Jul 1st, 12:00 AM

Virtual Sources for Spatio-temporal Monitoring Data Analysis

A. Nuzhny

E. Saveleva

S. Kazakov

S. Utkin

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Virtual Sources for Spatio-temporal Monitoring Data Analysis

A. Nuzhny, E. Saveleva, S. Kazakov, S. Utkin

Nuclear Safety Institute Russian Academy of Sciences, 52 B. Tulskaya, Moscow, 113191, Russia (nuzhny@inbox.ru, esav/kazak/uss@ibrae.ac.ru)

Abstract: The current work deals with application of independent virtual sources as a tool for analysis and storing groundwater monitoring data. Estimation of virtual sources is based on learning from data principle. Virtual sources allow to analyse the main features of the process. They can be useful for both data exploratory analysis and prediction. This work considers two methods for independent virtual sources construction: (1) based on using Central Limit Theorem and (2) based on the mixture of Gaussians. The methods were applied to real spatio-temporal data on groundwater level dynamics (2D spatial case) and groundwater contamination by radioactive nitrates (3D spatial case). These data sets are characterized by heterogeneous sample distribution both in space and time.

Keywords: Data mining, Independent sources, Gaussian distribution, Spatio-temporal data, Monitoring.

1. INTRODUCTION

Real monitoring systems usually deal with a problem of data analysis in spatio-temporal continuum One of possible ways is a data mining methodology featuring adaptive learning from data principle [Cherkassky and Muller, 1998]. Data mining provides a set of methods to handle, to model and to make predictions basing on measured data and additional knowledge somehow connected with data (physical laws, empirical rules, etc.). Diversity of data mining methods is very high covering all kinds of data analysis problems: regression, classification, clustering, etc.

The current work considers a data mining approach based on independent sources (independent components) [Lee T.-W.1998]. Generally independent component analysis solves the blind source separation problem by decorrelating the signals and reducing some higher-order statistical dependencies. Independent sources store enough information on the process for modelling and making predictions. Thus independent sources also can be useful as a compressing tool to store the monitoring data more effectively. Independent sources can be considered as one of possible non-linear generalizations of principle component analysis.

Classical independent sources problem is an inverse problem. A measured value (a receiver) is treated as a superposition of signals from some sources. The problem is to establish sources. As any inverse problem this one is an ill-posed problem with unlimited number of possible solutions. In this work independent sources are considered as virtual (not real) sources responsible for a process, thus any reliable solution is equally good.

This work discusses two approaches to find the independent virtual sources. Both of them deal with Gaussianity: (1) consider superposition more normally distributed in comparison with single signals (based on using Central Limit Theorem) and (2) consider the receiver

as a weighted superposition of Gaussians. Both methods have some theoretical limitations complicating their application to real data.

Methods were applied to analysis of real data from a groundwater monitoring system studying the vicinity of nuclear waste facilities.

2. THEORETICAL BACKGROUND

2.1 Method using the Central Limit Theorem

To formulate an independent sources problem let us consider values measured at spatio-temporal points x_i as receivers obtaining signals $c_i=c(x_i)$, which are a superposition of initial signals ($b_j=b(r_j)$) from some sources r_j. The task is to restore initial signals and their sources. Otherwise it is to find the coefficients w_{ij} in:

$$b_j = \sum_i w_{ij} c_i \quad (1)$$

under the additional constraint:

$$\sum_j w_{ij}^{\,2} = 1 \quad (2)$$

Such formulation of the independent sources problem requires the following assumptions: the sources are points, no spatial size; influence of sources on receivers is immediate; distribution of signals from sources is isotropic. The possibility to work under these assumptions should be considered before application of the method.

One of the possible ways to solve this problem is to use the Central Limit Theorem stating that the superposition of sources presents more normal distribution then single signals. The level of normality can be characterized by the excess kurtosis (the function of the fourth standardized moment) [Joanes and Gill, 1998]. Kurtosis equals to zero for normally distributed data, the positive kurtosis indicates more narrow probability density function in comparison with the normal one, and the negative kurtosis is due to wider probability density function then the normal one. Kurtosis (K) can be written in terms of matrix $\mathbf{W}=(w_{ij})$ from (1) as

$$K = \frac{\frac{1}{n}\sum_j \left(\sum_i w_{ij} c_i \right)^4}{\left(\frac{1}{n}\sum_j \left(\sum_i w_{ij} c_i \right)^2 \right)^2} - 3 \quad (3),$$

where n is a number of sources. Thus the problem is stated as the minimization of squared kurtosis (3) under the constraint (2).

Numerical optimisation under a constraint uses a penalty function. This function has zero value within the allowed range of values and it tends to infinity in the forbidden area. The penalty function for maximization under the constraint (2) looks as

$$\left(\sum_j w_{ij}^{\,2} - 1 \right)^2 \quad (4).$$

The final expression for optimization is thus:

$$\Phi = K^2 - \frac{1}{\mu}\left(\sum_j w_{ij}^{\,2} - 1 \right)^2 \quad (5),$$

with parameter μ monotonically decreasing during the optimization.

A number of parameters of the optimisation problem equals to a number of elements in matrix **W,** which strongly depends on the dimension of the initial data set. Together with the number of parameters grows the number of local optimisation solutions. The number of functional parameters of the optimisation problem can be reduced by a preliminary compressing data transformation (**AC**), for example, by transformation to principle components. The final transformation of data to independent sources in this case is performed by matrix

$$\boldsymbol{B} = \boldsymbol{WA} \quad (6)$$

Coefficients of matrix **B** describe the credit-debit dependencies between sources and measurement wells. A positive coefficient indicates a positive dependence: the increase of a source value causes an income of a value at the receiver. A negative coefficient is when a source value is increasing on account of the receiver. Thus matrix **B** is an important tool for analysis of the influence between sources and monitoring wells, especially in the case when sources can be associated with some facilities (for example, lakes or rivers in the case of water monitoring).

The matrix **B** also allows to estimate the locations of independent sources [Üzümcül et al., 2003]:

$$\boldsymbol{r}_j = \sum_i \boldsymbol{b}_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_0) + \boldsymbol{x}_0 \quad (7),$$

here x_0 is a centre of mass over all measurement locations.

2.2 Gaussian mixture approach

The process under study can be approximated as a weighted superposition of Gaussian-like signals:

$$\boldsymbol{c}(\bar{\boldsymbol{r}},\boldsymbol{t}) = \sum_{t-1}^{n} \boldsymbol{A}_i \, exp\left(-\frac{\alpha_i}{\boldsymbol{t}-\boldsymbol{t}_i} \left\| \bar{\boldsymbol{r}} - \bar{\boldsymbol{r}}_i \right\| \right) \quad (8),$$

where $\bar{r}$ indicates a spatial coordinate, t is a temporal coordinate, n is a number of sources (user-defined parameter), A_i, α_i, $\bar{r}_i$ and t_i, ($i=1,\ldots,n$) are fitting parameters. Parameters $\bar{r}_i$ and t_i indicate a spatio-temporal coordinates of sources.

The model fitting procedure leads to an optimization problem. The model parameters (weights) are estimated so as to minimize the sum of squared residuals at the sample locations:

$$\boldsymbol{D} = \sum_k \left(\boldsymbol{c}(\bar{\boldsymbol{r}}_k, \boldsymbol{t}_k) - \boldsymbol{y}_k \right)^2 = \sum_k \left(\sum_i \boldsymbol{A}_i \, exp\left(-\frac{\alpha_i}{\boldsymbol{t}-\boldsymbol{t}_i} \left\| \bar{\boldsymbol{r}}_k - \bar{\boldsymbol{r}}_i \right\| \right) - \boldsymbol{y}_k \right)^2 \quad (9)$$

here y_k are known values, $k=1,\ldots,K$ with K being a number of samples. Minimization was performed by RPROP [Riedmiller and Braun, 1993] method. RPROP is an effective and stable iterative optimization algorithm. The direction of a parameter's changes depends only on the sign of the objective function's derivative, and the learning rate depends on the previous state of the system. Such construction of iterations sometimes allows to jump over local extremes and significantly decreases the time of the approximation in the monotonic case.

The function (8) with fitted parameters allows to make predictions at any spatio-temporal locations. Any kinds of dependencies between sample locations and sources are also available. The main drawback of this approach is the assumption made on the form of

signals emitted by virtual sources. From this point of view the first approach is more general, as it does not require any assumption on the form of signals.

3. DATA DESCRIPTION

Groundwater monitoring in the region connected with radioactive storing facilities controls groundwater level dynamics and water quality. We consider two case studies describing two different events: a groundwater level dynamics and groundwater contamination by radioactive nitrate. The groundwater level is a unique value per a well at a time, thus it presents a two-dimensional in space problem. Nitrate contents also depend on the depth of a sample and consequently lead to a three-dimensional in space problem. The main problem is to extract the principle data features allowing their compact description.

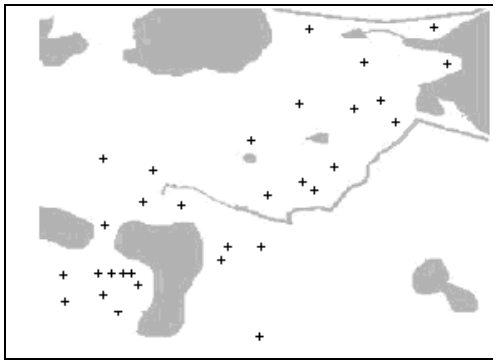

Figure 1. Spatial distribution of hydro-geologic wells

For our analysis we used 32 geological monitoring wells. Figure 1 presents spatial distribution of these wells together with the main hydrologic objects of the region. Each well contained at least one groundwater level measurement per month during period from April 1970 till January 2006. Total number of groundwater samples was 15673. In [Nuzhny et al, 2007] a linear feature extraction approach was discussed (principle component analysis). First principle component responsible for the averaging over the well gives satisfactorily good approximation for 32 spatially distributed wells. Other components did not significantly influence on that approximation. Thus, features with physical meaning were decided to be searched through non-linear description. Several not strong assumptions led us to a classical independent sources problem.

Nitrate samples were even sparser - an average period between two measurements in the same position is a year during the period with the highest sampling density (1964 - 2003 years). The total number of nitrate samples was 1576. Taking into the account the curse of dimensionality it is obvious that traditional data analysis fail for these data. The most reasonable approach in such case is a parametric one. Gaussian-like approximating functions are the most common for a parametric description.

4. RESULTS AND DISCUSSION

4.1 Virtual sources for groundwater levels

A temporal cut for a groundwater level can be approximated with 98% confidence using 10 principle components [Nuzhny et al, 2007]. Correlation of the first principle component with precipitation data shows the direct effect - instantaneous impact on the precipitation (maximum of correlation at zero). Other principle components present different structure of correlation, thus they are after other physical phenomena. So estimation of independent sources for groundwater level was performed by minimization of kurtosis (3) using 9 principle components. The purpose was to delete direct influence of precipitation and to focus on analysis of other dependencies.

As for theoretical assumptions of the method the rudest among them is the assumption of isotropy, as anisotropy is a usual case in hydrogeology. But still such assumption is often made for preliminary analysis and it does not seriously distort results. The immediate reply of the receiver on the source is fulfilled as we have a monthly step in our data, so instantaneous effect means the effect within the month.

Figure 2a shows spatial distribution of 9 independent sources, they are marked by pluses. All of them can be associated with surface water reservoirs (lakes and a river). Figure 2b shows histograms of lines from matrix **B** (6) corresponding to the wells marked in fig.2a as filled circles. Analysis of matrix **B** allows to get some information on possible influence of contaminated artificial reservoirs on the water in wells. In terms of groundwater levels problem matrix coefficients obtain real meaning: positive coefficient indicates that growing of water level in the source causes an income of water to the well; negative coefficient means that the source takes water from the well.

For example, the well number 18 gets impact from sources 2,3 and 7 associated with the same lake and the source 9 associated with the river. Sources 1 and 5 (both associated with contaminated lakes) take water from this well. So, we can conclude that water from contaminated lakes do not get to this well. But still sources 2,3 and 7 can be influenced by contamination via well 7 taking water from the source 5. Such results are the subject for further analysis by experts and decision makers working at the site.

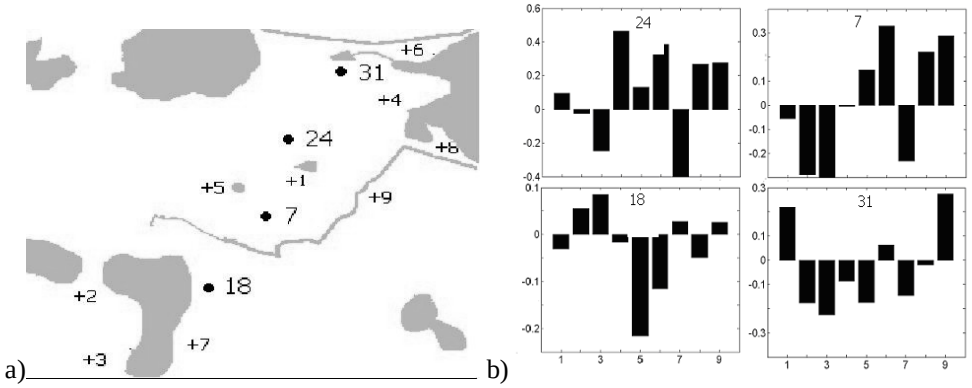

Figure 2. Results on the independent sources analysis: a) Schema of water reservoirs, 4 wells (filled circles) and 9 virtual sources (pluses); b) Histograms of lines of matrix **B** corresponding to 4 wells

So, we estimated virtual sources responsible for groundwater level dynamics. These sources appeared to be useful tools for studying influences between different water subsystems.

4.2 Virtual sources for nitrate contamination

Virtual sources for groundwater contamination by nitrates were estimated by Gaussian mixture approach (8). As initial values of parameters A_i and α_i we took small values (0.001), initial centres (sources) $\vec{r}_i$ were randomly distributed over the region. Examples of initial distributions of 8 and 4 centres are presented in figure 3 (rhombus), filled circles indicate spatial distribution of sample wells. Stability of RPROP algorithm guarantees the stability of the final result. Several initial random distributions of sources were used, and they led to very close results.

We started with 8 sources but the obtained result indicated surplus of such model - several sources united into one (fig. 4a). So finally the model with 4 sources (20 fitting parameters) was used (fig. 4b). The quality of the model can be checked by the relative error:

$$\boldsymbol{R} = \frac{\boldsymbol{D}}{\sum_{\boldsymbol{k}} \boldsymbol{y}_{\boldsymbol{k}}^2} \quad (10)$$

where ***D*** is defined by (9). Final relative error for model with 8 sources is ≈ 0.13, and with 4 sources - ≈ 0.2. Twice decreasing the complexity of the model we nearly preserved its quality.

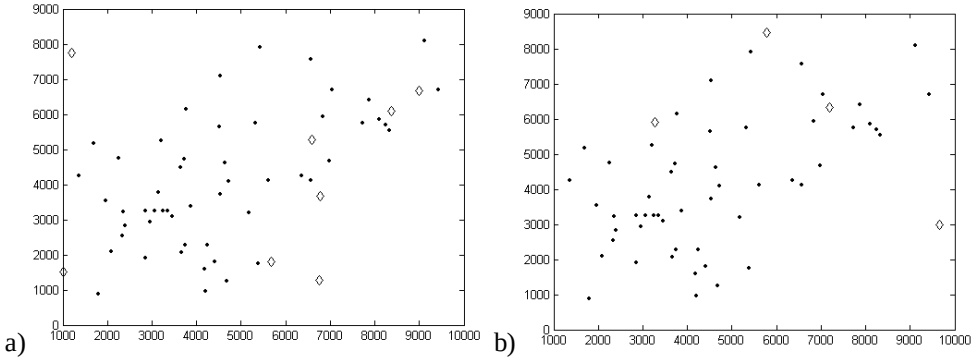

Figure 3. Examples of initial distribution of virtual sources (rhombus): 8 sources (a) and 4 sources (b). Filled circles indicate sample wells.

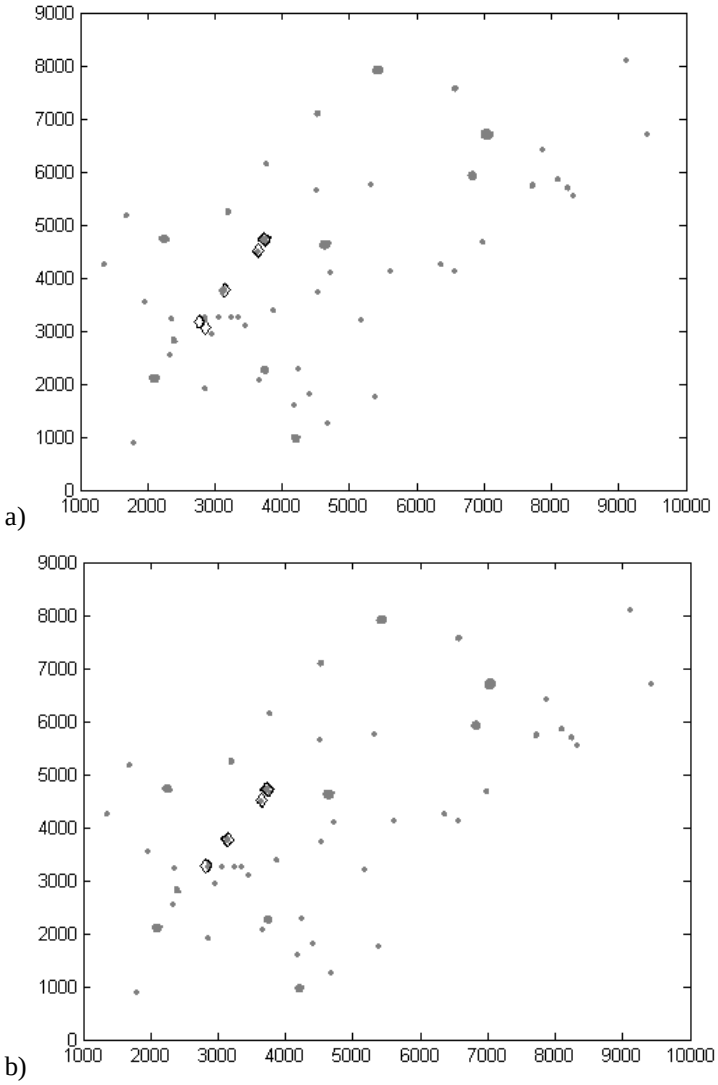

Figure 4. Final distribution of virtual sources (rhombus): 8 sources (a) and 4 sources (b). Grey filled circles present sample wells with diameter depending on local measurements' density

Positioning of virtual sources appeared to be very stable. But it is not because of denser sampling. In figure 4 monitoring wells are drawn as circles with diameters depending on the number of samples in the well. One can see rather frequent observed sample wells remote from virtual sources. Thus spatial location of virtual sources is due to the process.

Estimated virtual sources allow to describe the trend of the nitrate contamination by (8) and to make predictions on future development of the nitrate contamination. The 4 virtual sources indicate the zone responsible for the nitrate contamination process. Detailed monitoring in the vicinity of sources can be proposed to site administration.

4.3 Discussion on virtual sources

Two studied examples have different general features: groundwater level generally characterizes the water system; on the contrary the contamination process is caused by some real originating impact. This diversity arises within virtual sources analysis.

Considering obtained virtual sources due to these processes we find the approvement to such point of view. Groundwater level's virtual sources are distributed over the whole monitoring area somehow linking to surface water reservoirs and perhaps some groundwater ones.

Contamination virtual sources concentrate in the vicinity of the contaminating origin. All of these sources tend to the same area independently of their initial distribution. Geographically contamination virtual sources are located between two contaminated lakes, which can be the real origins of the groundwater contamination. Thus, virtual sources appeared to be not so virtual, but indicating to real features.

5. CONCLUSIONS

Independent virtual sources appeared to be useful tool allowing to describe

- dependencies between groundwater monitoring wells and surface water reservoirs. This information can be useful for analysis possibility of water migration in the neighbourhood of the contaminated water reservoirs.
- the localisation of the origin of the groundwater contamination.
- the model for groundwater contamination predictions.

Virtual sources also help to understand the general features on the process under study.

ACKNOWLEDGEMENTS

The work was partly supported by a grant 07-08-00257 of the Russian fund for fundamental researches (RFFI).

REFERENCES

Cherkasski V., Muller F., *Learning from data*, John Wiley Interscience, New York, pp. 437, 1998.

Joanes, D. N., Gill, C. A., Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society (Series D): The Statistician* V. 47: 183–189, 1998.

Lee T.-W., *Independent Component Analysis. Theory and Application*. Kluwer Academic Publishers, Boston, pp. 210, 1998.

Nuzhny A., Savelieva E., Jastrebkov A., Statistical Analysis for Extracting Features on the Groundwater Level Dynamics, in *Proc. of IAMG2007, Geomathematics and GIS analysis of resources, Environment and Hazards*, eds. P. Zhao, F. Agterberg, Q. Cheng, pp.723 - 726, 2007.

Riedmiller M., Braun H., A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm, in *Proc. of the IEEE Intl. Conf. on Neural Networks*, 1993

Üzümcül M., Frangi A.F., Reiber J.H.C., Lelieveldt B.P.F., Independent Component Analysis in Statistical Shape Models. Medical Imaging 2003, in *Proceed. of SPIE*, V.5032: 375-83, 2003.