Analysis of Spatial Structure of Latent Effects Governing Hydrogeological Phenomena [1]

László Márkus, Olaf Berke*, József Kovács** and Wolfgang Urfer*,

Department of Probability Theory and Statistics, Eötvös Loránd University, Budapest

* Department of Statistics, University of Dortmund

** Department of Applied and Environmental Geology, Eötvös Loránd University, Budapest

Key Words: Autoregressive Processes; Dynamic Factor Analysis; Geostatistics; Groundwater;Hydrogeology; Infiltration; Spatially Referenced Time Series; Universal Kriging;Variogram.

1. Latent Effects in a Hydrogeological Problem

In the present study we investigate the data provided by the karstwater level monitoring system, set up in the Transdanubian Mountains, more precisely in the Bakony, the Keszthelyi Mountains and the Balaton-Highland. (Here, like in the sequel, the term karstwater is used for groundwater in karstic areas.) The detailed description of the monitoring system itself and the geological and hydrogeological situation in which the system was planned to function and collect data about the water level, can be found in Márkus *et al.* (1997), as well as the results of our previous study in determinig the underlying (called also latent or background) effects driving the karstwater fluctuations. Some changes in the available data and supplementary informations made it inevitable to recalculate the previously obtained results. At the

[1]This research was supported in part by the Joint Acad. Research Project of DAAD and MÖB, project No.97, the Hung. Nat. Sci. Research Fund OTKA, grant No.: T 016665 and Project for Research Development in Higher Education FKFP grant No.:

same time these recalculated results paved the way to the exploration of the spatial structure of the underlying effects. This purpose required the combined use of dynamic factor analysis and certain kriging procedures.

By the courtesy of the Hungarian Water Resources Research Center and more personally of András Csepregi, we had access to data of some more monitoring wells of the above mentioned system, and the complete geological layer-description registered at the sinking of the wells. On the basis of these new informations some of the wells, considered in the previous study, had to be excluded because they do not measure the water level in the *main* (Triassic) karstic aquifer. So, the final number of monitoring wells, providing data for the present investigation, was 64. As it can be seen on Figure 1, their areal distribution was rather scattered. The water levels in these wells were registered more or less permanently during the period of 1970-90. Observations after 1990, though available, are not taken into consideration in this study, since the social and economic changes taking place that time in Hungary triggered sharp decline in mining and mounting concern towards environmental damages, and that led to significant changes in the human interference into the water resources of the area. The recovery processes of nature are of different character than human interference, so they need distinct study.

The empirical time series of the registered water level - called *hydrograph* - is the most important hydrogeological characteristic of a monitoring well. Therefore, it is plausible to assume that wells with similarly fluctuating hydrographs are in similar hydrogeological situation, consequently, they are under the influence of the same underlying effects. Since our final goal is the analysis of the underlying or latent effects, hydrogaphs have to be the main objects of our study. Unfortunately, the timing of the observations was not systematic. One way to overcome this difficulty is to compute yearly averages from the non time-equidistant raw data, and consider hydrographs of yearly averages. Although it inevitably causes some loss of information, but in the given case this was reasonable mainly because data available for comparison was also of the same structure.

Since we now had access to all the registered data, we could control the averaging. Compensating for the high irregularity in the frequencies of observations it is more adequate to take first monthly averages, and calculate yearly averages only afterwards, from these monthly averages. This method ensures, that the different time intervals (e.g. the seasons) represent more equal weights in the yearly averages.

In spite of a more complete data set, we were confronted missing data problems. As the previous investigations showed Márkus *et al.* (1997), the usual simple methods of adjusting for missing data, such as replacing it from overall means, or interpolating from adjacent points, or predicting values from linear trend regression etc. were not satisfactory. Instead, in the present situation, the high spatial correlations can be used, and the missing values can be replaced with the prediction from (multiple) linear regression based on the observed highly correlating time series of the surrounding wells. The correlations between dependent and independent variables were higher (often much higher) than 0.98 .

STAT.	Correlations, Casewise MD deletion, N=20				
Variable	M4	NV7	NV9	NVAM	NVAM
M4	1,00000	,97726	,93717	,99722	,99161
NV7	,97726	1,00000	,98860	,97739	,99469
NV9	,93717	,98860	1,00000	,94076	,97105
NVAM	,99722	,97739	,94076	1,00000	,99259
NVAM	,99161	,99469	,97105	,99259	1,00000

Figure 1: Correlation among hydrographs

Correlations and close locations were the most important factors in the choice of the independent variables for multiple regression, while simultanously, the similarity of the geological structure had also been carefully checked. As a rule, this latter coincided for close and highly correlating wells, so we used it for discrimination only when the number of suitable candidiates for regression had to be lowered. Figure (1)shows correlations among 5 wells, one of which, NVAM2, has a missing value at the year 1971. On the base of correlations the wells MARKO4, NV7 and NVAM1 can be selected for the multiple regression. Hydrogeological reasoning goes against taking MARKO4 into the independent variable list, so the replacement can be determined by the prediction from the regression with NV7and NVAM1. The regression results can be seen on Figure (2).

STAT.	Regression Summary for Dependent Variable: NVAM2 R= ,99936264 R²= ,99872569 Adjusted R²= ,99861950 F(2,18)=9404,9 p<,00000 Std.Error of estimate: 1,2407					
N=21	BETA	St. Err. of BETA	B	St. Err. of B	t(18)	p-level
Intercept			-8,36204	1,956358	-4,27429	,000263
NV7	,551537	,034178	,62808	,038921	16,13741	,000000
NVAM1	,453564	,034178	,45571	,034339	13,27081	,000000

Figure 2: Results of regression

All in all 21 missing from the 1344 data had to be recovered this way, keeping thus the proportion well below 2% of the overall data number.

What we are interested in is how the fluctuation pattern of an individual hydrograph can be obtained from the linear combination of several *basic* fluctuation patterns, that may correspond to the underlying effects, driving the fluctuation of the water level. Since the karst is mainly a capillary system, the water level in it follows - with some easing - the configuration of the terrain, hence the overall magnitude of the level, - in other words the *mean* of the hydrographs, - does not provide valuable information, and is completely ignorable in the analysis of the latent effects. Any kind of *detrending* the hydrographs can be misleading, because it would partially remove important information about the underlying effects, and the remainder would be meaningless for further analysis. On the other hand, computing the *variance* of time series without removing its trend does not seem to be a meaningful tool to capture information on the underlying effects.

In order to avoid the situation when wells with greater means or variance would predominate, we *standardised the hydrographs* for each and every well to mean zero and variance one. This did not change the *proportion* of weights in the linear combination of basic fluctuation patterns that estimates the hydrograph of the given well.

The next step in the previous study was the grouping of the wells according to fluctuation patterns in the hydrographs. This was motivated by the idea that the latent effects may vary some throughout the area, but wells with similar hydrographs have to be under the influence of the same underlying effects. So, when concentrating to the utmost exactness in the determination of the effects then group by group considerations seem more adequate, whereas a unified approach seems more fruitful in getting a compatible and comparable weight system at every location, even at the price that the obtained factors reflect only the overall tendencies of the effects. It has to be noted, that if the group is too homogeneous, then instead of finding independent fluctuation patterns that reflect the realistic latent effects, one may find certain linear combination of it, characteristic of the whole group. It is well known from ordinary factor analysis that the presence of marker variables is needed for finding a well interpretable factor solution.

It was a priori clear that two main effects driving the fluctuations are the infiltration of the water from the precipitation, and the water extraction of the bauxite and coal mines and the communal water pumps. So, within the limits of the present paper, we intended to determine and identify these effects and describe the spatial structure of the infiltration.

Considering all the water extraction data of the area it appeared that the amount of water, pumped out at Nyirád was three or four times more than all the rest. The remaining extractions were scattered in location and devided in amount all over the territory, thus their effect was negligible. For this reason it proved to be sufficient to use the data of Nyirád for comparison.

Infiltration itself is a very complex process, because the amount of the infiltrating water depends on the geomorphology, the water conductivity of the superficial rocks, the air temperature, the amount, duration and the physical state of the precipitation, and also the vegetation and many other factors. This makes it clear, that the effect of infiltration changes throughout the mountains, hence the importance of creating homogeneous groups. Nevertheless the area is not so large that the main tendencies in the amount of yearly infiltrated water would vary too heavily among locations. This circumstance justified the validity of the above mentioned unified approach in order to obtain compatible weights of the infiltration effect in wells.

After these considerations we are in the position to perform dynamic factor analysis for the correct identification of the underlying effects.

2. Dynamic Factor Analysis

When observations of multiple time series are considered it is often plausible to assume that there are common driving forces behind them, a few latent effects or factors, which determine the behaviour of the individual observations. A common goal in statistics is to identify these latent effects or factor time series. The conventional tool to determine latent variables from observed samples is factor analysis. As Anderson (1963) pointed out, direct application of factor analysis to multidimensional time series often produces unreliable or misleading results, especially, when delayed interdependence occurs among the components. The reason is that conventional factor analysis has been elaborated for independent observations and independence is not the case for time series. This fact requires the elaboration of a new technique, capable to take into account the dynamic structure of the observations. This technique has been named dynamic factor analysis after Geweke (1977).

Since then different approaches became known, several of which are mentioned in Márkus *et al.* (1997).

The idea we use goes back to Bánkövi, Veliczky and Ziermann (1979, 1992), though it can be related up to a certain degree to the work of Box and Tiao (1977) on canonical transformations of time series vectors, too. This factor analysis model requires the factor time series to fit well to an autoregressive model and to minimise a cost function, which is a linear combination of the conditional variance of the prediction error and the state estimation error. So factors should be well predictable and the original time series should be reproduced well from them. This optimisation problem can be solved by an iterative method rather than directly, that gives the advantage of easy programmability (cf. Ziermann and Michaletzky, 1995). In our previous studies (cf. Márkus, Kovács and Csepregi, 1997, Márkus, Berke, Kovács, Urfer 1997) we used an optimisation algorithm for quadratic forms described by Michaletzky, Tusnády, Ziermann and Bolla (1997) and the computer program developed by them on the base of this algorithm.

Let $\mathbf{Y}(\mathrm{t})$ be a sample from the observation of the multidimensional time series or time series vectors.

$$\mathbf{Y}(t) = (\,Y_1(t), \ldots, Y_N(t)\,)' \quad , \qquad 0 \le t \le T \;.$$

Further assume $\mathbf{Y}(\mathrm{t})$ to be weakly stationary apart from a possible linear trend, and satisfying

$$\mathbf{Y}(t) = \mathbf{A} \cdot \boldsymbol{\mathcal{F}} + \boldsymbol{\epsilon}(t) \tag{2.1}$$

with the N $\times$ M matrix $\mathbf{A}$, the time series vector $\boldsymbol{\mathcal{F}}(t)$ of M *independent* time series

$$\boldsymbol{\mathcal{F}} = (\,\mathcal{F}_1(t), , \ldots, \mathcal{F}_M(t)\,)' \quad , \qquad 0 \le t \le T \;,$$

and the Gaussian white noise vector

$$\boldsymbol{\epsilon}(t) = (\,\epsilon_1(t), \ldots, \epsilon_N(t)\,)' \quad , \qquad 0 \le t \le T \;.$$

Now it is aimed to find optimal, in a certain sense, estimators $F_j(t) = \widehat{\mathcal{F}}_j(t)$-s of the *factor time series* $\mathcal{F}_j(t)$. Speaking somewhat heuristically, the estimation of our model should focus on the following three requirements:

(i) The estimators $\mathbf{F}(\mathrm{t}) = [F_j(t)]$ of the factors $\boldsymbol{\mathcal{F}}$ should be a *time-independent* linear transformation of the observation $\mathbf{Y}$.

(ii) The factor time series $\mathcal{F}_j$ should be linearly well predictable from their past.

(iii) Time-independent linear transformation of the factors should provide a "good" estimator $\widehat{\mathbf{Y}}$ of the observations $\mathbf{Y}$.

In order to meet these requirements and give them a more precise meaning

(i) consider only the class of homogeneous linear estimators of $\boldsymbol{\mathcal{F}}$,

$$\widehat{\mathcal{F}}_j(t) = F_j(t) = \sum_{i=1}^{N} b_{j,i} \cdot Y_i(t) \qquad ,$$

(ii) suppose $\boldsymbol{\mathcal{F}}$ to consist of independent *autoregressive* processes $\mathcal{F}_j(t)$ of order L_j with a possible linear trend:

$$\mathcal{F}_j(t) = c_{j,0} + \sum_{k=1}^{L_j} c_{j,k} \cdot \mathcal{F}_j(t-k) + \delta_j(t) \tag{2.2}$$

and with the Gaussian white noises $\boldsymbol{\delta} = (\,\delta_1(t), \ldots, \delta_M(t)\,)'$, independent from each other and from $\boldsymbol{\epsilon}(t)$. Plugging in now the estimated factors into the best predictor of the autoregressions one gets $\widehat{F}_j(t)$ Empirical best the predictor of $F_j(t)$ as

$$\widehat{F}_j(t) = c_{j,0} + \sum_{k=1}^{L_j} c_{j,k} \cdot F_j(t-k) \qquad . \tag{2.3}$$

Though the optimality of this predictor cannot be preserved for the plug in, we will use $\widehat{\mathbf{F}}(t) = \left(\widehat{F}_1(t), \ldots, \widehat{F}_M(t)\right)'$ in 2.3 for the prediction of the estimator $\mathbf{F}(t)$ of the factors $\boldsymbol{\mathcal{F}}$, given the past of the *estimator*. Since the observations and thus the estimations of the factors are known for $0 \le t \le T$, it is possible to compare the predictor with the estimator itself, and by centring, get an unbiased estimator $\widehat{\delta}_j(t)$ of the noise $\delta_j(t)$ as

$$\widehat{\delta}_j(t) = F_j(t) - \widehat{F}_j(t) - \left[\ \overline{F_j - \widehat{F}_j}\ \right] \qquad .$$

(For any $X(t)$ $\overline{X}$ denotes the average $\frac{1}{T+1}\sum_{t=0}^{T} X(t)$.) The squared sum $\mathcal{E}^{(d)}$ or the weighted squared sum $\mathcal{E}_w^{(d)}$ of $\widehat{\delta}_j(t)$ is called the estimated *dynamic error* or estimated *weighted dynamic error*, respectively:

$$\mathcal{E}^{(d)} = \sum_{j=1}^{M}\sum_{t=0}^{T}\widehat{\delta}_j(t)^2 \quad , \quad \mathcal{E}_w^{(d)} = \sum_{j=1}^{M} w_j^{(d)} \sum_{t=0}^{T}\widehat{\delta}_j(t)^2 \qquad .$$

where the constants $w_j^{(d)}$ are the dynamic weights.

As of (iii), from the estimated factors one can produce a linear estimator

$$\widehat{\mathbf{Y}}(t) = \left(\widehat{Y}_1(t), \ldots, \widehat{Y}_N(t)\right)' \ , \quad 0 \le t \le T \ ,$$

$$\widehat{Y}_i(t) = d_{0,i} + \sum_{j=1}^{M} d_{i,j} \cdot F_j(t)$$

estimating the observed time series $\mathbf{Y}(t)$, the realisations will be called the *reestimation of the observation.* The "reestimator" opens the way to estimate $\boldsymbol{\epsilon}(t)$, the noise in 2.1 as

$$\widehat{\epsilon}_i(t) = Y_i(t) - \widehat{Y}_i(t) - \left[\ \overline{Y_i - \widehat{Y}_i}\ \right] \quad ,$$

the squared sum $\mathcal{E}^{(s)}$ or the weighted squared sum $\mathcal{E}_w^{(s)}$ of which is called the estimated *static error* or estimated *weighted static error* respectively:

$$\mathcal{E}^{(s)} = \sum_{i=1}^{N}\sum_{t=0}^{T}\widehat{\epsilon}_i(t)^2 \quad , \quad \mathcal{E}_w^{(s)} = \sum_{i=1}^{N} w_i^{(s)} \cdot \sum_{t=0}^{T}\widehat{\epsilon}_i(t)^2 \ .$$

where the constants $w_i^{(s)}$ are the static weights.

The estimation of the model means to find the matrices

$$\mathbf{B} = [b_{j,i}] \quad , \quad \mathbf{C} = [c_{j,k}] \quad , \quad \mathbf{D} = [d_{i,j}] \ ,$$

$$i = 1, ..., N \quad , \quad j = 1, ..., M \ , \quad k = 1, ..., L = \max_j L_j$$

and the vectors

$$\mathbf{c}_0 = [c_{j,0}], \quad \mathbf{d}_0 = [d_{0,i}],$$

of which $\mathbf{B}$ determines the estimator of the factors, $\mathbf{C}$ and $\mathbf{c}_0$ the predictor of the estimator, $\mathbf{D}$ and $\mathbf{d}_0$ the reestimation of the observation.

To fulfil the requirement given in (iii), the estimation of the model, that is the matrices $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ and the vectors $\mathbf{c}_0$, $\mathbf{d}_0$, is regarded to be "good" if the sum of the estimated (weighted) static and the dynamic errors is minimal. This means the minimisation of the following functional:

$$\Psi_w(t) = \mathcal{E}_w^{(s)} + \mathcal{E}_w^{(d)} = \sum_{i=1}^{N} w_i^{(s)} \cdot \sum_{t=0}^{T} \widehat{\epsilon}_i(t)^2 + \sum_{j=1}^{M} w_j^{(d)} \sum_{t=0}^{T} \widehat{\delta}_j(t)^2 \quad . \tag{2.4}$$

on the constraints

$$var(\mathbf{F}) = \Sigma_F = I_M \quad . \tag{2.5}$$

Let us remark that the ML estimator of the matrices $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ seems too complicated to compute, even though one can determine the density function, but to find its place of global maximum seems too difficult. To overcome these difficulties we used the algorithmic solution, mentioned above.

Rather than find a direct optimal solution of 2.4 and 2.5, what seems to be almost impossible, an iterative approximation is suggested in Bánkövi, Veliczky and Ziermann (1979), a criss-cross algorithm, which was further developed by Michaletzky *et al.* (1997). This means to find for fixed $\mathbf{B}$, $\mathbf{D}$ the optimal $\mathbf{C}$, then for fixed $\mathbf{C}$ determine $\mathbf{B}$, $\mathbf{D}$ and iterating this procedure until the norm of the difference of this matrices will be sufficiently small. A more detailed description of the procedure can be found in Márkus *et al.* (1997). It is known (Bánkövi *et al.* (1992)), that neither steps of this procedure could increase the value of the Ψ functional, but a full analysis of this iteration proved to be too difficult as yet. Practical experiences and simulations indicate however, that the algorithm is sufficiently stable, and under broad conditions should provide a unique solution.

Starting from first order AR processes for factors, almost no change could be experienced when the order was increased. However, we got much better results if we dropped the autoregressive term of the *first factor*, regarding only the 2nd and 3rd factors as AR 1 processes. The reason possibly is the clear downward trend, caused by the water extraction. We applied the above mentioned iterative approximation of the Ψ functional until the change of its value between two consecutive steps of the iteration exceeded 0.001. For the computation we used a program developed by Michaletzky, Tusnády, Zierman and Bolla. Since the hydrographs were standardised, and there was no reason to emphasise the behaviour of this or that monitoring well, or this or that factor, we chose equal unit static and dynamic weights in the static and dynamic error terms, as well. Three effects, that is the corresponding three dynamic factors explain about 84% of the overall variance, although it has to be noted that the generalisation for the case of dynamic factors of this customary characterisation of the result of ordinary factor analysis is not quite straightforward. We were able to identify the first two factors, and it seems for us that the third is describing a small territorial variance in infiltration.

Naturally, as well as on the basis of the previous investigation, the water extraction and the infiltration are expected to be the main underlying effects. In Márkus *et al.* (1997) detailed reasoning is given, here we address only briefly the identification. It is not the aim of the present paper to analyse the first factor but, for the sake of completeness, evidence is given that it represents the water extraction.

STAT.	Correlations of Factor 1 and Wells Influenced by Depression								
Wells	HGN 22	HGN 26	HGN 37	HGN 46	HGN 49	HGN 51	HGN 55	HGN 63	HGN 68
FACTOR1	.984	.995	.988	.983	.978	.994	.977	.981	.991

Figure 3: Correlations of factor 1 and the hydrographs influenced by depression

Very strong correlation can be found between the dominating first factor and the hydrographs of monitoring wells just next to the water extraction site (see Figure (3). These are the wells influenced severely by the depression caused by water extraction. On Figure (4)the graphs of the above mentioned wells and of the first factor also show that factor 1 can be regarded as the effect, representing the water extraction at Nyirád.

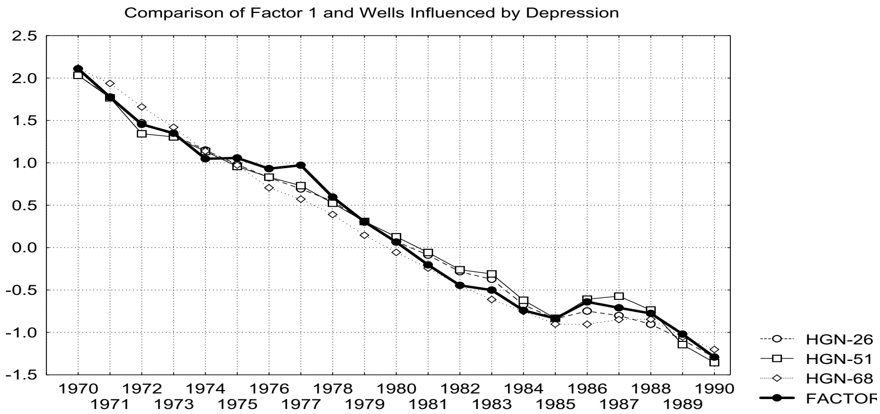

Figure 4: Graphs of factor 1 and the hydrographs influenced by depression

The second dynamic factor represent the overall tendencies of the infiltration throughout the whole investigated area. In order to check it, infiltration has to be computed by conventional methods. Different empirical methods exist for the computation of the infiltration. These methods have it in common that all calculations are based on the precipitation data of meteorological stations, and do not depend on the location, or the local geological structure. For the comparisons we used infiltrations, calculated by four of these methods, called by the names of their creators as Böcker (1974), Kessler (1954), Maucha (1990) and Morton (1983), see also Csepregi (1995). As it was pointed out in Markus *et al.* (1997) the different methods produce pretty different results, correlating only 0.5-0.7 among themselves.

To identify the second factor, it has to be compared with the computed infiltration. In Figure (5) the correlations are given between the second factor and the infiltrations, computed by Kessler's method. Kessler's infiltrations give the highest correlations with the 2nd factor just like in the previous investigation (see Márkus *et al.* (1997)).

STAT. Meteoro- logical	Correlations of Factor 2 and Infiltrations Calculated by Kessler's Method								
Stations	A.FA	BVÁR	BSZ.	HER.	KH.	NYIR	ÚRK	V.PR	ZIRC
FACTOR2	.838	.726	.728	.845	.673	.713	.827	.716	.755

Figure 5: Correlations of factor 2 and Kessler's infiltrations

The corresponding graph (see Figure (6)) fit even better than suggested by the correlations. This is due to the fact that in calculated infiltrations much higher fluctuations can be observed that in empirical ones, this calls for further investigations and possibly corrections in the methods.

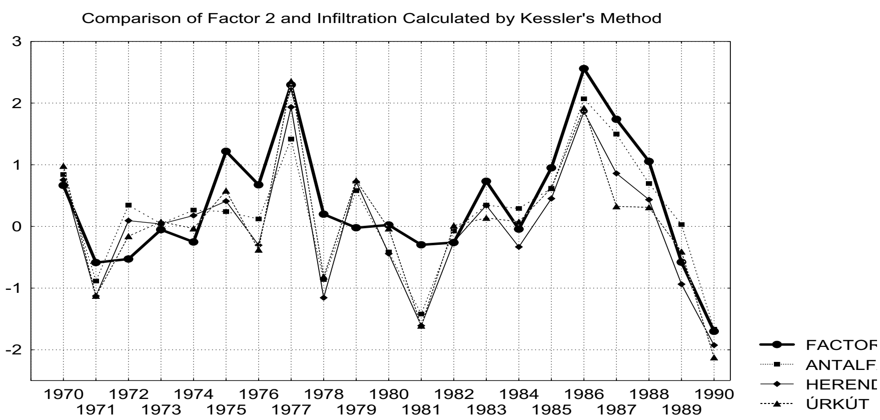

Figure 6: Graphs of factor 2 and Kessler's infiltration

On the Balaton Highland, where the effect of water extraction can be excluded because of the geological structure, undisturbed wells can be observed, the water level of which is influenced exclusively by the infiltration. So, it is a natural idea to compare the hydrographs of these wells straight with the 2nd factor. This can be seen on (8) Figure 5 while the corresponding correlations are given in(7) In Table 4. The good fit of the graphs, and the high correlations strengthen the statement, that factor 2 represent infiltration.

STAT.	Correlations of Factor 2 and Water Levels in Undisturbed Wells						
Variable	B.AKAL	B.FÜR1	B.FÜR2	KÁDÁR.	K.KÁL	PÉCS.	ZÁNKA
FACTOR2_	.8488	.9150	.8947	.9104	.8702	.9189	.9067

Figure 7: Correlations of factor 2 and the hydrographs of undisturbed wells

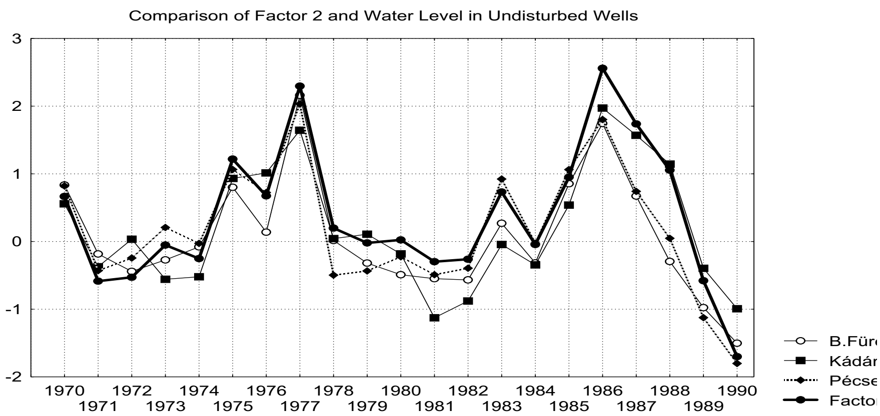

Figure 8: Graphs of factor 2 and hydrographs of undisturbed wells

3. Spatial Modelling for the Weights of the Infiltration

The method of dynamic factor analysis determined the empirical weights of the i-th factor in the reestimation of the j-th observations - these are $d_{i,j}$, the elements of the $\mathbf{D}$ matrix. These weights are straightforward analogues of the factor loadings in ordinary factor analysis, performed according to the covariance matrix, hence $\mathbf{D}$ will also be called the matrix of dynamic factor loadings or sometimes dynamic loadings for short. In the given case dynamic factor analysis was performed not just for unstructured multiple time series but for spatially referenced ones. Therefore a unique location can be associated to each and every dynamic loading; $d_{i,j}$ is naturally associated with the location of the i-th observation. The obtained empirical dynamic loadings are estimations of the theoretical ones, incorporating this way random errors, and being thus random variables. The original observations Y_i represent a finite and thus incomplete sample from an infinite collection of possible sample variables distributing continuously all over the investigation area, that is from the spatial (stochastic) process of karstwater levels. Theoretically, this spatial process can be decomposed into the linear combinations of *the same* factors at every location, hence the dynamic loadings have obvious meaning at intermediate points as well , that is at locations other than observation sites. Generally speaking, the factors may also have spatial structure, but the considered phenomenon can realisticly be modelled by fixed factor time series. To account for the spatial variability of the factors themselves seem a difficult but interesting problem that we intend to investigate in the sequel. These considerations make sense to study the spatial process of the empirical dynamic factor loadings. To stress spatial dependence introduce the notation

$$\mathcal{Z}_j = \left\{ Z_j(s) : s \in \mathcal{D} \subset I\!R^2 \right\}.$$

for the spatial process of the empirical dynamic loadings of the j-th factor. This means that the $d_{i,j}$ coefficients give the observed value of the variables $Z_j(s_i)$. This way data have been obtained at N locations, $s_1, ..., s_N$ and $Z_j(s_1), ..., Z_j(s_n)$ represent a finite sample from a continuously parametrised spatial stochastic process all over $\mathcal{D}$. This layout is often referred to as the statistical analysis of spatial data called geostatistics, see Cressie (1993, p. 10) for more details. In the present work we shall investigate the spatial structure of the dynamic loadings of the second factor - the one corresponding to infiltration in the application - only, and this fact enables us to omit in what follows the lower index of the spatial processes Z_j, by putting $Z(s) = Z_2(s)$.

In general the random variables of spatial processes possess some kind of spatial dependence. The spatial dependence of any two variables $Z(s)$ and $Z(s+h)$ is structured according to the shift h between the corresponding locations s and $s+h \in \mathcal{D}$. The shift represents both distance and direction. For so called isotropic spatial processes the dependence structure is direction invariant.

The objectives of the spatial analysis within this context of geostatistics are the following.

(i) Spatial modeling: find an appropriate model for the phenomenon under study.

(ii) Parameter estimation: estimate the parameters of the model which define the distribution of the spatial process.

(iii) Kriging: predict a realization of the interesting process for any location in the investigation area where data are needed or for locations on a certain grid for mapping purposes.

The modeling approach presumes (often unverifiably) that the spatial process $Z(s)$, $s \in \mathcal{D}$, can be decomposed into two components which account additively for different scales of variation. The first component $\mu(s)$, $s \in \mathcal{D}$, is the large scale variation component describing the mean or trend of the spatial process. The second component $\delta(s)$, $s \in \mathcal{D}$ is the correlated error process accounting for *all* the random variation (often classified as the smooth small-scale, microscale and error variations, cf. Cressie (1993, p. 113)) of the process. With this the process is represented by

$$Z(s) = \mu(s) + \delta(s), \qquad s \in \mathcal{D}.$$

By use of the matrix notation for the sample vector $\mathbf{Z} = (\, Z(s_1), ..., Z(s_n)\,)'$ with mean vector $\boldsymbol{\mu} = (\, \mu(s_1), ..., \mu(s_n)\,)'$ and spatially correlated random disturbances vector $\boldsymbol{\delta} = (\, \delta(s_1), ..., \delta(s_n)\,)'$ the sample can be represented by the general or spatial linear model

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\delta}, \quad E(\,\boldsymbol{\delta}\,) = \mathbf{0}, \quad Cov(\,\boldsymbol{\delta}\,) = \boldsymbol{\Sigma}.$$

The existence of the moments is supposed. To complete the spatial linear model the mean and covariances can be structured according to some parametric functions.

The large scale variation is expressed as a linear combination of k regressors $\mathbf{f}(s)'$ with the coefficient vector $\boldsymbol{\beta}$:

$$\mu(s) = \mathbf{f}(s)'\boldsymbol{\beta},$$

where $\mathbf{f}(s) = (\, f_1(s), ... f_k(s)\,)'$ is composed by functions which correspond to explanatory variables such as location coordinates, distances to certain locations, grouping variables, and so forth (cf. Cressie, 1993, p. 151). The coefficient vector $\boldsymbol{\beta}$ is unknown and has to be estimated on the basis of the observed values $(\, Z(s_1), ..., Z(s_n)\,)'$ of the spatial process $\mathbf{Z}$. In modelling the spatial trend of the dynamic loadings, polynomials in the location coordinates will be used for regressors. Define $\mathbf{F} := (\, \mathbf{f}(s_1)', ..., \mathbf{f}(s_n)'\,)'$ the regressor matrix for the sample variables, and $\mathbf{f} = \mathbf{f}(s_0)$ the vector of regressors for a random variable at location $s_0 \in \mathcal{D}$. Hence, the expectation of the sample variables and the variable $Z(s_0)$ are given by $E(\,\mathbf{Z}\,) = \mathbf{F}\boldsymbol{\beta}$ and $E(\, Z(s_0)\,) = \mathbf{f}'\boldsymbol{\beta}$, respectively.

To represent the spatial dependence structure of the dynamic loading process the semivariogram

$$\gamma(h, \boldsymbol{\theta}) := \frac{1}{2} Var\Big(Z(s) - Z(s+h)\Big)$$

is used - as it is usual in geostatistics - instead of the autocovariance function or covariogram

$$\sigma(h, \boldsymbol{\theta}) := Cov\Big(Z(s),\; Z(s+h)\Big).$$

For ergodic processes these two functions are connected by

$$\gamma(h, \boldsymbol{\theta}) = \sigma(0, \boldsymbol{\theta}) - \sigma(h, \boldsymbol{\theta}).$$

The dependence of the covariance matrix of the sample vector $\mathbf{Z}$ on $\boldsymbol{\theta}$ is sometimes expressed by the notation $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma} = \left[Cov(\, Z(s_i),\, Z(s_j)\,)\right]_{i,j=1}^{n}$. Further, the vector of covariance's between $Z(s_0)$ and the sample vector $\mathbf{Z}$ is denoted by $\boldsymbol{\sigma}(\boldsymbol{\theta}) = \boldsymbol{\sigma} = \left[Cov(\, Z(s_0),\, Z(s_i)\,)\right]_{i=1}^{n}$. Valid parametric models for the semivariogram and the covariogram are presented in Cressie (1993, pp. 61-68). There, an interpretation of the components of the vector $\boldsymbol{\theta}$ is also presented. In the Gaussian case the model parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ define the first and second order moments and thus the whole process $\mathcal{Z}$.

After the model identification, i.e. choice of the parametric model, the unknown model parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ must be estimated. The estimation of the spatial mean and prediction of the spatial process at any location is possible.

The various known variogram estimators cannot be used directly for spatial prediction (kriging) what we intend to carry through, because they are not conditionally negative definite, which is the well-known property of theoretical variograms. The idea then is to search for a valid variogram that, as the measure of the spatial dependence structure, is closest to the estimated spatial dependence in the dynamic loading process. The space of all valid variogams is a large set over which to search, so a parametric family of variograms

$$\boldsymbol{\mathcal{V}} = \{2\gamma : 2\gamma(\cdot) = 2\gamma(\cdot, \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta)$$

(e.g. linear, exponential, spherical etc.) is chosen. Several goodness-of-fit criteria for finding the best element of $\boldsymbol{\mathcal{V}}$ can be proposed. Cressie (1993, p. 91) notes that the various least squares criteria require the fewest assumption on the distribution of $\mathbf{Z}$. In the given case of dynamic loadings, not only the exact but even the asymptotic distribution of the estimators is too difficult to calculate, and only simulations seem to be perspectivical in obtaining information on the distribution of $\mathbf{Z}$. Thus the distribution can easily be misspecified, calling for special care to be taken for the robustness of the used methods. Least squares fittings in general possess these properties (see e.g. Carroll and Ruppert 1982). Furthermore, in view of the bias of variogram estimations when spatial trend is present in the model, we decided to settle at the weighted least squares method of variogram fitting as described in Cressie (1993, p. 95).

The estimation of the spatial mean is known as trend surface estimation. Assuming knowledge about the structure parameter $\boldsymbol{\theta}$ the trend surface estimator or best linear unbiased estimator (BLUE) is given by

$$\widehat{\mu}(s_0) = \mathbf{f}'\widehat{\boldsymbol{\beta}}, \qquad s_0 \in \mathcal{D}.$$

where for $\widehat{\boldsymbol{\beta}}$ seems to be reasonable to use the generalised least squares estimate of $\boldsymbol{\beta}$ plugging in $\widehat{\boldsymbol{\theta}}_{WLS}$ instead of the true parameter $\boldsymbol{\theta}$. Thus the "WLS" estimate for the trend parameter $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}}_{WLS} = \left(\mathbf{F}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_{WLS})^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}_{WLS})^{-1}\mathbf{Z}. \tag{3.2}$$

Evaluating the trend surface estimator by use of the *estimated* structure parameter $\widehat{\boldsymbol{\theta}}$, the empirical best linear unbiased estimator (EBLUE) is obtained.

The prediction of spatial processes is known as kriging. Adopting the parametric trend model $\mu(s) = \mathbf{f}(s)'\boldsymbol{\beta}$, $s \in \mathcal{D}$, the geostatistical method of best linear unbiased prediction (BLUP) is called universal kriging. The universal kriging predictor is given by

$$\widehat{Z}(s_0) = \mathbf{f}'\widehat{\boldsymbol{\beta}} + \boldsymbol{\sigma}'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{F}\widehat{\boldsymbol{\beta}}), \qquad s_0 \in \mathcal{D}.$$

Similar to the EBLUE, plugging in the estimated parameters for the true but unknown structure parameter results in empirical best linear unbiased prediction (EBLUP).

Let us note, that optimality properties of the BLUE and BLUP may not be shared by the EBLUE and EBLUP, respectively. The uncertainty of the EBLUE and EBLUP is measured by the corresponding mean squared error (MSE) and mean squared prediction error (MSPE). Approximations for these quantities do not exist in general. A remarkable result, however, is known as the underestimation problem. This result states that using the estimate $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ gives the empirical MSE and empirical MSPE of the EBLUE and the EBLUP, which have the tendency to underestimate (in mean) the true MSE of the BLUE and true MSPE of the BLUP. And by the optimality of the BLUE and BLUP, the corresponding mean squared error and mean squared prediction error can not be larger than the one of the EBLUE and EBLUP, respectively (cf. Cressie, 1993, pp. 296-299).

It is clear, that on the one hand the structure parameter $\boldsymbol{\theta}$ cannot be estimated independently of the trend parameter, on the other hand (3.2) shows that the estimation of the trend parameter $\boldsymbol{\beta}$ requires knowledge of the structure parameter. A possible procedure to overcome this problem is to start with the ordinary least squares estimate of $\boldsymbol{\beta}$, compute the structure parameter estimation from the residuals, fit a variogram model, obtain a generalised least squares estimator of $\boldsymbol{\beta}$ based on the fitted model, and so forth, the procedure can be itterated. However, it can easily be illustrated on certain examples (Cressie, 1993, p. 166.) that residuals based on the most efficient estimator of the trend parameter yield a biased estimator of the variogram. It is generally true that the bias of a residual based variogram estimator is small at lags near the origin but more substantial at distant lags. Now, since our variogram model is fitted by weighted least squares, which puts most weight on the estimator at small lags, the effect of the bias should be small. Moreover, because kriging is carried out in local neighbourhoods, the fitted variogram is only evaluated at smaller lags, precisely where it has been well fitted. The bias is more likely to affect the estimated kriging variance, which may be smaller than it should be. The exactness of the universal kriging predictor as well as the estimated kriging variance may be far more influenced by the distribution of $\boldsymbol{\delta}$ in the model if that is mistakenly specified as Gaussian. So in the end of the itterative calculation of the kriging predictor we carefuly payed attention to the distribution of the residuals.

The actual calculations of the kriging predictions were carried out using the SpatialStats module of the S-Plus program package. Ordinary kriging immediately proved to be inadequate for the spatial prediction of the dynamic loadings of the second factor. It is further emphasized by the highly non-normal distribution of the dynamic loadings. Afterwards, universal kriging has been performed with linear, quadratic, and cubic polynomials as regressors of the hypothesised trend surface. The spherical family of variograms sems appropriate for fitting to the empirical variograms of the residuals. Compared to the linear case the fitted variogram for the residuals of cubic trend removal does not show great improvements, although the crucial fit around the origin improves somewhat (compare Figures (9), (10)). The sill value decreases to one third when universal kriging is performed instead of ordinary one, however, it does not lowers significantly further for the cubic case compared to the linear trend removal. On the other hand the number of estimated parameters increases - only x^3 has insignificant coefficients, thus only this term can be dropped. What still may argue for the cubic trend removal is the distribution of the residuals that gets much closer to normal (see Figure (14)), although in terms of skewness and curtosis it is still not sufficiently good (cf. Figure (15)). The obtained kriging predictions (see Figures 11, 12 are not very different, and reflect well the geological - sedimentological and tectonic - structure, however the detailed analysis from these point of view exeeds the framework of the present paper. The map of the MSPE of the cubic case is given on Figure (13) There is hope to get more information on the distribution of the dynamic loadings by simulations and then trans-Gaussian kriging may be used.

Acknowledgements

The authors wish to thank György Michaletzky for providing the elaborated program package and for the helpful discussions on the theory of dynamic factor analysis. We should also like to thank András Csepregi for providing valuable hydrogeological

informations.

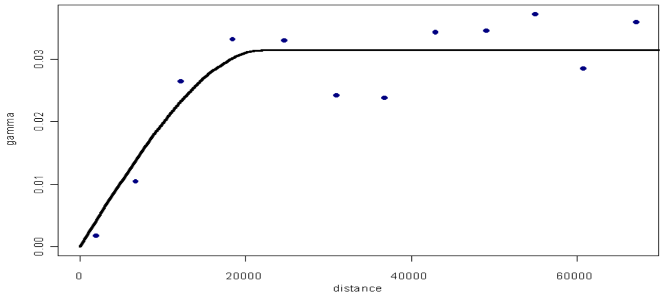

Figure 9: Variogram fitted by WLS to the residuals from *linear* trend fitting

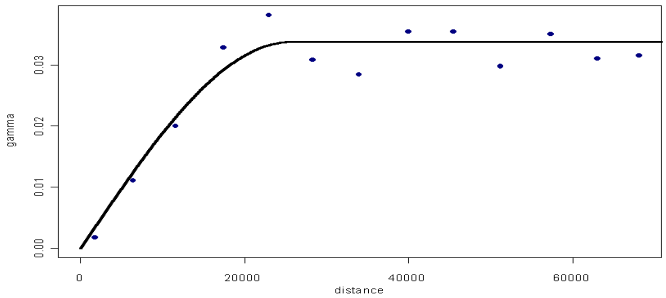

Figure 10: Variogram fitted by WLS to the residuals from *cubic trend* fitting

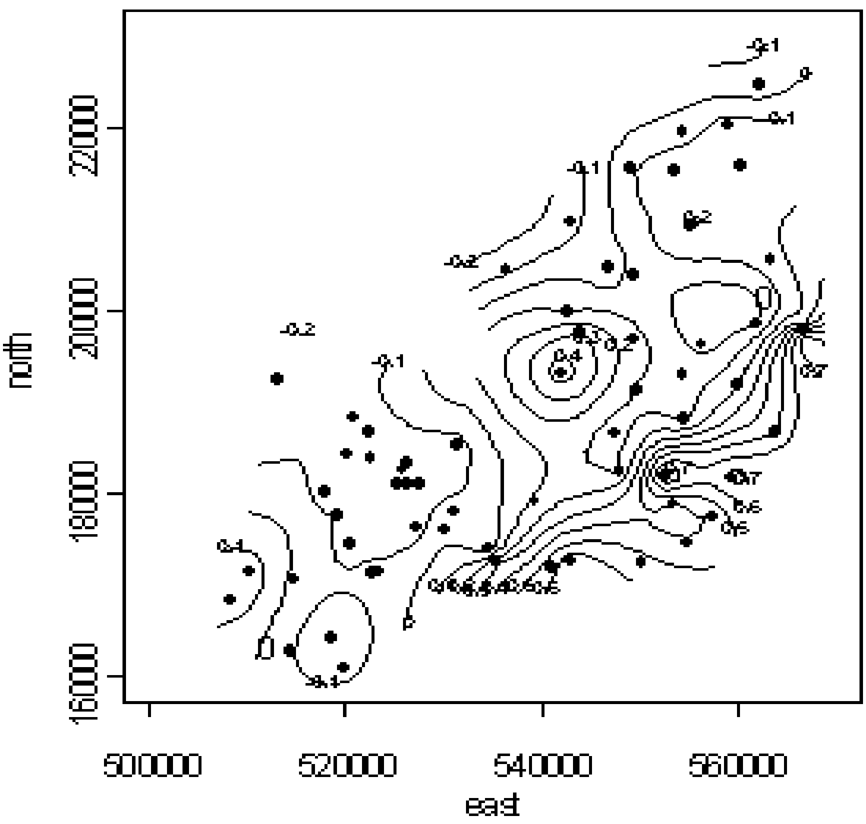

Figure 11: Universal kriging prediction of dynamic loadings with *linear* trend fitting

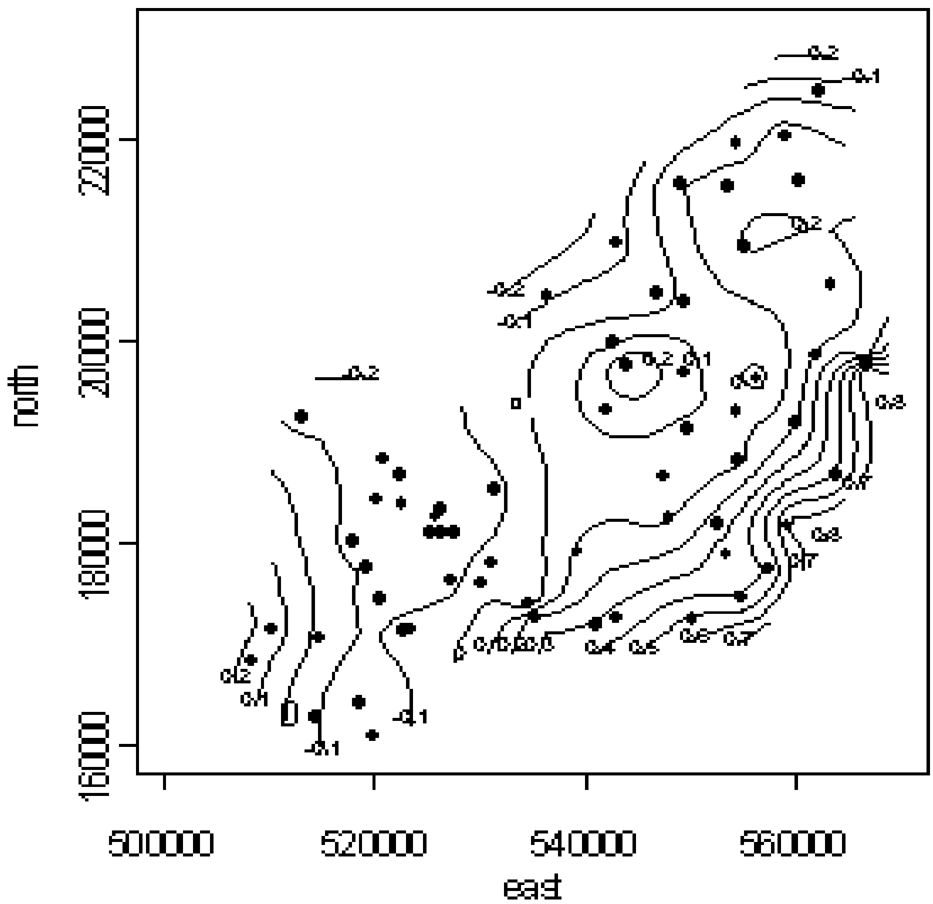

Figure 12: Universal kriging prediction of dynamic loadings with *cubic* trend fitting

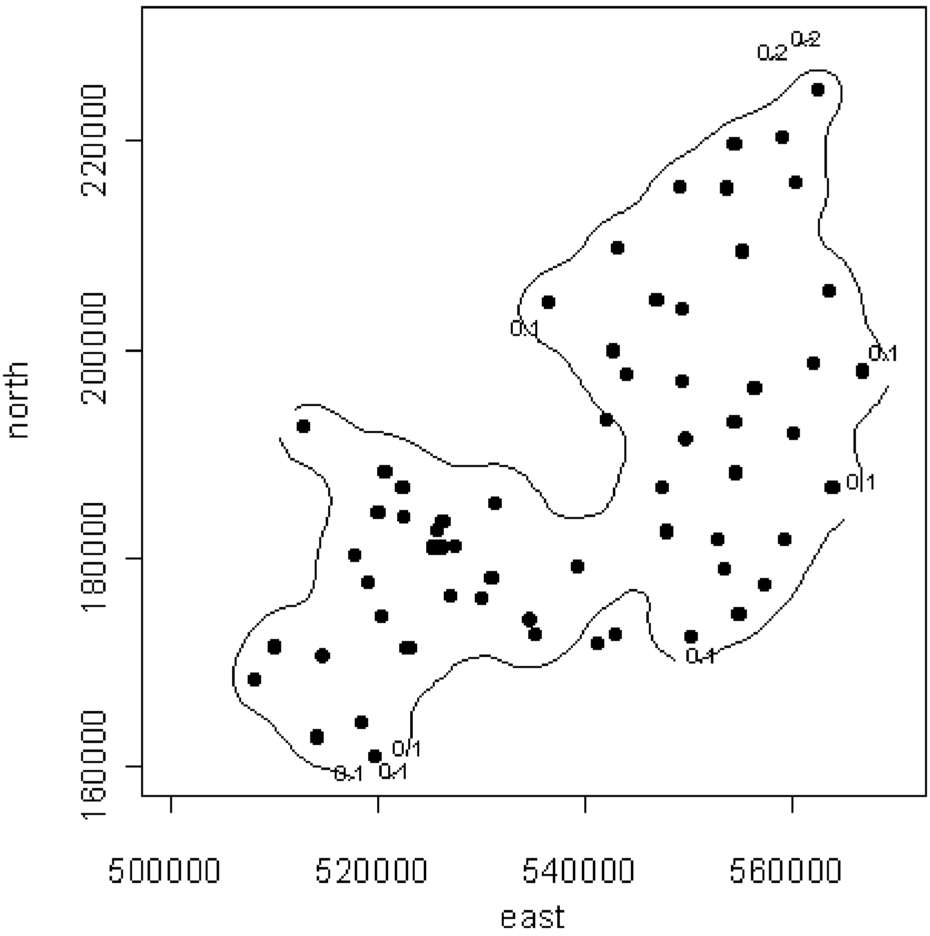

Figure 13: Standard error of universal kriging prediction of dynamic loadings with *cubic* trend fitting

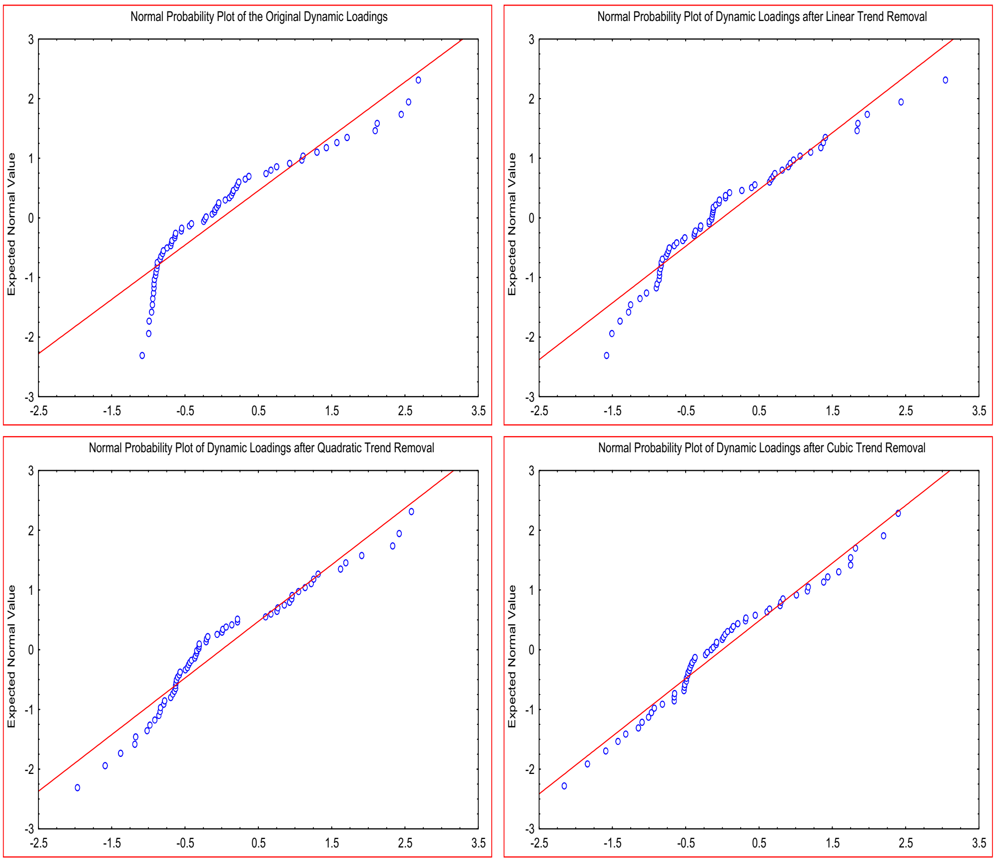

Figure 14: Normal plot of observations and residuals of linear, quadratic and cubic trend fitting

STAT. Variable	Descriptive Statistics						
	V. N	Mean	Min	Max	Std. Dev.	Skew ness	Kurt osis
ORIGINAL	64	0.00	-1.1	2.68	1.00	1.16	.54
LIN. TREND RES.	64	0.00	-1.6	3.05	1.00	.87	.48
QUAD.TREND RES.	64	0.00	-2.0	2.59	1.00	.78	.14
CUB. TREND RES.	64	0.00	-2.1	2.58	1.00	.76	.59

Figure 15: Skewness and kurtosis of standardised residuals

References

Anderson, R. L. (1963) The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika*, 28, 1-25.

Bánkövi, Gy., Veliczky, J. and Ziermann, M. (1979) Dynamic models for prediction of the development of national economies. In: Janssen, J. M. L., Pau, L. F. and Straszak, A. (eds) *Models and Decision Making in National Economies.* North-Holland, Amsterdam. 257-267.

Bánkövi, Gy., Veliczky, J. and Ziermann, M. (1992) *Dynamic Factor Analysis* (in Hungarian). Karl Marx University of Economics, Budapest. 1-81.

Böcker, T. (1974) Determination of infiltration in karst-terrain by the method of quarterly treshold precipitations (in Hungarian). *Report of the Hungarian Water Researches Institute.*

Box, G. E. P. and Tiao, G. C. (1977) A canonical analysis of multiple time series. *Biometrika*, 64, 355-405.

Cressie, N. (1993) *Statistics for Spatial Data*, Revised Edition. Wiley, New York.

Csepregi, A. (1995) Investigation of groundwater recharge on the Hungarian Plain by precipitation and areal evapotranspiration data (in Hungarian). *Research Report of the Hungarian Water Researches Institute.*

Geweke, J. F. (1977) The Dynamic factor analysis of economic time series models. In: Aigner, D. J. and Goldberger, A. S. (eds) *Latent Variables in Socio-economic Models.* North-Holland, Amsterdam. 365-382.

Kessler, H. (1954) The percentage of infiltration and the permanently extractable water in karst-terrains (in Hungarian). *Vízügyi Közlemények (Journal of Hydraulic Engineering)*, 36.

Márkus, L., Berke, O.Urfer, W. and Kovács, J. (1997) Investigation of Latent Dynamic Characteristics of Space-Time Phenomena *Forschungsbericht 97/5.* Dortmund University press pp. 1-39.

Márkus, L., Kovács, J. and Csepregi, A. (1997) Identification of underlying effects driving multiple time series; an application to karstwater level fluctuations. *Journal of Environmental and Ecological Statistics.* (Submitted for publication).

Maucha, L. (1990) Calculation of karstic infiltration (in Hungarian). *Hidrogeológiai Közlöny (Journal of the Hungarian Hydrogeological Society)*, 70, 153-161.

Michaletzky, Gy., Tusnády, G., Ziermann, M. and Bolla, M. (1997) Extrema of sums of heterogeneous quadratic forms. *Journal of Linear Algebra and Its Applications.* .

Morton, F. I. (1983) Operational estimates of areal evapotranspiration and their significance to the science and practice. *Journal of Hydrology*, 66, 1-76.

Ziermann, M. and Michaletzky, Gy. (1995) Factor analysis of time series (in Hungarian). *Szigma, Journal of the Society for Economic Modeling*, 26, 77- 91.