Models for detecting excess of contaminants

A. F. Militino, M. D. Ugarte and B. Ibáñez
Departamento de Estadística e Investigación Operativa.
Universidad Pública de Navarra

Keywords: linear mixed model; nitrate concentration; spatial linear model; spatial statistics.

Introduction

Detecting areas with high nitrate concentration in groundwater is of public concern because an excess of these nitrogen compounds in any area represents serious environmental pollution. Groundwater is essential to agricultural operations and an important source of drinking water. Nitrate is one of the most problematic and widespread of the vast number of potential groundwater contaminants, and therefore high levels of nitrate in a given region may lead to undesirable effects on users and curtailment of groundwater usage. Due to the present difficulty of obtaining potable groundwater for industrial or drinking purposes, the European Union has enacted laws limiting the presence of nitrate levels to 50 mg/l. These laws require member states to identify vulnerable regions of nitrate concentration and to develop programs for reducing the pollution (91/696/CEE). Hence, governmental agencies require geostatistical researchers to develop methods for assessing nitrate concentration. The distribution of NO_3 and its displacement are a function of preferential paths and hence, its spatial dependence must be taken into account. But transport and retention capabilities for water and chemicals in an agricultural field are not only spatial but temporally variable. Therefore, a longitudinal model that incorporates spatial and temporal variability could explain better the behavior of these pollutants in groundwater.

Methods

Kriging has been the preferred method for modelling spatial variability, because of its optimality under correct assumptions on normality and proper knowledge of the covariance function (Cressie, 1993; Militino y Ugarte, 2001). In the last two decades, the interest has been on the inclusion of the temporal variable to produce more complete studies. Unfortunately, the spatio-temporal kriging extension is not straightforward, so a wide range of approaches have been attempted by several authors depending on both the type of data to be analyzed and the objective of the study. A review of geostatistical spatio-temporal models can be found in Kiriakidis y Journel (1999). A common approach considered by several authors (see, for example, Bogaert y Christakos (1997); Cressie y Majure (1997) is the use of spatio-temporal variograms that treat time as a third dimension. However, some of them present serious difficulties because, frequently, time is not only a new dimension but a new effect where the proposed covariance space-time models sometimes yield unrealistic results (Heuvelink y Webster, 2001). Recent studies show promising procedures such as the spatio-temporal Kalman filter (Mardia et al., 1998), flexible spatio-temporal variogram models (Fernández-Casal et al., 2003), and other semiparametric approaches (Angulo et al., 1998). In most of these works, information in time is assumed to be fairly exhaustive; and the aim is to provide predictions at new locations at any given time. However, situations found in practice require alternative methods, particularly when the focus is on providing global predictions in a given area from data collected in time.

Linear mixed models are linear models incorporating both fixed and random effects. Fixed effects allow for modelling the response variable mean and random effects allow for modelling its variance-covariance structure. The inclusion of random effects reduces the number of parameters in the specification of the variances and gives more flexible models (McCullogh y Searle, 2001). In particular, when we associate the same random effects to the same group of data, we are expressing a kind of correlation within this group of data. Longitudinal or repeated measures data are frequently modelled with linear mixed models (Verbeke y Molenberghs, 2000) because grouping data taken at different times can share the same random effect. Analogously, data spatially dependent may share the same covariance matrix of the error term as it is done in kriging.

Here, we are interested in providing an appropriate approach for modelling any measurements of environmental pollution that have been collected in time and space, by using this linear mixed model theory. Within this context, the possibility of obtaining a map where predicted global measurements are displayed on the area under study is of particular concern. It would help environmental agencies to study the spatial distribution of the given pollution measurement having taken into account all the measurements in time. To achieve this purpose, predictions at new locations are needed, which can be obtained by means of predicting new error terms in the linear mixed model.

Results

The data set consists of nitrate concentration measured in milligrams per liter (mg/l) sampled at 63 wells pouring into the Ebro river in a southern region of Navarra, Spain. The area is an extensive agricultural zone that accumulates large quantities of fertilizers and pesticides, being this the main reason of the presence of nitrates. Because of the seasonality that may be present in the fertilization process, data were collected in the four seasons during year 2000, leading to a total of $N = 252$ observations. In this application, there are only four observations per sampling location, one for each climatological season, hence, the use of the aforementioned spatio-temporal models is inappropriate. As an alternative, we propose the use of linear mixed models, taking advantage of the fact that they allow us to make a longitudinal analysis of spatially correlated data. In this way, we are able to provide a map of overall prediction of nitrate concentration, which could be used by environmental agencies and governments to take legislative and preventive actions in those regions displaying high levels of nitrate concentration.

Conclusions

The prediction procedure has appealing characteristics such as simplicity of formulation and computation, as it may be easily implemented using standard software such as SAS or S-PLUS. The model includes a fixed linear term where information about location is considered, a random term incorporating the repeated measurements collected at each location in time, and a random error allowing for different spatial dependence structures for each season. For the data considered here, the linear trend does not seem to present significant differences among the four periods considered. But, Winter and Fall present an almost negligible presence of spatial correlation in contrast with the highest values of Spring and Summer. The model takes into consideration these differences and presents a smooth approach for the whole year providing a unified map of nitrate concentration in the four seasons, with a measure of accuracy. Environmental researchers can derive practical benefits from this study by analyzing the map that provides the location of vulnerable regions, and by taking appropriate actions to minimize the problem.

Acknowledgements

The authors would like to thank the “Departamento de Medio Ambiente, Ordenación del Territorio y Vivienda del Gobierno de Navarra”, Spain, for providing the data. The work has been supported by the “Departamento de Salud” (project Res. 1878/2001) and the Spanish Ministry of Science and Technology (project AGL2000-0978).

Referencias

Angulo, J. M., González-Manteiga, W., Febrero-Bande, M. y Alonso, F. J. (1998). Semi-parametric statistical approaches for space-time process prediction. *Environmental and Ecological Statistics* 5:297–316.

Bogaert, P. y Christakos, G. (1997). Spatio-temporal analysis and processing of thermometric data over Belgium. *Journal of Geophysical Resarch-Atmospheres* 102(D22), 25:831–846.

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

Cressie, N. A. y Majure, J. J. (1997). Spatio-temporal modeling of livestock waste in streams. *Journal of the Agricultural, Biological and Environmental Statistics* 2:24–47.

Fernández-Casal, R., González-Manteiga, W. y Febrero-Bande, M. (2003). Flexible spatio-temporal stationary variogram models. *Statistics and Computing* 13:127–136.

Heuvelink, G. B. M. y Webster, R. (2001). Modeling soil variation: past, present, and future. *Geoderma* 100:269–301.

Kiriakidis, P. C. y Journel, A. G. (1999). Geostatistical space-time models: a review. *Mathematical Geology* 31(6):651–684.

Mardia, K. V., Goodall, C., Redfern, E. J. y Alonso, F. J. (1998). The kriged Kalman filter. *Test* 7(2):217–277.

McCullogh, C. E. y Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Interscience.

Militino, A. F. y Ugarte, M. D. (2001). Assessing the covariance function in geostatistics. *Statistics and Probability Letters* 52:199–206.

Verbeke, G. y Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag.