sampling frequency estimation, monitoring system calibration, variogram, Kis-Balaton Water Protection System,

Összefoglalás

Alábbiakban leírt kutatásban a mintavételezési gyakoriság becslése kerül bemutatásra a Kis-Balaton Vízvédelmi Rendszer (KBVR) adatsorán.

A kutatás elsődleges célja, hogy rávilágítson a variogramok segítségével történő mintavételezési gyakoriság becslés menetének problémáira. Első lépésként a mintavételezési gyakoriság becslésének rendszer kalibrálásban betöltött lényeges szerepe kerül kiemelésre a fenntartható fejlődés szemszögéből, majd maga a becslési módszer kerül bemutatásra. A mintavételezési gyakoriság becslését a geostatisztikában több függvénnyel is el lehet végezni, jelen esetben az empirikus félvariogram került alkalmazásra.

A variogram vizsgálatok alatt számos problémába merülhet fel, elsősorban az adatok előkészítése során, amikor a periódus, mint egy „speciális trend" kerül eltávolításra. Ezen problémák és lehetséges megoldásai is bemutatásra kerülnek.

A KBVR összes foszfor paraméterének variogram vizsgálatának eredménye kimutatta, hogy 7 napos vagy annál kisebb mintavételezési gyakoriság szükséges ahhoz, hogy a paraméter adatsoraiból visszaállíthatóak legyenek a vizsgált területen zajló folyamatok. (Ennek a követelménynek jelenleg is eleget tesz a Kis-Balaton Üzemmérnökség Laboratóriuma).

A leírt módszer a tudományok minden területén alkalmazható ahol mintavételezésből származó adatsorok állnak rendelkezésre és lényegi kérdés a pontos mintavételezési gyakoriság meghatározása, hogy a kutatásokból reprezentatív eredmények és szakmailag megalapozott döntések születhessenek.

1. Introduction

"Sustainable development is a pattern of resource use that aims to meet human needs while preserving the environment so that these needs can be met not only in the present, but also for generations to come (Anonymous, 1987)". This was one of the concepts that urged scientists to make environmental monitoring a basic element of every case study. Therefore, during the past few decades a great "mass of data" has been produced describing the environments we live in.

The question is whether or not the data produced in diverse fields of science and by different sampling methods conforms to standards so it would produce a representative data set. This is a highly important problem because these data define and limit the mathematical methods that can be

applied to them, and, in this way, the conclusions and interventions as well.

Following the outline described above, the sample itself has to be defined from a statistical perspective. The statistical sample is a random variable X with distribution F, a random sample of length n =1,2,3,... is a set of n independent, identically distributed random variables with distribution F (Samuel, 1962). The data in fields, such as meteorology, hydrology, and limnology are not always independent. For example, an annual one pH sample from 10 consecutive years should be independent; however, when the data follow each other at short intervals (daily sampling) they are not independent because they are too close to each other in time. Examining the results of such daily sampling results in a time series is received where each datum cannot be interchanged with another. In many cases, because of environmental and man-made impacts, the samples cannot be described with the same distribution. In conclusion, the sampling of a certain parameter should be as frequent as needed for the sample to include each important property of the aggregate. If all of these requirements are fulfilled the data set could allow the estimation of the expected value.

The final question is, what should the sampling frequency of a certain process be, so that the aims of the study could be achieved and estimates could be given regarding future events?

Viewed from this perspective, it is obvious that higher the parameter's variability within h (time or space) distance, the more frequent the sampling should be. Many functions are known that are able to describe the variability of a parameter in h (time of space) distance. In this case, the basic function of spatial statistics, the variogram was used, in determining the optimal sampling frequency (Füst, 1997; Márkus et al., 1999; Dryden, 2004) during system calibration (Füst and Geiger, 2010, 2011).

2. Materials and methods

2.1 Data-series acquired from the Kis-Balaton Water Protection System (KBWPS)

In the course of this research, the Total phosphorous (mg l^{-1}) parameter of the KBWPS's Z11 (Balatonhídvég) sampling location was examined for the time interval 01.01.1988-31.12.2006. The data was received from the laboratory of the West Transdanubian Water Authority's Kis-Balaton Department, where daily sampling was conducted following the water authority's national code of practice, and analyzed in the same laboratory during the investigated time period (Kovács et al., 2010).

2.2 The variogram

Three functions are used in geostatistics for describing the spatial or the temporal correlation of observations: these are the correlogram, the covariance and the semivariogram. The variogram and the semivariogram originated from the variogram can be described mathematically as follows (Füst, 2004; Molnár and Füst, 2002; Molnár et al., 2010). Let $Z(x)$ and $Z(x+h)$ be the values of two sampled parameters in h distance from each other. Distance *h* can be measured in time or space. The variance of the difference of values $Z(x)$ and $Z(x+h)$ is $D^2[Z(x)-Z(x+h)]=D^2[Z(x)]+D^2[Z(x+h)]-2COV[Z(x),Z(x+h)]$. In case of samples taken from the same population we could assume that $D^2[Z(x)]=D^2[Z(x+h)]$ so $D^2[Z(x)-Z(x+h)]=2D^2[Z(x)]-2COV[Z(x),Z(x+h)]=2\gamma(h)$. Function $2\gamma(h)$ is called the parameter's variogram, while $\gamma(h)$ is its semivariogram. If we introduce the simplified notation $D^2[Z(x)]=D^2(x)$, then $\gamma(h)=D^2(x)-g(h)$. If the sample size is *N* in case of discrete samples from a nominally distributed population, the semivariogram could be calculated with the

Matheron algorithm (Matheron, 1965): $\gamma(h)=\frac{1}{2N(h)}\sum_{i=1}^{N(h)}\left[Z(x_i)-Z(x_{i+h})\right]^2$.

In the case of non-nominal distribution, there are different transformations that can ensure it. Many publications in the geostatistical literature however refer to this as an unimportant distribution type (Clark, 1979., Cressie, 1993).

In practice $Z(x_i)\geq 0 \quad (i=1,2,...,n) \quad \sigma^2[Z(x)]\geq g(h)\geq 0$, so theoretically the semivariogram can only take values from the $0\leq\gamma(h)\leq\sigma^2[Z(x)]$ range. The most important properties of the function are as follows *(Fig. 1)*:

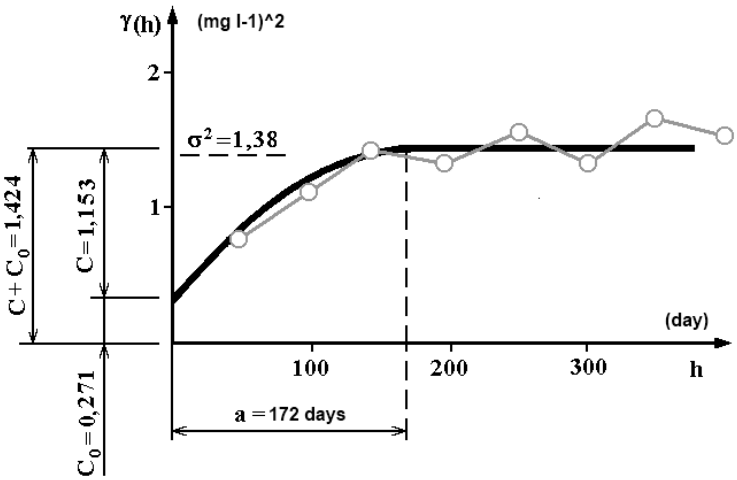


Fig. 1.

Properties of the variogram, with the curve (with the circles) indicating the empiric and the other the theoretical semivariogram (based on Füst, 1997)

Continuity can be seen from the $\gamma(h)$ function's accession. If the function does not start from the origin of the coordinates, there must have been drastic changes in the parameters' processes. This is called the

nugget effect. It is the height of the jump ($C_0 \geq 0$) of the variogram at the discontinuity at the origin.

If the semivariogram does not have an uprising part, the empirical semivariogram's points (the circles on the graph) will align along an *h* line parallel to the abscissa. If this occurs, the continuity has fully ceased.

Sill ($C + C_0$) is the limit of the variogram tending to infinity lag distances. However *C* itself is the reduced sill.

Range is the distance in which the difference of the semivariogram from the sill becomes negligible. In models with a fixed sill, it is the distance at which this is first reached; for models with an asymptotic sill, it is conventionally taken to be the distance when the semi variance first reaches 95% of the sill.

If the semivariogram stabilizes at $h \to \infty$ after a fast ascent, then the parameter is stationary. However if $\gamma(h)$ is an increasing function (if $h \to \infty$ then $\gamma(h) \to \infty$), the parameter is non-stationary.

The empirical semivariograms can be approximated with many theoretical functions. However, discussing these is not an aim of the study (Füst, 2004; Molnár és Füst, 2002; Molnár et al., 2010).

Our estimation of sampling frequency is based on the fact, that samples outside the range (let it be temporal or spatial) are -in practice- independent. In other words, the samples taken outside the range (temporal or spatial) can only describe the vicinity of their environment, and in this way cannot provide scientists with information regarding the processes' deeper structure.

3. Example of range estimation in practice on the KBWPS' data series

As described previously the daily sampled total phosphorous (mg l^{-1}; TP) parameter of the KBWPS was analyzed using variograms. A part of the original data series can be seen in *Fig.2,* while the empiric semivariogram of the total time interval on *Fig.3*.

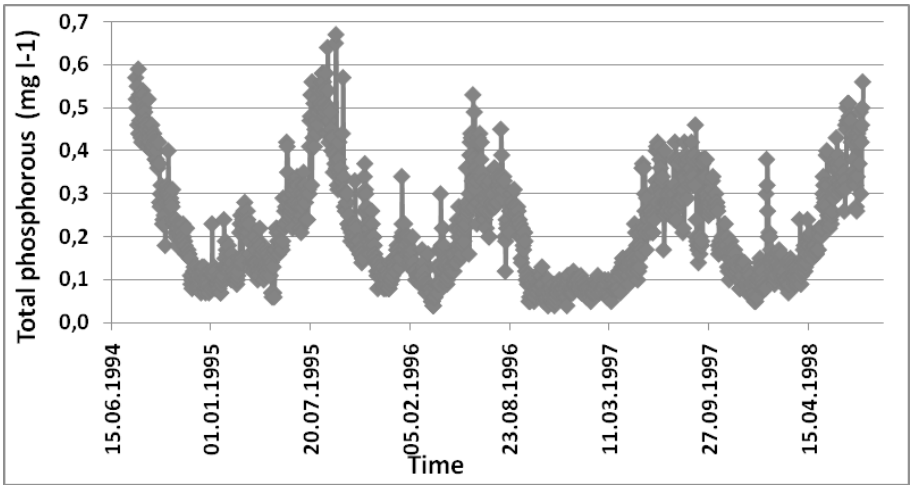


Fig. 2.
A short period of the original total phosphorous (mg l^{-1}) data series, showing the periodical structure of the signal.

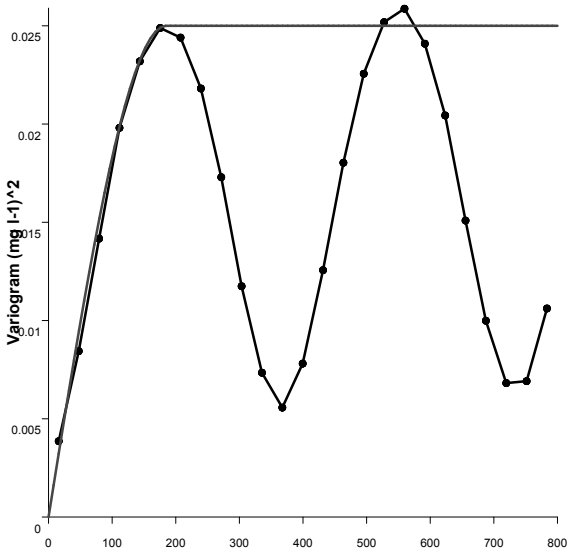


Fig. 3 Empiric (dotted line) and theoretical semivariograms of the total investigated time period (01.01.1988-31.12.2006) of the total phosphorous parameter (mg l^{-1}), indicating a 182 day range.

The fitted semivariogram is spherical. In this case, the nugget effect is zero. This is important because the nugget effect is the characteristic that describes the margin of error caused by the sampling method or instrument, the parameters' accuracy, and the parameter's change in time or space. The estimated range in *Fig. 3* is 182 days; this concurs with the facts stated in the Nyquist–Shannon sampling theorem for processes describable with annual periodicity (Shannon, 1998). If we evaluate the results from a sampling perspective, it becomes clear that a sampling frequency of 182 days can only give us information about annual processes. (Here we must state that in most cases of the KBWPS's parameters the 182 day range was found, which indicates annual periodicity). If there is a need to observe processes describable only with periods smaller than one year, the sampling frequency must be adjusted accordingly to them. Hence, the smallest range acquired from the empiric semivariogram should be used.

If we look at the TP's semivariogram for an interval of 40 days *(Fig.4)* no range can be seen; it is covered by the "special trend" caused by periodicity seen in Fig. 3.

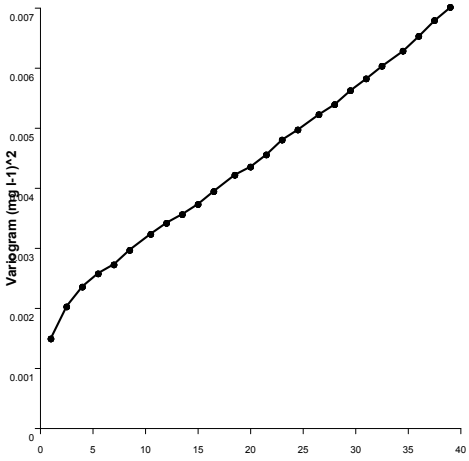
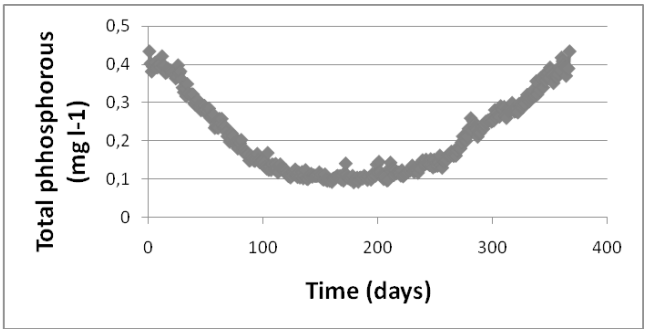


Fig.4. The total phosphorous' empiric semivariogram for an interval of 40 lags (days), where no range can be seen.

So the periodicity -as a special trend- must be removed, because we are interested in the more detailed processes within the residual. The most obvious method would be to remove a repeated average function, (generated from the average of the same day of every year *(Fig. 5)* from the realization of the parameter's time series. The variogram fitted on the residuals can be seen in *Fig. 6.*



Annual period generated from the average of the same day of every year investigated

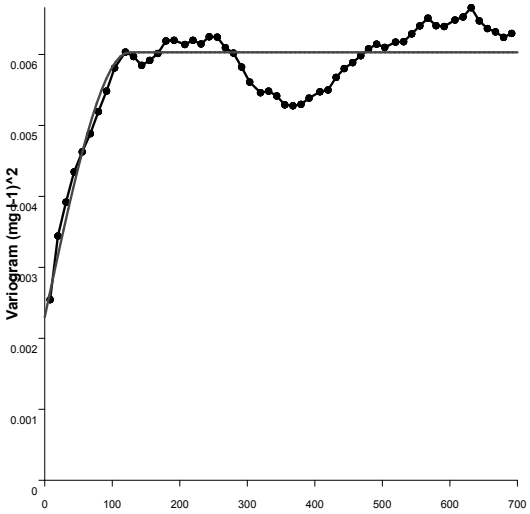


Fig. 6.

The empiric (dotted line) and theoretical semivariograms fitted on the total phosphorous residuals, after the removal of the generated average annual period

According to *Fig. 6* a long period can still be found in the signal. In other words the trend removal was unsuccessful. We must ask ourselves the question why?

The answer can be found on *Fig. 7*, which describes the time elapsed between the peaks of the periods.

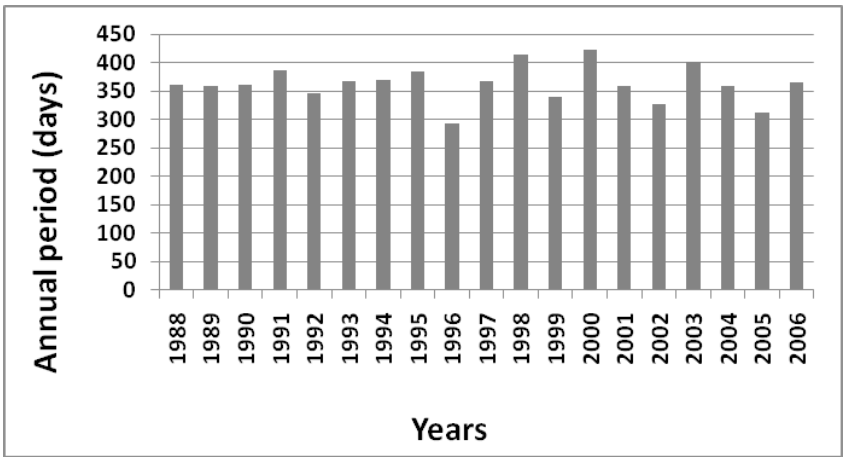


Fig. 7.
Time elapsed (days) between the peaks of the periods.

It is true that the average of these values is 365, but the smallest is 292, and the highest is 422 days. The conclusion is that the period's length is not constant, so the removal of the average period *(Fig. 5)* from the whole time series can only be successful if all of the years consist of 365 day periods. If not, than we are the ones implanting the periodic process into the signal. This is known as the Slutsky effect (Slutsky, 1937).

Taking these facts into consideration, the range (time) examination of a more than one year interval (01.06.1996.-19.08.1997.) seemed the most appropriate. In contrast to the previously discussed trend removal, in this case a polynomial trend was estimated, and because the interval was longer than one year, instead of a quadratic, a polynomial (n=5) trend was removed *(Fig. 8)*.

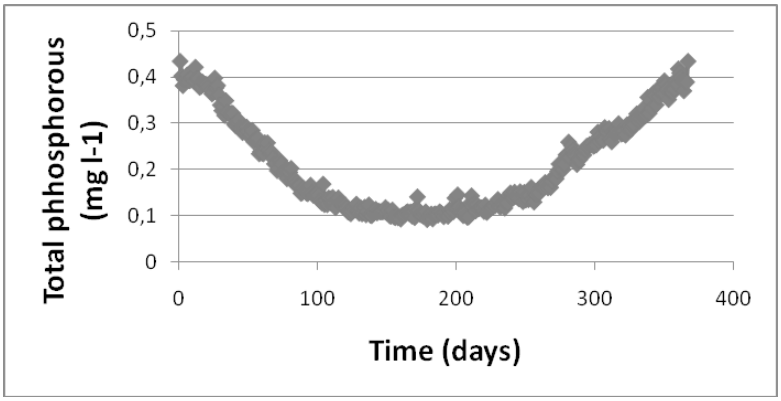


Fig. 8. The x-y scatter plot of the examined period (01.06.1996.-19.08.1997.) with the fitted polynomial (n=5) regression line.

The variogram fitted to the residual can be seen on *Fig. 9.* The method was repeated for many other time periods and similar results were produced. The principal range was to be found at 8-10 day lags. These results were acceptable. The slight distortions are caused by the range's characteristic that it is an estimated aleatory variable, because all of its estimated values depend on the examined interval's data, as in a sample realization.

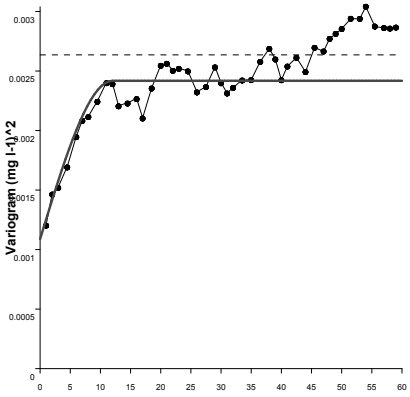


Fig. 9. Empiric and theoretical (dotted line) semivariograms fitted to the total phosphorous residuals indicating an 8-10 day range after the removal of the polynomial (n=5) trend.

To ensure reliability in future sampling, the employment of a lower frequency, than the one in the results is advised. In regard to the TP on sampling location Z11 a frequency of 7 days is suggested. A minor break can be seen in *Fig. 9* at 3 day lag. We assumed that there are significant changes in the parameters processes after three days, and that a second 3 day range may be determined as well. However further investigations (using more effective trend removal methods) should be carried out to answer this question.

4. Summary

With the method described above we were able to determine exactly the most appropriate sampling frequency for the TP parameter of the KBWPS's Z11 sampling location. Fortunately this suggested sampling frequency is equal to the one already applied (7 days).

Knowledge of the ideal sampling frequency is a key question in monitoring practice. If it is not chosen wisely, the data obtained is useless, or can only be used for certain analyses. This is what science nowadays cannot afford, because first of all data is valuable whether expressed in terms of goodwill or money, and secondly data not representing the area it was received from can lead to unprofessional interventions and, as a result, further damage to the environment.

Acknowledgements

This work was sponsored by the project of TÁMOP-4.2.2/B-2010-0025.

References

Anonymous, 1987. Our Common Future, Report of the World Commission on Environment and Development, World Commission on Environment and Development. Published as Annex to General Assembly document A/42/427, Development and International Co-operation: Environment August 2, 1987.

Clark, I., 1979. Practical geostatistics. Applied Science Publishers LTD. London, 151p.

Cressie, N., 1993. Statistics for Spatial Data, Revised Edition, Wiley, New York, 928p.

Dryden, I. L., Márkus, L., Taylor, C. C., Kovács, J., 2010. Non-stationary spatio-temporal analysis of karst water levels, *Applied Statistics*, in press.

Füst A., 1997. Geostatisztika, Eötvös Kiadó, 232p.

Füst A., Geiger J., 2010. Monitoringtervezés és -értékelés geostatisztikai módszerekkel I.: Szakértői véleményen alapuló, "igazoló" mintázás geostatisztikai támogatása. Földtani Közlöny140 vol. 3. Sz. pp. 303-312.

Füst A., Geiger J., 2011, Monitoring tervezés és -értékelés geostatisztikai módszerekkel II. Monitoring hálózatok kalibrációja. Kézirat.

Füst A., 2004. Short Course of Geostatistics. Manuscript. Szent István University, Department of Informatics. 56p.

Kovács J., Hatvani I.G., Korponai J., Kovácsné Sz.I., 2010. Morlet wavelet and autocorrelaiton analysis of long term data series of the Kis-Balaton Water Protection System (KBWPS). *Ecol. Eng*., 36, pp. 1469-1477.

Márkus, L., Berke, O., Kovács, J., Urfer, W., 1999. Analysis of spatial structure of latent effects governing hydrogeological phenomena, *Environmetrics* 10 pp. 633-654.

Matheron, G., 1965. Les Variables Regionaliées et leur Estimation. Masson at Cie. Editeurs, Paris, 305p.

Molnár S., és Füst A., 2002. Környezet-informatikai modellek I. Szent István Egyetem, Gépészmérnöki Kar, Informatika Tanszék, Gödöllő, 81p.

Molnár S., Füst A., Szidarovszky F., Molnár M., 2010. Környezetinformatikai modellek.Szent István Egyetem, Gödöllő, 191p.

Samuel S. Wilks, Mathematical Statistics, John Wiley, 1962, Section 8.1

Shannon, C. E., 1998. Communication in the Presence of Noise – Proceedings Of The IEEE, Vol. 86, No. 2, February 1998, pp. 447-457

Slutsky E., 1937. The summation of random causes as the source of cyclic processes. *Econometrica*; 5 pp. 105-46.

Wackernagel, H., 2003. Multivariate Geostatistics, Springer-Verlag, Berlin, 357p.