# Spatial uncertainty of groundwater-vulnerability predictions assessed by a cross-validation strategy: an application to nitrate concentrations in the Province of Milan, northern Italy

A. G. Fabbri[1], A. Cavallin[1], M. Masetti[2], S. Poli[1], S. Sterlacchini[3] & C. J. Chung[4]
[1]*DISAT, Università di Milano-Bicocca, Italy*
[2]*Dipartimento di Scienze della Terra, Università di Milano, Italy*
[3]*IDPA, Consiglio Nazionale della Ricerche, Italy*
[4]*Department of Earth Sciences, University of Ottawa, Canada*

## Abstract

Natural and anthropogenic factors are identified as critical in characterizing aquifer vulnerability in the Milan Province study area, where the impact of elevated concentrations of $NO_3^-$ is being assessed. In this contribution, map versions of continuous and categorical data layers are used to establish relationships between map units and the location of 305 water wells with nitrate levels either clearly above a threshold of 25 mg/l (impacted wells), or with wells clearly below that (non-impacted wells). The natural and anthropogenic data layers that are assumed to reflect (a) potential sources of nitrate, and (b) the relative ease with which nitrate may migrate in groundwater, are: population density, nitrogen fertilizer loading, precipitation and irrigation, the protective capacity of soils, land use, vadose zone permeability, groundwater depth, and groundwater velocity.

The water wells are separated first into the two groups to locate and recognize sites to be used to map high vulnerabilities using a prediction model based on the empirical likelihood ratio, ELR. Further partitions of the two sub-groups into prediction and validation wells allows setting up blind tests to cross-validate the predictions of relative vulnerability classes (ranks). Prediction-rate tables are

WIT Transactions on Information and Communication Technologies, Vol 43, ©2010 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/RISK100421

obtained and visualized either as histograms or as cumulative proportions of the study area in decreasing order of predicted vulnerability class versus the corresponding relative proportion of impacted validation wells, i.e., not used to predict. Predictions are thus compared and interpreted and repeated predictions are obtained using different sub-sets of prediction and validation wells in the two regions to obtain maps of uncertainty of the prediction classes. The target of the strategy used is not only to assess the goodness of predictions but also to estimate their reliability levels. In this application the uncertainty of the classes in the prediction map happens to be relatively high, which is due to the small number of water wells available in the spatial database.

*Keywords: spatial prediction modeling, spatial uncertainty, empirical likelihood ratio, aquifer vulnerability, nitrate concentration, cross-validation.*

# 1 Introduction

This contribution modifies a previous approach to the estimation of the vulnerability of groundwater to nitrate concentration [1]. The application problem in a study area around the city of Milan, in northern Italy, was to use the values of nitrate concentration in wells as an effective indicator of groundwater surficial contamination and of its vulnerability to further impact. This is in line with the European Council Directives 91/976/EC (Nitrate Directive) [2], which aims to protect surficial and groundwater against pollution caused by nitrate from agricultural practices. This is to encourage the designation of Nitrate Vulnerable Zones. Such zones can also be affected by non-agricultural activities, such as the presence of septic tanks and leaking municipal sewers in urbanized areas. Those authors used a database of several hundreds of water wells with measured nitrate concentration. The distribution of high concentration wells was related with a set of digital maps representing natural and anthropogenic factors in order to describe the potential sources of nitrate and their relative ease to migrate to groundwater. Their approach was based on a modeling technique termed “Weights-of-Evidence,” WoE, and in the application the digital maps were systematically binarized to facilitate the establishment of the spatial relationships.

In a recent paper, Fabbri and Chung [3] criticized these and other similar approaches due to either the absence of, or the incorrect use of, cross-validation to interpret the relative quality of the prediction maps generated by spatial modeling. The same database used by Masetti *et al.* [1] was kindly provided by those authors and it was used in this contribution.

Two major drawbacks were found in the original application of prediction modeling by those authors: (i) a very weak spatial support, and (ii) the lack of cross-validation of the prediction results. The approach was improved also by using a more sophisticated version the same model that avoids data binarization. An analytical strategy and modeling technique, based on cross-validation by blind tests, was used to obtain maps of relative vulnerability of groundwater to nitrate pollution. The study area database is described next, followed by the analytical methodology applied: the Empirical Likelihood Ratio function. The new resulting prediction maps and associated fitting- and prediction-rate curves are then discussed. They allow drawing conclusions on the quality and uncertainty of the relative measures of aquifer vulnerability interpreted using spatial cross-validation.

## 2 The study area database

The Milan Province study area covers approximately 2000 km$^2$ and contains different agricultural and industrial land uses. Its hydro-geologic setting is characterized by many aquifers with complex interactions and recharge conditions due to a large variety of soils, land uses and irrigation networks. Figure 1 shows the location of the study area and of the city of Milan. Figure 2a shows the location of the water wells. The main aquifer in the area is described in [1]. It is termed Traditional Aquifer, consists of Pliocene–Pleistocene sediments, and is unconfined. It has a transmissivity between 5x10$^{-2}$ and 1x10$^{-3}$ m$^2$/s, and permeability between 5x10$^{-3}$ and 1x10$^{-8}$ m/s with a thickness between 60 and 120 m. The composition is of gravels and sands with an increase of clay-silt layers southward. The regional flow is also southward and the groundwater depth varies from 30 m to the north to 5 m to the south. Over 300 water wells are uniformly distributed throughout the area that are monitoring four times a year the nitrate concentration. That oscillates between minima around 1.0 mg/l and maxima around 70 mg/l, with a median close to 20 mg/l. According to European Community standards [2], the guide value of nitrate in soil is of 25 mg/l. The most impacted part of the study area is in the northeast with values exceeding 50 mg/l. The concentration decreases southward to lower values less than 10 mg/l.

According to those authors, the concentration monitored throughout time appears constant without temporal trends and with differences between the northern and the southern parts of the study area. Statistical methods for regional groundwater vulnerability assessment can be used to correlate the measured occurrence of contaminants with the distribution of natural and man-induced factors represented as maps. For this purpose, a spatial database of

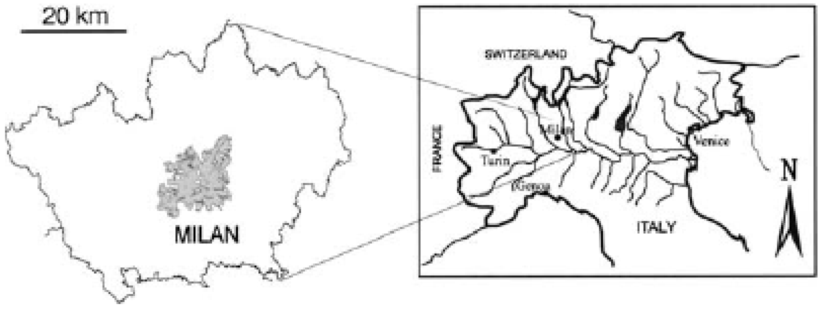


Figure 1: Location of the study area in northern Italy (modified from [1]).

those factor maps was constructed by Masetti *et al.* [1], which included the following maps described in Table 1 and in Figure 2: occurrence of impacted and non impacted wells, indicator map of study area, groundwater recharge, land use, soil protection capacity (as categorical maps), groundwater depth, groundwater velocity, main annual irrigation, nitrogen fertilizer loading, population density, and rainfall (as continuous value maps).

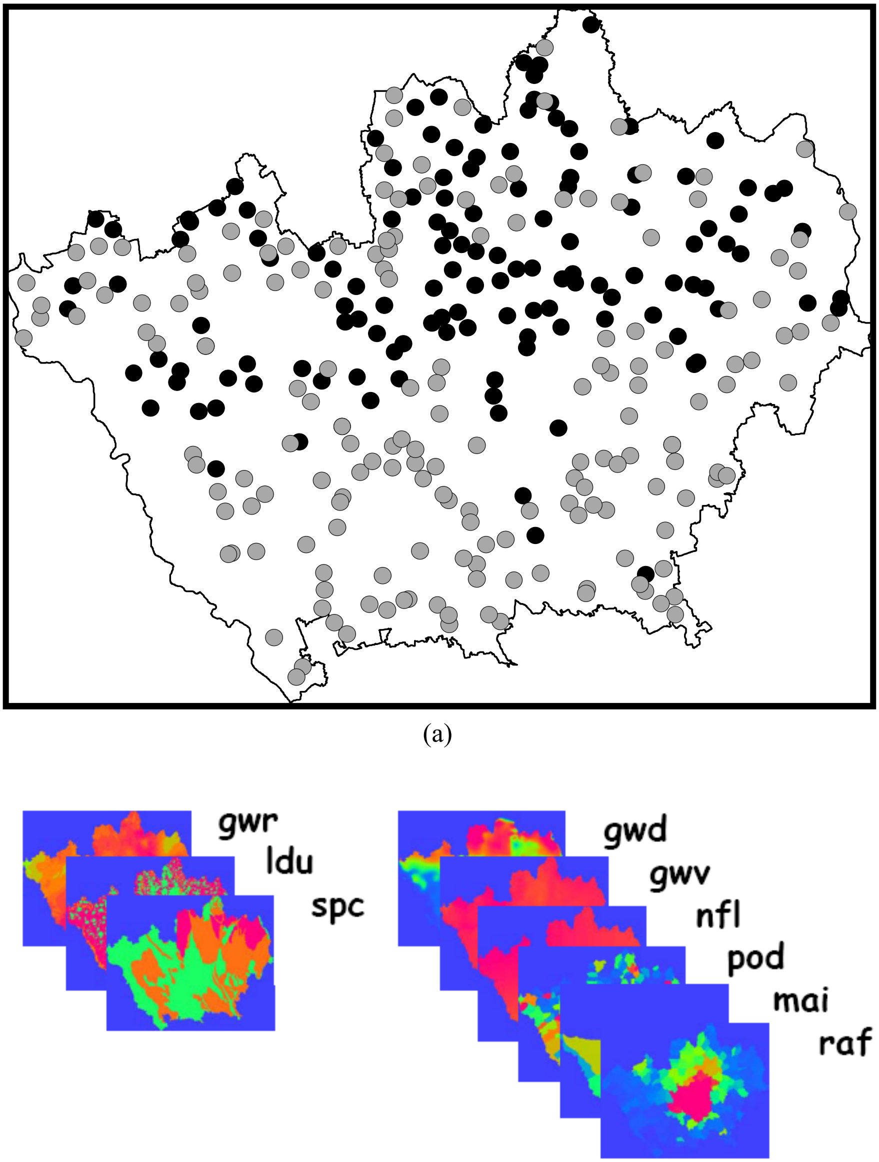
In practice, the spatial database consists of three sets of digital maps/images of the same pixel resolution of 20 m and of size of 3300 pixels and 2665 lines. Within that image space, the study area occupies 4,908,305 pixels and the remaining 3,886,095 pixels are outside the study area.

The first set of images shows the distribution of 133 wells with high nitrate concentration, ≥ 25 mg/l, and that of 172 wells with lower concentration, ≤ 24 mg/l. The well location is assigned to single pixels, 133 and 172, respectively. Together, the high and the low concentration wells correspond to 305 different pixels in the study area. Each well pixel is assigned a sequential numeric label with values 1 to 133 and 1 to 172 for the high and low concentration wells. A third image indicates with value 1 the study area and with value 0 the outside.

The second set of images consists of the three categorical maps of Table 1 and Figure 2(b), with short names **gwr**, **ldu** and **spc**, with their respective value ranges.

Table 1: Wells, natural and anthropogenic factors in the study area database.

| **Occurrence data and study area, *DSP*s** | | |
|---|---|---|
| Factor map | Data range | Description |
| 133h<br>172l<br>Study_area | 1 to 133<br>1 to 172<br>1 and 0 | Index of water well ≥ 25 mg/l $NO_3^-$<br>Index of water well ≤ 24 mg/l $NO_3^-$<br>Study area indicator is 1, outside is 0 |
| **Categorical data, *ISPs*** | | |
| Factor map | Data range | Description |
| **gwr** | classes 6 to 15 | Combination of **raf** and **mai** x a function of **spc** as infiltration coefficient |
| **ldu** | classes 1 to 3 | Urban, agricultural and woods |
| **spc** | classes 1 to 3 | Low, moderate and high |
| **Continuous data, *ISP*s** | | |
| Factor map | Data range | Description |
| **gwd** | 1-50 | m |
| **gwv** | 11.2-18.1 | - ln m/s |
| **nfl** | 1-428 | kg/h/y |
| **ma** | 1-790 | mm |
| **pod** | 43-7933 | inhabitants/km$^2$ |
| **raf** | 808-1253 | mm/y |



(a)

(b)

Figure 2: The spatial database for this study described in Table 1. (a) The distribution of impacted water wells (black circles), and of non-impacted water wells (gray circles). The impacted water wells represent the direct supporting pattern ***DSP***. (b) The data layers represent the indirect supporting patterns, ***ISP***s.

The third set of images contains the continuous value maps of Table 1, **gwd**, **gwv**, **nfl**, **mai, pod** and **raf**, with their respective value ranges. Altogether, the database initially consists of 12 images.

## 3 Analytical method and strategy

The modeling of spatially distributed data targets a relationship established in a spatial database in order to apply it to identify in a study area classes of pixels that to a different degree satisfy that relationship. Chung and Fabbri [4] and Chung [5] used an approach they termed Favorability Function modeling, **FF**. The basic concept is that "*a value at a pixel in the final prediction map can be assumed to be computed as a mathematical function generating the target pattern, given the information of the supporting patterns at the pixel.*" In that general approach they proposed: data structures functional to prediction modeling, prediction models, and strategies for processing and modeling interpretation in a sequence of steps. In simple terms, their approach is summarized next.

### 3.1 Data structures for spatial prediction modeling

A basic element in mathematical modeling is to frame a problem in terms of a proposition, i.e., a mathematical statement that can be established as true or false. For instance, we can use our aquifer vulnerability by nitrate concentration in the following proposition,

$$T_p\text{: a point } p \text{ is impacted by nitrate concentration.} \qquad (1)$$

In support of the proposition in (1) we can use the information available in our spatial database, consisting of the spatial data, described in the previous section, that now we can term as: (i) the ***direct supporting patterns, DSP*** (the 133 pixels with wells with ≥ 25 mg/l nitrate, or alternatively the 172 pixels with wells with ≤ 24 mg/l nitrate), and (ii) the ***indirect supporting patterns, ISP*** (the natural and anthropogenic factor maps listed in Table 1, with either categorical or continuous data). We will be using a mathematical model to establish the spatial relationships between the ***DSP*** and the ***ISP***s and use them to obtain a ***prediction image***, ***PI***, based on some function related with the frequency of occurrence of those relationships in the database. Such image will have to be displayed by recoding the predicted values into a ***prediction pattern***, ***PP***, so that appropriate colors can be assigned. At this point we would like to interpret the relative quality of the ***PI*** and/or ***PP***, for instance by verifying how the pixels of the ***DSP*** are distributed among the prediction classes obtained. For this we can generate fitting-rate tables, histograms or cumulative curves. Such fitting rates, however, may not tell us much of the prediction power of our modeling, unless we have another different ***DSP*** to use to validate our ***PI***. The distribution of the pixels in this second ***DSP*** among the prediction classes would provide us with a prediction-rate table, generally different from the fitting-rate table. As a matter of fact, we can consider the observed ***DSP*** as part of a larger pattern that we can term ***Target pattern***, ***TP***. The ***TP*** is what we would like to have from our modeling but of it we only have a part represented by the ***DSP*** and it is this ***DSP*** that we can use to establish the spatial relationships. Our aim is to obtain a ***TP*** that estimates the degree that our proposition in (1) is true for the entire study area. For that we will need some assumptions later on but first we will need some models to establish the spatial relationships.

### 3.2 Three favorability-function models based on probability theory

Here we will consider the empirical likelihood ratio function, ***ELR***, for spatial prediction modeling. It is one of the most suitable functions among several other ones that can be used as a favorability function. To establish the ***ELR*** function for the spatial prediction model, let us define the ***TP***, as the area with pixels to be impacted by nitrate concentration. Suppose hypothetically, that the study area is divided into two exclusively disjoint sub-areas, ***M***, the areas impacted by nitrate concentration, and $\overline{\boldsymbol{M}}$ , the remaining non-impacted areas. Consider two joint multivariate frequency distribution functions (m dimensional because we have m indirect supporting patterns, ***ISP***s) of m supporting patterns from ***M*** and $\overline{\boldsymbol{M}}$ . The two m-dimensional multivariate frequency distribution functions at the pixel p with m pixel values, ($c_1, \cdots , c_m$ ) are expressed by $\boldsymbol{f}(c_1, \cdots , c_m \mid \boldsymbol{M})$ from the impacted area, ***M*** and $\boldsymbol{f}(c_1, \cdots , c_m \mid \overline{\boldsymbol{M}})$ from the non impacted area, $\overline{\boldsymbol{M}}$ . The ***likelihood ratio function*** (see [6–8]) at p is defined as the ratio:

$$\lambda(p \mid c_1, \cdots, c_m) = \frac{\boldsymbol{f}(c_1, \cdots, c_m \mid \boldsymbol{M})}{\boldsymbol{f}(c_1, \cdots, c_m \mid \overline{\boldsymbol{M}})} . \tag{2}$$

The same likelihood ratio function is also commonly used in discriminant analysis for classification in statistical analysis [9]. If the m pixel values, ($c_1, \cdots , c_m$) at p provide useful information for identifying areas likely to contain the ***TP***, then $\boldsymbol{f}(c_1, \cdots, c_m \mid \boldsymbol{M})$ is likely larger than $\boldsymbol{f}(c_1, \cdots, c_m \mid \overline{\boldsymbol{M}})$ . That means that the frequency (probability) that the pixel has the m values, ($c_1, \cdots , c_m$ **),** assuming that the pixel belongs to ***M*****,** should be larger than the frequency (probability) that the pixel has the same m values, assuming that it comes from $\overline{\boldsymbol{M}}$ . In this case, we have that $\lambda(p \mid c_1, \cdots, c_m) > 1$. On the other hand, if the pixel p is likely to belong to $\overline{\boldsymbol{M}}$ , then $\lambda(p \mid c_1, \cdots, c_m) < 1$. The range of $\lambda$ goes from 0 to $\infty$ .

As discussed in Heckerman [10] and Pearl [11], $\lambda(p \mid c_1, \cdots, c_m)$ , or any of the monotone non-decreasing functions of $\lambda$, can be used as good measurements for expressing the likelihood of containing the target pattern, ***TP***, at each pixel. One of the simplest monotone functions for the target mapping at each pixel p is the logarithm of the likelihood ratio function:

$$\boldsymbol{WoE}\{ p \mid c_1, \cdots , c_m \} = \log_e \lambda(p \mid c_1, \cdots, c_m) , \tag{3}$$

which was termed as the "***weights of evidence***" model by Peirce [12], (see also Spiegelhalter [13]); and it ranges from $-\infty$ to $+\infty$. While ***WoE***$\{$ p$|$ $c_1, \cdots, c_m \} > 0$ if the pixel p is likely to contain the ***TP***, if it is not so, ***WoE***$\{$ p$|$ $c_1, \cdots, c_m \} < 0$.

Another monotone function of $\lambda$ is the certainty factor function. The certainty factor, ***CF***, was first introduced, by Shortliffe and Buchanan [14], as a model for representing and combining evidences in medicine. Chung and Fabbri [5] adapted the method to represent GIS information in the application to landslide hazard prediction. The following definition of the certainty factor is one of them:

$$\boldsymbol{CF}\{\, \mathrm{p} \mid \mathrm{c}_1, \cdots, \mathrm{c}_m \} = \frac{\lambda(\mathrm{p} \mid \mathrm{c}_1, \cdots, \mathrm{c}_m) - 1}{\lambda(\mathrm{p} \mid \mathrm{c}_1, \cdots, \mathrm{c}_m) + 1}, \tag{4}$$

which represents the level of likelihood that pixel p is impacted by nitrate concentration given the m pixel values, ($c_1 \cdots, c_m$) and it ranges between -1 and 1. ***CF*** is equal to zero if the likelihood ratio function is equal to 1 and the value of ***CF*** increases to 1 if the pixel p is likely to be impacted by nitrate concentration. Thus, ***CF*** can be used as a measure of the likelihood. An excellent discussion on this subject was provided by Heckerman [10].

Any one of the three functions in (2), (3), and (4) can be used as a favorability function, and it complies with the following two properties: (**p1**) it represents a relative level of likelihood that a pixel p contains a part of the ***TP***, and (**p2**) using the known part of the ***TP*** in the training area we should be able to construct a favorability function and to establish the uncertainty of the function. If we construct an **FF** satisfying (**p1**), then we can generate a prediction image, ***PI***, by computing a function value at every pixel in the study area. Property (**p2**) states that we can generate the favorability function and its corresponding uncertainty function from the training area.

In addition we need to see that the following three assumptions are reasonable: (**a1**) the known impacted pixels in the training area are a random selection of all known and future impacted pixels; (**a2**) the supporting patterns are correlated with the ***TP***; and (**a3**) the processes generating the impacted pixels is not a random process but it follows a certain rule. With assumption (**a1**) we are allowed to possibly extend the **FF**, which we have estimated in the training area, to the other pixels in the rest of the study area. Assumption (**a3**) allows us to model the **FF**, and assumption (**a2**) allows estimate the **FF** using the known part of the ***TP*** in the training area.

These three functions have been extensively utilized to express and propagate quantitative reasoning, knowledge and uncertainties through a complex inference network for artificial intelligence computer systems [11, 15].

All three functions are dependent on two frequency distribution functions, namely, $\boldsymbol{f}(\mathrm{c}_1, \cdots, \mathrm{c}_m | \, \boldsymbol{M})$ and $\boldsymbol{f}(\mathrm{c}_1, \cdots, \mathrm{c}_m | \, \overline{\boldsymbol{M}})$. Without knowing all the ***TP***, it would not be possible to obtain these two frequency distribution functions. The frequency distribution functions and the ratio can be estimated from m indirect supporting patterns, ***ISP***s, and the known part of the ***TP***, the ***DSP***, in the training area.

We have categorical and continuous ***ISP***s and in practice the likelihood ration function can be estimated as a multiple of two estimated likelihood ration functions, one for categorical data and one for continuous data. The different ***ISP***s are then integrated for each pixel using the combination rules of the model selected: in our case the ones for ***ELR*** that differ from the ones for the ***WoE*** or the ones for the ***CF***. Chung [4] discussed in detail the use of the likelihood ration function for spatial prediction modeling. Chung and Fabbri [16] compared the results of applying the three models to the same spatial database and observed that they generate identical ranks of the prediction values.

### 3.3 Strategy of processing

In the previous sections, to assist in modeling, a terminology was presented: ***ISP***, ***DSP***, ***PI***, ***PP***, ***TP*** and fitting and prediction rates. In addition, it was mentioned that to obtain the ***TP*** we have to consider the two properties, (**p1**) and (**p2**) and the three assumptions, (**a1**), (**a2**) and (**a3**). In our case the ***DSP*** consists of the 133 impacted pixels and/or the 172 non-impacted pixels, which together comprise a set of 305 pixels of 20 m resolution. Correspondingly, the ***ISP*** consists of the 9 values at each of the 305 pixels of the map factors described in Table 1. These are the basic spatial support of the database. The entire database for the study area, however, consists of 4,908,305 pixels and in it we have 305 pixels for which the ***DSP-ISP***s spatial relationships can be established. How should we develop a processing strategy?

A first obvious strategy in such a situation is to consider a subset of the database with the 305 pixels corresponding to the impacted and non-impacted pixels as a training area and the rest of the database as a study area. Apply the models to the training area and the statistics obtained from it is used to classify the pixels in the study area. This will generate a ***PI*** and a ***PP*** to be then interpreted. How good would they be? How good can we consider the prediction? The distribution of the 133 impacted pixels in the various classes of predicted values is not telling us much about the prediction power of the modeling. It only expresses how well they aggregate on the ordered classes.

To study the relative “goodness” of the prediction classes we have to use some form of empirical validation. For example, we can proceed by repeating the prediction within the 305 pixel database using a randomized subset of the training area (for instance, 80% of 133 = 106 impacted pixels to predict the remainder impacted pixels, 20% of 133 = 27). We could then repeat the random selection n times, pretending each time not to know the 20% of the impacted pixels, i.e., a blind tests, thus obtaining n new prediction images and the corresponding prediction rates. The analysis of those prediction rates will help us in assessing the relative goodness of the ***PI*** generated using all the 133 impacted pixels. The ***PI*** hopefully estimates the ***TP***.

A second strategy could be to subdivide the training area and the study area into a W-region and an E-region, and then use the training area of one region to

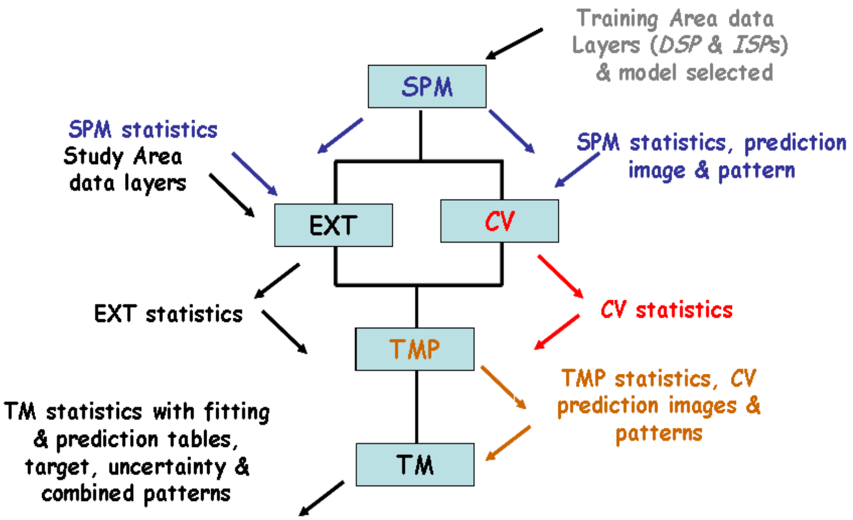


Figure 3: Strategy of processing with **STM**. Routine names are as follows: SPM, Spatial Prediction Modeling, CV, Cross-validation, TMP, Target Mapping Preparation, TM Target Mapping, and EXT, Extension of prediction statistics to study area.

establish the spatial relationship and extend the prediction to the other region. This will generate a pair of ***PI***s, one for the E-region and another for the W-region, with the corresponding prediction rates. The first strategy has been used in the application described in the next section. There was no reason to apply the second strategy in this case.

### 3.4 Software for spatial target mapping

The processing strategy is applied to the database by means of interactive and iterative procedures using the Spatial Target Mapping software, **STM** [17], in conjunction with a general-purpose spreadsheet and optionally a geographic information system. Fabbri and Chung [18], Fabbri *et al.* [19] and Chung *et al.* [20] discussed earlier versions of the software.

The main steps in the Spatial Target Mapping analytical procedure are described in Figure 3, as a flowchart with five **STM** routines (namely, SPM, CV, EXT, TMP, and TM) for which inputs and outputs are specified in the illustration. The SPM routines are run first to obtain a ***PI*** for the training area and the related statistics. Then, the CV routines are run to perform a cross-validation to interpret the ***PI*** generated by SPM. If the study area is the same as the training area, the TMP routines are run next to obtain a file of all the ***PI***s requested by the statistics of the CV routines. If the study area is not equal to the training area, the EXT routines are run after the SPM routines to use the statistics from the training area, to generate a ***PI*** in the study area. Outputs of Target Mapping, TM, are the ***PI***, the ***PP***, the associated uncertainty pattern, ***UP***, and the prediction-uncertainty combination pattern, ***CP***. In addition, tables are generated with the distribution of the future (validation) occurrences of impacted wells, in the prediction classes. A visualization of this distribution, the prediction-rate curve, describes the "goodness" of the prediction.

## 4 Results and interpretation

The application of the **ELR** model was made using a subset of the study area. We used a *training area* consisting of the 305 pixels with wells with known nitrate concentration, in which 133 pixels represent the distribution of wells with ≥ 25 mg/l of $NO_3^-$, the ***DSP*** (i.e., the known part of the ***TP***). In addition, the *training area* comprises the various ***ISP***s with the corresponding 9 values for each of the 305 pixels. The result of the modeling is a prediction image, ***PI***, of 305 pixel with predicted values, ranging between +∞ and -∞. Figure 4 shows the corresponding prediction pattern, ***PP***, of 305 colored isolated pixels in a background of the study area of nearly 5 million pixels. It is almost impossible

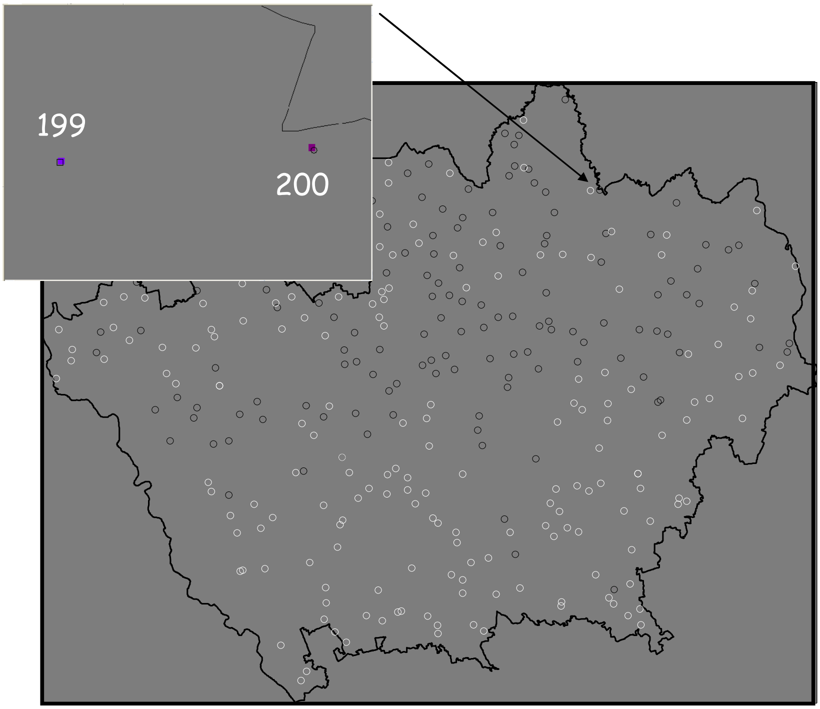
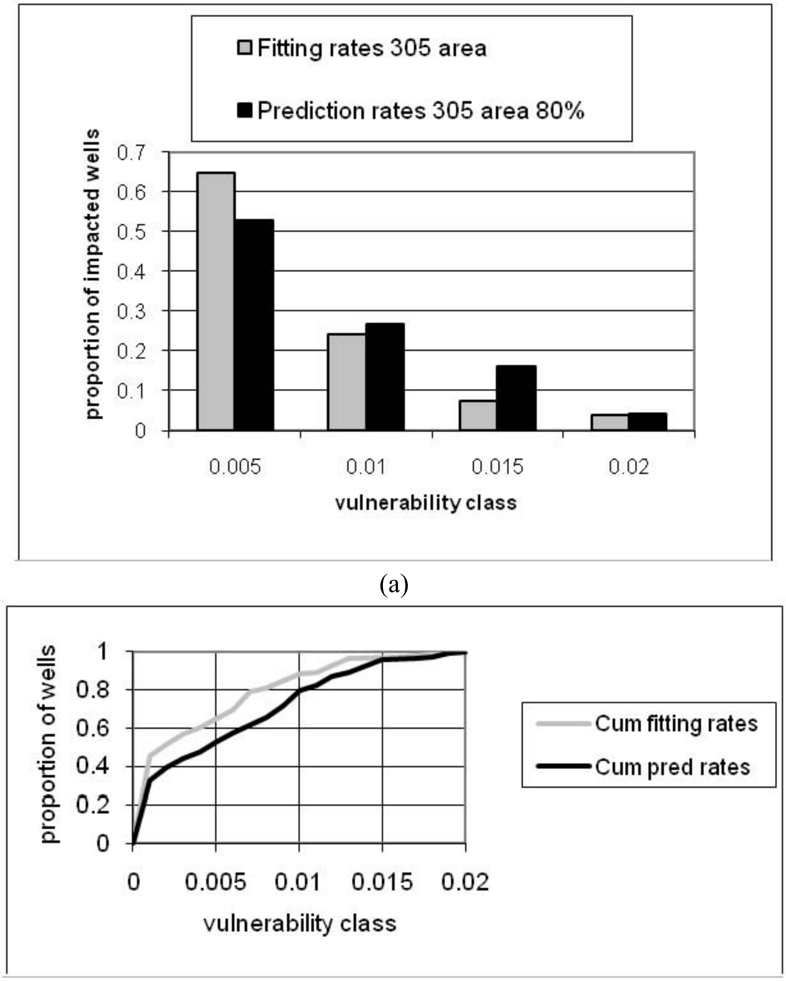


Figure 4: Prediction pattern of the training area of 305 pixels overlaid with the distribution of impacted and non-impacted wells for added visibility. The magnified inset shows the ranked prediction values of two pixels.



(a)

(b)

Figure 5: Fitting and prediction rates for the 305 pixel training area. (a) Histogram for four equal area classes with the highest values obtained. (b) Cumulative curves for the same rates.

to see the 305 pixels so that an enlarged inset displays two of them with predicted values as ranks 199 and 200, respectively. The pixel values were arranged in decreasing order to generate a histogram for an arbitrary number (here 200) of equal area classes corresponding to their ranks. Figure 5(a) shows such a histogram as gray columns that represents how well the 133 impacted wells are all contained within the four highest-value classes, each of 0.5% of the training area. In this case the ranks of the fitting rates are all within the 180-200 range. We term these rates as fitting rates because they indicate the fitting of the impacted wells in the classes. As shown by the gray curve in Figure 5(b), a cumulative fitting-rate curve with the same equal area classes as those in Figure 5(a) can also describe the distribution of the impacted wells in successively lower predicted classes. Such a curve is a different way of visualizing the prediction results.

After applying the **ELR** model, the statistics of the spatial relationships computed for the 305 pixels of the training area, can be extended to the entire study area to generate a new ***PI*** for the remaining 4,908,000 pixels in it. The corresponding ***PP*** is shown in Figure 6(a), where for visualization the predicted values have been sequenced in descending order and transformed into 200 equal-area classes (0.5% of the study area, i.e., 19,540 pixels each class). In Figure 6(a) to the pixels within the rank value range of 200 a pseudo-color look-up table was assigned so that convenient groups of classes could be visualized as indicated by the legend.

The ***PP***, however, even if it is the best prediction we can generate, because it uses all the impacted wells available, the ***DSP***, does not show its prediction "goodness" or prediction power. For this, short of waiting for future further impacts to the more vulnerable areas, or of drilling more water wells, we can empirically assess the prediction quality by repeating the predictions many times with different partitions of the impacted wells. For instance, we can repeat the analysis 10 times, each time using only a randomized subset of 80% of the 133 impacted pixels (106 wells), generate the prediction classes and verify which predicted classes contain the remaining 20% impacted pixels (27 wells). Of course, we can select higher number of iteration and different partitions of the wells, depending on the known characteristics of the database.

This generates 10 new prediction images, and 10 new prediction-rate curves can be obtained by looking each time at the proportional distribution of the 27 impacted wells now used for validation and not for prediction. The histogram in Figure 5(a) shows the prediction rates as black columns and the black curve in Figure 5(b) shows the average of the 10 prediction rates obtained. In the diagram, the horizontal axis indicates the cumulative proportion of study area and the vertical axis the corresponding cumulative proportion of the impacted wells (133 for the fitting rates, and different sets of 27 for the prediction rates).

The prediction-rate curve shows that the prediction power is not what we would expect from the fitting rate curve.

We can explore the distribution of the uncertainty values for the different classes of the ***PP***. The cross-validation analysis, in this case of 10 predictions with 10 different 80% subsets of randomized occurrences, provides us with a measure of the uncertainty associated with the class assignments of the pixels.

We can generate the statistics of the variation of values of the pixels in the 10 prediction images computed, and obtain an uncertainty pattern, ***UP***, shown in Figure 6(b). It shows the relative percentage of variation among the 10 predicted values for each pixel of the ***PP***. Figure 7 shows a threshold of the ***PP*** in Figure 6(a) at levels below the 5% of the uncertainty, the variance, from the ***PP*** in Figure 6(b). As can be observed comparing Figure 6 with the combined pattern, ***CP***, in Figure 7, some of the highest prediction classes and some of the lowest

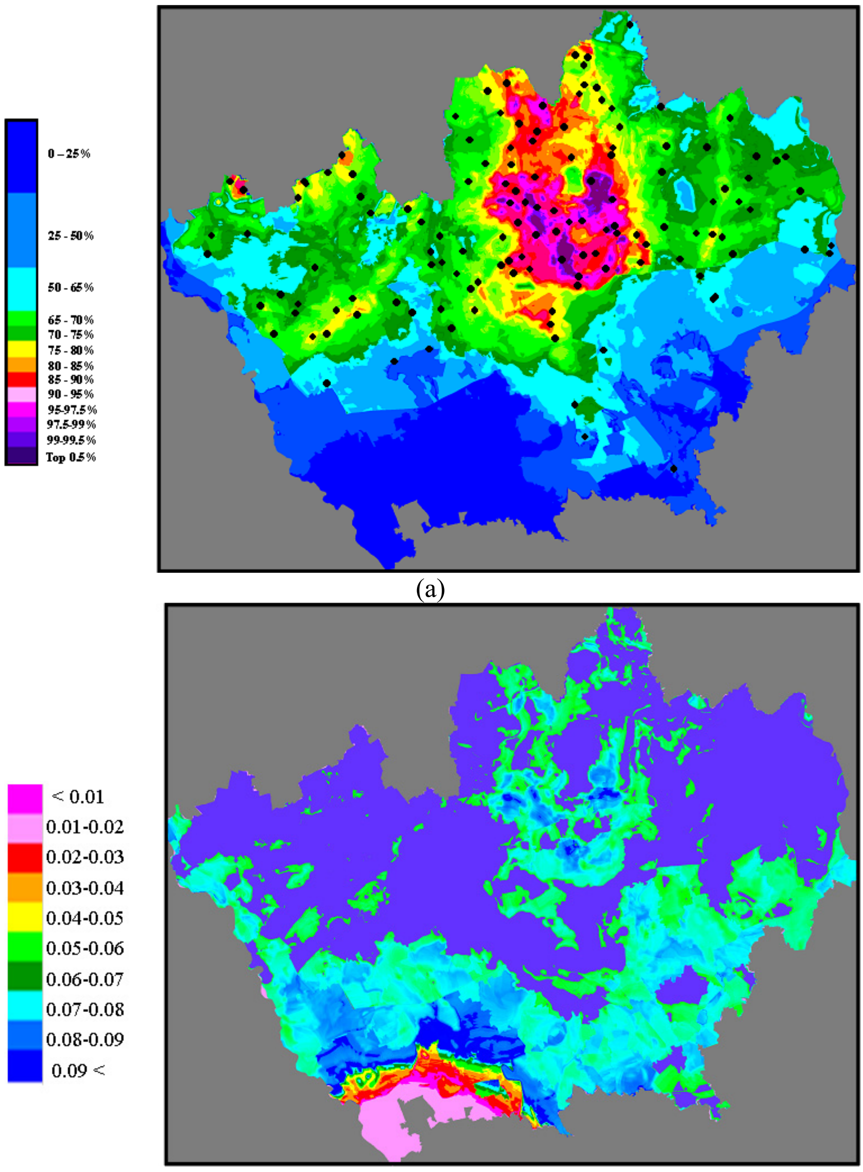


Figure 6: The prediction pattern, ***PP***, obtained extending the prediction from the 305-pixel training area to the study area in (a), with differently colored percentages of ranks and the distribution of the impacted wells. In (b), the corresponding uncertainty pattern, ***UP***, is shown with differently colored intervals of relative variance.

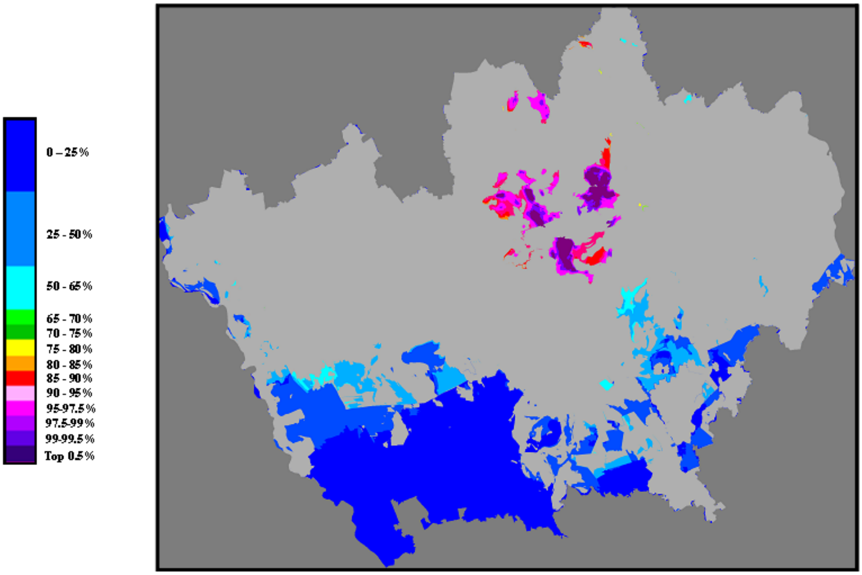


Figure 7: The combination of prediction pattern and uncertainty pattern, ***UP***, at 5% of the relative variance. The color legend is the same as in Figure 6(a).

correspond to very high uncertainty, i.e., with high relative variance, well over 5%.

The ***CP*** in Figure 7, and not the ***PP*** in Figure 6(a), should be considered the meaningful result of the analysis, i.e., the acceptable estimation of the ***TP***.

Figure 8 shows the 10 prediction-rate curves together with the average prediction-rate curve also displayed in Figure 5(b), as a black curve. The prediction has a remarkably wide range of variation, considering that 2% of the training area contains all the predicted wells.

The ***PP*** with the respective ***UP***, ***CP*** and prediction-rate curve, describe the relative "goodness" of the database and its capability to represent aquifer vulnerability to pollution by $NO_3^-$. The spatial support of only 305 pixels in the training area to establish spatial relationships and extend them to 4,908,000 pixels of the study area is clearly very weak. This can be observed in the ***CP*** of Figure 7 and the variability of prediction-rate curves in Figure 8. For this reason, not much confidence can be assigned to the predicted values outside the area visible in the ***CP*** of Figure 7.

These are the basic properties of the spatial database. Any other prediction or interpretation will have to be related with these results.

The **STM** software makes it possible to assess the basic properties of the database. Any other prediction model application or analytical strategy will have to be take these results into consideration. **STM** implies an analytical strategy based on cross-validation of the prediction results, so that their relative

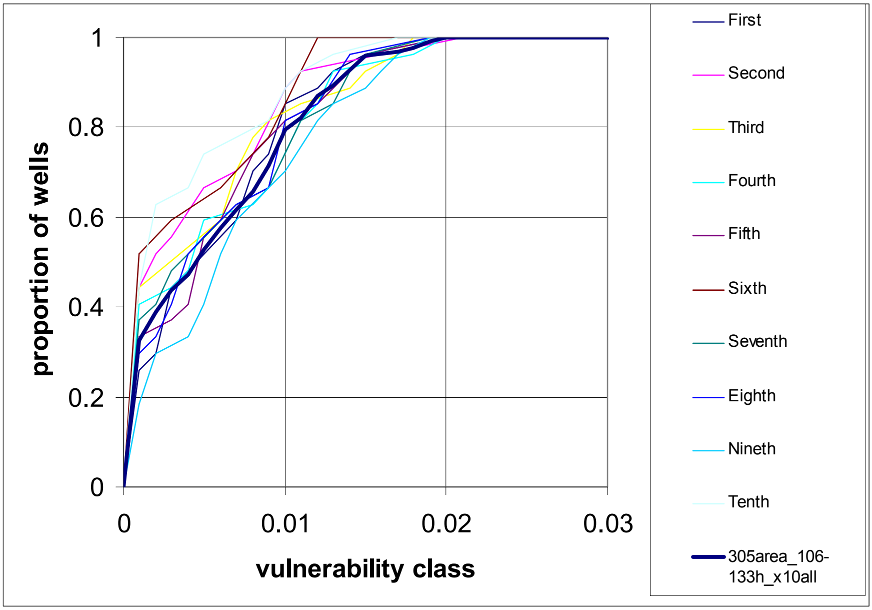


Figure 8: The ten cumulative prediction-rate curves obtained by cross-validation and their average curve as a heavier line. They correspond to the 10 predictions described in the text.

"goodness" and the robustness of the predicted values can be assessed. Had such a strategy been used earlier in this application, it would have guided to a more cautious interpretation of the significance of the database.

## 5 Concluding remarks

In this study we have revisited the application of spatial modeling by Masetti *et al.* [1], who provided their database for aquifer vulnerability assessment in a study area around the city of Milan, in northern Italy.

Two major drawbacks were found in the original application: (i) nitrate concentrations were observed at 305 wells only, 133 wells were treated as impacted pixels, 172 as non-impacted pixels, and in addition, the remaining study area was also assumed as a set of non-impacted pixels to establish the spatial relationships; and (ii) the lack of cross-validation to establish the uncertainty of estimation. Those drawbacks were corrected in this analysis and an improvement was made by using the empirical likelihood model, **ELR**, instead of its simpler version, the **WoE,** in which spatial data are generally binarized.

A different analytical approach was used, that redefined data structures for modeling, included three favorability function models based on probability theory, a processing strategy based on the properties of the database, and a new spatial target mapping software, **STM**. The three models are known to generate identical ranks of predicted values. The analysis could only be based on the very weak spatial support of a training area of 305 pixels to generate a prediction in a study area of almost 5 million pixels. The resulting prediction pattern is affected by high uncertainty that was assessed by cross-validation via the 10 blind-tests of 10 iterations of predictions with subsets of the training area database. The **STM** software permitted the suitability assessment of the entire database for aquifer vulnerability prediction.

The results are in line with the views of Chung and Fabbri [3] on the need to apply cross-validation via blind-tests to explore the viability of spatial databases in prediction modeling. It is commendable that scientists share their data to broaden the interpretability of the prediction patterns and the understanding of the prediction models.

In this application we have not discussed the causes of the impacts on the water wells in line with the new results obtained. This will be the subject of a future contribution.

## Acknowledgement

SpatialModels Inc., Ottawa, Canada, has kindly provided a beta version of its new Spatial Target Mapping System, **STM** (www.spatiamodels.com).

## References

[1] Masetti, M., Poli, S. & Sterlacchini, S., The use of the weights-of-evidence modeling technique to estimate the vulnerability of groundwater to nitrate contamination. *Natural Resources Research*, **16(2)**, pp. 109-119, 2007.

[2] European Community, Council Directive 91/676/EEC of 12 December 1991 concerning the protection of waters against pollution caused by nitrates from agricultural sources, (Nitrate Directive) OJ L 375, 31.12.1991: pp. 1–8, 1991.

[3] Fabbri, A.G. & Chung C.F., On blind tests and spatial prediction models. *Natural Resources Research*, **17(2)**, pp. 107-118, 2008.

[4] Chung, C.F. & Fabbri, A.G., Representation of geoscience data for information integration. *Jour of Non-renewable Resources*, **2(2)**, pp. 122-139, 1993.

[5] Chung, C.F., Using likelihood ratio functions for modeling the conditional probability of occurrence of future landslide for risk assessment. *Computers & Geosciences*. **32**, pp.1052-1065, 2006.

[6] Duda, R.O., Hart, P.E. & Nilsson, N.J., Subjective Bayesian methods for rule-based inference systems. *Procs. Natl. Computer Conf., 1976,* pp. 1075-1082, 1978.

[7] Kshirsagar, A.M., *Multivariate Analysis*, Marcel Dekker Inc., New York, 534p., 1972

[8] Press, S.J., *Applied Multivariate Analysis*, Holt, Rinehart and Winston, Inc., New York, 521 p., 1972.

[9] Cacoullos, T., *Discriminant Analysis and Applications*, Academic Press, New York, 434 p., 1973

[10] Heckerman, D., Probabilistic interpretations for MYCIN's certainty factors. *Uncertainty in Artificial Intelligence,* eds. L.N. Kanal & J.F. Lemmer, Elsevier Science Pub., North-Holland, pp. 167-196, 1986.

[11] Pearl, J., *Probabilistic reasoning in intelligent systems,* Morgan Kaufmann Pub., San Mateo, California, 552 p., 1988.

[12] Peirce, C.S., *Collected Papers 1931-1935*, Cambridge, Harvard University Press, 1978.

[13] Spiegelhalter, D.J., A statistical view of uncertainty in expert systems. *Artificial Intelligence and Statistics*, ed. W.A. Gale, Addison-Wesley Pub., Reading, Mass., pp. 17-55, 1986.

[14] Shortliffe, E.H. & Buchanan, B.G., A model of inexact reasoning in medicine. *Math. Biosciences*, **23**, pp. 351-379, 1975.

[15] Kanal L.N. & Lemmer, J.F., *Uncertainty in Artificial Intelligence*. Elsevier Science Pub., North-Holland, 509 p., 1986.

[16] Chung, C.F. & Fabbri, A.G., Three Bayesian prediction models for landslide hazard., *Proceedings of International Association for Mathematical Geology 1998 Annual Meeting (IAMG'98)*, ed. A. Buccianti Ischia, Italy, October 3-7, 1998, p.204-211, 1988.

[17] **STM**, www.spatialmodels.com

[18] Fabbri A.G. & Chung C.J., Training decision-makers in hazard spatial prediction and risk assessment: ideas, tools, strategies and challenges. *Disaster Management and Human Health Risk,* eds. K. Duncan & C. A. Brebbia, Southampton, WIT Press, p. 285-296, or WIT Transactions on the Built Environment, www.witpress.com, ISSN 1743-3509 (on-line) doi:10_2495/DMAN09025, 2009.

[19] Fabbri, A.G., Chung, C.F. & Dong-Ho Jang, 2004, A software approach to spatial predictions of natural hazards and consequent risks. *Risk Analysis IV,* ed. C.A. Brebbia, Southampton, Boston, WIT Press, p. 289-305, 2004.

[20] Chung, C.F., Fabbri, A.G., Jang, D.H. & Scholten, H.J., Risk assessment using spatial prediction model for natural disaster preparedness. *Geo-information for Disaster Management*, eds. P. van Oosterom, S. Zlatanova & E.M. Fendel, Berlin, Springer, pp. 619-640. *Procs. of Gi4DM, The First Symposium on Geo-information for Disaster Management*, Delft, Netherlands, March 21-23, 2005.