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

Learning conditional probabilities in event bushes with temporal labels

S. I. Nikolenko [a], C. A. Pshenichny [b], R. Carniel[c]

[a]*Steklov Mathematical Institute, St. Petersburg, Russia (sergey@logic.pdmi.ras.ru)*

[b] *VNII "Okeangeologia", St. Petersburg, Russia (pshenichny@yandex.ru)*

[c] *University of Udine, Udine, Italy (roberto.carniel@uniud.it)*
Departamento de Geofisica, UNAM / UNiversidad Autonoma de Mexico, D.F. Mexico

Abstract: The event bush is a recently developed formalism for knowledge representation optimized specifically for Earth science knowledge. It graphically represents logical and/or probabilistic dependencies between its nodes (events). Numerical data in geosciences often come in the form of time series: therefore, it is a very common situation to have tables with time-related numerical data for the variables which correspond to the events represented in an event bush. In this work, we present a way to use these numerical data to learn conditional probabilities that have to be specified in the intermediate event bush created on the basis of an event bush with temporal labels. The same idea works also if the bush also has spatial labels.

Keywords: event bush; time series; probabilistic inference; risk assessment

1 INTRODUCTION

The event bush is a newly developed formalism for knowledge representation optimized specifically for Earth science knowledge. It was first proposed by Pshenichny [2003] as a new approach to probabilistic knowledge representation. Event bushes underlie Bayesian Belief Networks (BBN) and perform probabilistic inference via reducing to a BBN, see Pshenichny et al. [2008] for more details.

Latest research has shown that the event bush formalism is extremely viable; we have developed ways to incorporate temporal and spatial data immediately into an event bush and have begun developing learning algorithms for event bushes based on time series data. Thus the event bush becomes an information technology linking time series and geodata, or, to put it more explicitly, bringing time series data into the geoinformation system format. The novelty of this approach is semantic and conceptual strictness, which is a virtue of event bushes. In the present paper we continue this work towards automatic learning of probabilities in event bushes.

For instance, a volcanic environment in mountainous area in a temperate climatic zone implies ice caps topping mountains including active volcanoes, deep valleys hosting rivers, and forested mountain slopes. Even slight warming under the volcano, or a regional earthquake may trigger ice avalanches turning into lahars downslope, which become yet more devastating as they capture tree trunks. Lahars may pond rivers and cause floods, which, in turn, evoke other dangerous consequences. Such processes are successfully modeled by event bush, and all of them can be attributed spatial or temporal values, or both.

Let us assume that the whole mapped area is divided into finite number of regions (e.g., by principles of geomorphology), so that each node of event bush is attributed a subset of regions on a map.

Inversely, for any region an event bush may be plotted that represents either a full, or, commonly, a reduced version of the general bush.

Importantly, (i) a process that originates in one region (e.g., heat flow in the summit area of the volcano) may need to last long enough to cause another process in the same place (ice melting). Otherwise, (ii) a lahar should travel long enough from its source area to capture tree trunks. Finally, (iii) a flood should spread well upstream the gullies and last for quite a time to cause slumping of gully sides. Obviously, we have "shifts" of cause-effect relationship (i) in time, (ii) in space, and (iii) both. These can be assessed from appropriate physical models or from observations, if only feasible. In the latter case we have time series or geostatistical datasets. Both can, and, ideally, should, be used for learning conditional probabilities. For the sake of simplicity, we consider learning based on one-dimensional data, i.e., the time series, but take into account that these data are regionalized.

Let us describe how we proceed from the initial structure of event bush through a time-and-space-labeled bush to a BBN. First, our algorithm creates copies of the event bush structure for each region. Then, we need to estimate whether an event X in region R influences event Y in region Q. By a simple learning algorithm, we can find the time delay (or a distribution over a set of possible delays) between X and Y in each region. Thus, we can complete the construction of an intermediate event bush (ready to convert into a BBN) inside each region.

In this paper, we concentrate on the issues of temporal modeling. In Section 2, we review the basics of event bush theory, namely the method to convert event bushes to BBNs and the basics of temporal data processing in event bushes. Section 2 should not be confused with a thorough introduction of the event bush formalism. Such an introduction may be found in Pshenichny et al. [2008] and even right here, under the same cover, in Pshenichny et al. [2007a]. We also do not deal with the question of how to build the structure of an event bush or a Bayesian Belief network. This can be done by either expert elicitation or machine learning, see e.g. Jensen [2002], Neapolitan [2004], Tulupyev et al. [2006].

Instead, in the present paper we dive into the details of a question on using time series data for the event bush formalism. Section 3 contains the new results introduced in this paper: how to use time series data (often available in geosciences) for learning probabilities. Section 4 describes a particular example of applying our methods based on the data recorded at the Stromboli volcano. Finally, Section 5 concludes the paper with a discussion of further work possibilities.

2 Mathematical aspects of event bushes

2.1 Formal aspects of event bushes

Formally speaking, an event bush is a directed acyclic multigraph $B = \{V, E, S, M\}$, where

- V is the set of events (nodes);
- $E \subseteq 2^V$ is the set of edges;
- S is the set of operations (in the current case, logical connectives $\wedge$ and $\vee$);
- $M : E \to S$ is the set of labels on edges (in our case, they represent probabilities).

In the current event bush architecture there are two possible types of multiedges: $\vee$–edges with one source node and any number of target nodes and $\wedge$–edges with two source nodes and two target nodes. The semantics for an $\vee$–edge is "if source then one of the targets"; the semantics for an $\wedge$–edge is "if $\mathrm{src}_1 \wedge \mathrm{src}_2$ then tgt_1; if $\mathrm{src}_1 \wedge \neg\mathrm{src}_2$ then tgt_2". The semantics and other logical aspects of this construction, including its detailed motivation, can be found in Pshenichny et al. [2008].

As shown above, an event bush can have labels on edges. In this case, we consider probability labels that define conditional probabilities of the target nodes conditional on the source nodes. This gives rise to a general task of Bayesian inference on an event bush. This task (which is currently the primary numerical task solved by the event bush formalism) is solved by reducing an event bush to a BBN which is equivalent to the event bush, see Pshenichny et al. [2008] for more details.

The next developments which are absolutely crucial to make event bushes a success for geosciences would be adding time and space to the bushes. Once we have done that, it will be straightforward to incorporate the event bush technology into existing geoinformatic systems.

2.2 Temporal data in event bushes

It often happens that the logical structure of the event flow is not complete unless time is taken into account. Geological data are commonly related to time: to know *when* is no less important than to know *what*.

As shown above, the event bush is a qualitative tool of organization of a domain of knowledge. Together with the BBN formalism it forms a powerful tool for geohazard assessment. Thus, it it absolutely necessary to incorporate time into the structure of an event bush. BBNs have several known approaches to time handling. These can be divided into two main classes: discrete time (where it usually comes down to copying the nodes of the network several times corresponding to the time slices) and continuous time. For the time being, we do not touch continuous time handling in event bushes; but how do we account for discrete time?

This problem has been already addressed by Pshenichny and Nikolenko [2007]; Pshenichny et al. [2007b]. In this section we remind our progress and continue from there. We describe here the most basic approach to handling time-related information in an event bush and transforming such event bushes into BBNs and back. We take the most common (and sufficient for most practical purposes) approach of discrete time handling. That is, we divide the time scale into discrete periods with lengths known in advance. Each primary or secondary event in an event bush may have a time interval when it occurs (e.g. “magma ascent takes place from day 2 to day 5”, where time is measured from some arbitrary point when the modeled processes begin, see also Jaquet and Carniel [2003]). If it is not specified, we take that the event occurs throughout the whole time scale (e.g. it is a landscape peculiarity).

Apart from the time interval, there are two other questions that should be answered about each of the nodes of the event bush in order to incorporate time properly. First, how fast does this event influence its successors? E.g. if magma ascent leads to lava doming, does it occur immediately, in an hour, or in two hours? There is room for intermediate options here: magma ascent may lead to lava dome in one hour with probability 0.4 and in two hours with probability 0.6. This can also be easily accounted for in the corresponding BBN (in the next section, we will describe this construction in detail and show how one can learn the delays from data given as time series).

Second, for some events it may be impossible to say when they end (we detect that magma ascent has begun, but it has not yet ended). In these cases it should be noted how an event occurring now influences the probability of the same event occurring in the next time interval. E.g. if magma is ascending, how probable is it that it will be ascending in the next hour? Usually answers to these questions depend also on the duration of an event; this can also be taken into consideration.

The simplest approach to both these problems would be to ask the user for missing data. If a user can input the duration of an event if it occurs or the time of its beginning and how probable it is for it to continue for different durations, this is all we need to completely specify the BBN structure. In the next section, we describe a more automated approach that can learn cause-effect delays from data.

The event bush "loaded" with time values is translated then into a BBN (where the actual propagation will take place). We are given the value of a discrete time step (e.g. an hour) and information described above about the nodes of the event bush. In order to build a time-specific BBN, we first build a regular BBN on this event bush, discarding for the moment all time-related information (note that we also discard cause-and-effect relationships that do not take place immediately). Afterwards, we copy this BBN n times, where n is the number of time periods under consideration (if it is undefined, we may take the maximum of all possible durations or just set a reasonable upper bound for our modeling).

In the resulting BBN, we keep the structure of the initial BBN but add edges corresponding to time-related relationships. Specifically, we add directed edges between time-related causes and effects. Of course, the edges should be directed from the earlier event to the later. The additional edges may connect similar events with different time ("if magma ascends at time x, it will with probability 80% ascend at time $x + 1$") or different time-separated events ("if magma ascends at time x, with probability 80% there will be a lava dome or a lava flow at time $x + 1$"). The BBN in which actual computation takes place is transformed back into an event bush to display results to the user.

Time-specific event bushes will allow to adequately capture monitoring data, fully incorporate abundant physical models of eruptive scenarios, seismic unrest, flow propagation, and others, and will be tested on a number of various hazardous objects (volcanoes first of all), for which good time-series exist.

3 Learning temporal information from time series data

3.1 Time series as input

In the previous sections we came to the conclusion that implementing temporal information in expert systems requires some additional data. In particular, apart from the logical structure of the event bush, one is required to specify how events represented in the bush relate to each other with respect to time (i.e. how long is the delay between cause and effect). In this section we offer a new technique that will allow to answer this question automatically.

It is common for geosciences to have input data represented as time series. Consider Fig. 1. On the left we have a series of observations of the value of X, while on the right observations of Y are depicted. Suppose that in our event bush logical structure there is an arc from X to Y (meaning that X affects Y). By looking at these graphs, it is fairly obvious that Y indeed happens after X, and a sudden increase in X leads to an increase in Y after about three units of time. Thus, we can enter the delay quantity as 3, and the intermediate event bush will have arcs from "X at time t" to "Y at time $t + 3$".

Naturally, if we have time series as data, we can try to extract the delays automatically. We distinguish two major approaches to this, both based on bayesian estimation, but leading to different structures of the intermediate bush graphs.

3.2 Detecting the maximum likelihood delay

One approach is to fix the delays for every cause-effect pair. We would like to estimate which delay is most probable (most likely) for a given pair of time series. The simplest algorithm would do the following.

1. Fix a reasonable upper bound on the delay.

2. For every possible value of the delay t find the correlation between the effect Y and the

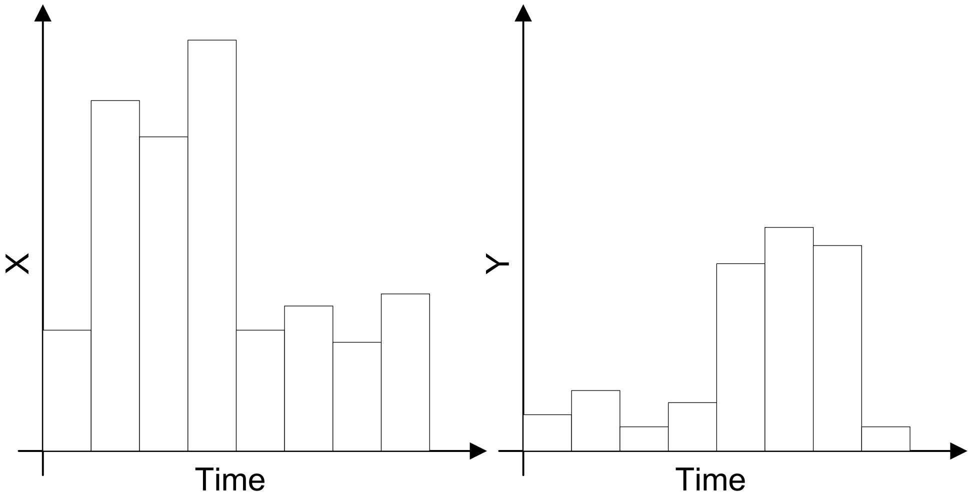

Figure 1: Uniquely determined delays.

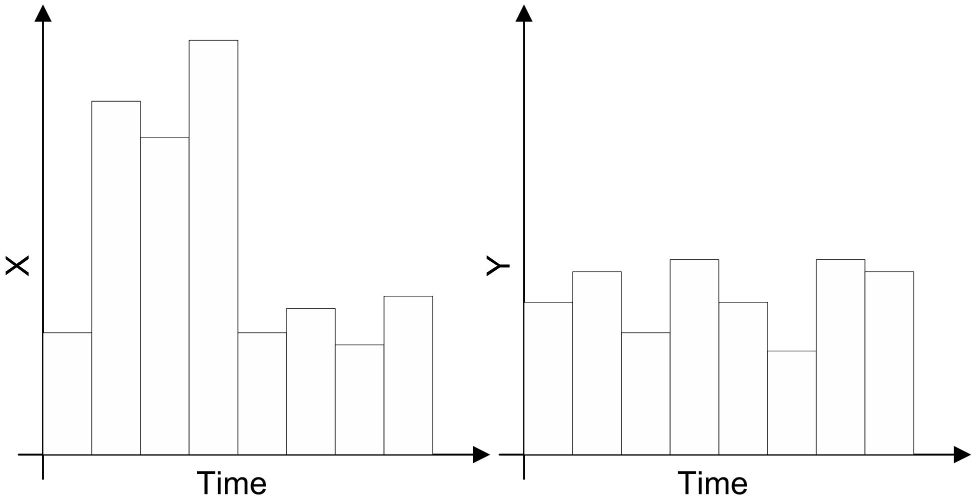

Figure 2: No reasonable correlation.

cause X shifted by t. We can use the Pearson correlation coefficient:

$$r_t = \frac{\sum (x_{i+t} - \bar{x})(y_i - \bar{y})}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}. \tag{1}$$

3. Choose the delay with maximal correlation:

$$d = \mathrm{argmax}_t r_t. \tag{2}$$

This simple algorithm allows to find the delay which maximizes correlation between the time series of the effect and the shifted cause. Therefore, the resulting delay value is most likely given these time series.

In the case when the delay value is clear and unique, like on Figure 1, this algorithm works fine. However, in other situations it has serious disadvantages.

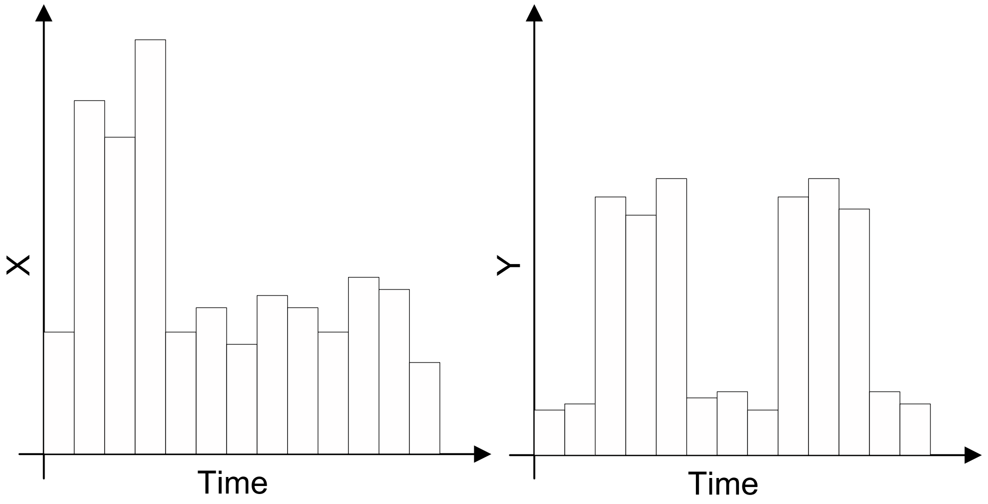

Figure 3: Two equally possible delays.

For example, on Fig. 2 there appears to be simply no reasonable correlation of the two time series (at least during the time span shown). In this case we would like our system to warn the user that these two nodes do not appear to be connected, while the above algorithm will just choose the most correlated delay (depending probably on the noise in data). However, this problem is easy to overcome: we can simply set a default threshold and on step 3 of the algorithm output d only if it exceeds this threshold.

But there are other problems. Look at Fig. 3. There are two plausible delays here: one time unit and seven time units. The algorithm above would choose one of them based on some miniscule correlation advantage which would most probably be accounted for by the noise in the data. In fact, we would like the system to accept the fact that it doesnt know what the delay really is. Of course, we could (and in some situations should) ask the user to choose one of the possibilities. But if the user does not have enough information to perform the choice, we should look for alternative ways.

3.3 Probability labels in the intermediate graph

Fortunately, our formalism is rich enough to handle this situation in the most natural way — consider all delays as possible and range the probabilities of these delays with respect to their correlations:

$$p_t = \frac{r_t}{\sum_t r_t}. \tag{3}$$

Let us make the assumption that there is only one value of the delay (that is, a single event cannot trigger another event twice). In this case, we can view the events “Y at time t” as mutually incompatible. On the level of the BBN this means that we can view Y as a node with T states corresponding to T statements: “Y at time 1”, …, “Y at time T”. And we already know the probabilities we would like to assign:

$$p(Y \text{ at time } t_0 + t \mid X \text{ at time } t) = p_t = \frac{r_t}{\sum_t r_t}. \tag{4}$$

This BBN is shown on the left of Fig. 4. The right part of the same figure depicts the construction in the case when events are not instantaneous, but can last for several time intervals.

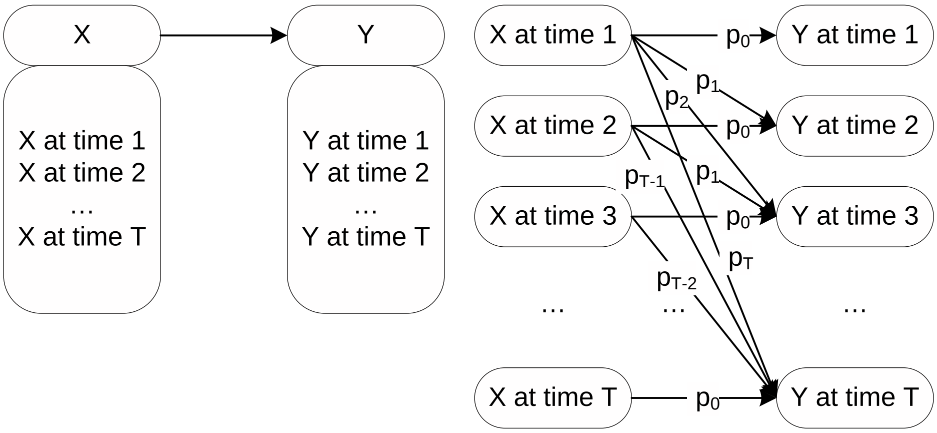

Figure 4: Intermediate event bush.

There is a natural modification of this algorithm if we do not want to have so many edges in the intermediate event bush. We can set a threshold of the correlation and assign nonzero probabilities only to the delays that exceed this threshold (assigning probabilities proportional to the correlations, as above). This way there will be much fewer edges in the intermediate graph with no substantial loss in representativeness.

3.4 Using variograms to estimate time intervals

Learning conditional probabilities is only a part of the job with specifying time labels on an event bush. Another important part would be to specify which time intervals to consider and how to break them up into smaller intervals. We propose a way to learn these attributes automatically from variograms of the input time series data.

Let us define for a stochastic process $V(t)$ its *variogram* as the following expression:

$$\gamma(\tau) = \frac{1}{2}\mathrm{E}\left[(V(t+\tau) - V(t))^2\right]. \tag{5}$$

Informally, a variogram, first applied to the estimation of volcanic hazard by Jaquet et al. [2000], captures the memory of the stochastic process V; it measures the correlation of V with itself shifted in time. If V has no memory of its past, $\gamma(\tau)$ will be simply equal to the variance of V for all $\tau > 0$. However, if V does depend on its own history, γ will gradually rise from 0 (note that $\gamma(0) = 0$) to the variance; as soon as $\gamma(\tau)$ stabilizes around the variance of V, this means that the process has effectively forgotten all about the zero point. For a much more detailed discussion of variograms see Carniel et al. [2008] and references therein.

For example, Carniel et al. [2008] depict the memory of the seismic noise recorded in March, 2005, on Tenerife, close to Las Cañadas caldera. The variogram reaches the variance after about 80 hours and remains at about the same level afterwards. This means that the process causing this seismic noise depended only on the history of its last 80 hours.

How does all this help us create event bushes? When we define the time intervals for nodes in event bushes, it would only make sense to define it as long as previous history may affect the current events. For example, we could go as far as 200 hours on the seismic noise, but since the autocorrelation of this noise reaches zero after 80 hours, there would be no meaningful links from the noise that far ago to the present time. Thus, for seismic noise it would suffice to consider a scale of 80 hours.

Note, however, that the overall time interval depends on all nodes in the bush. Perhaps there is

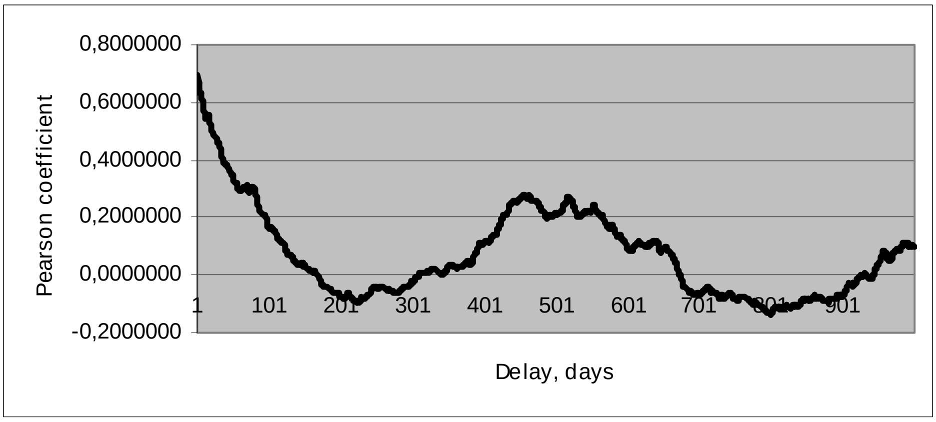

Figure 5: Correlation between events per day and tremor as a function of the delay in data.

some other event affected by seismic noise that has a longer memory (say, 200 hours). In this case, we would have to consider seismic noise from 200 hours ago because it might affect the current situation *via* this event. In general, however, calculating variograms provides us with an easy automatic way to estimate time scales of different events, and its use can be extended to inter-event dependence by the so called cross variogram (see e.g. Jaquet and Carniel [2003], Matheron [1962]).

4 An example

As an example of our approach we took the data from Stromboli volcano (Italy) that has been thoroughly studied in Jaquet and Carniel [2001], Jaquet and Carniel [2003], Carniel et al. [2008]. These studies have shown that Stromboli exhibits an extraordinary long memory of its previous activity. Variogram analysis performed by Jaquet and Carniel [2001] showed that the memory could last longer than 400 days.

For our study we took the records collected from 1992 till 2001 of the number of volcanic event per day registered at Stromboli together with the average tremor intensity recorded on that day. The graph shown on Fig. 5 depicts the Pearson coefficient between these data sets as a function of the delay between the two measured variables.

The graph shows that the memory of these two mutually dependent datasets is probably limited in this case to about 150 days (this is when the correlation first reaches zero). However, one would lose much of the structure if he would have to limit the model to immediate consequences, discarding the delayed correlations at all. Thus, in this case it would be sensible to set a minimal threshold of the correlation (say, 0.2) and relate a node corresponding to tremor at time t to nodes corresponding to the number of events up to time $t + \Delta t$, where Δt is the period of time after which the Pearson coefficient reaches the threshold for the first time.

As for computational issues, if an event bush has a relatively simple structure, then multiplying it by 150 would not be unfeasible. However, if the bush has many nodes, the direct day–by–day copying method could fail. In this case, we suggest merging neighboring intermediate nodes (that is, enlarging the time unit). However, in the areas where the correlation is high the time unit may be split up further. For example, in this case we could use day–by–day nodes up until about day 50. After that, we could average the data over 5 or even 10 days without losing too much of the expressive power (and common sense).

5 Conclusion and further work

We have presented some basic ideas on how to use time series data, very common in geosciences, for learning probabilities in event bushes with temporal and spatial labels. Note that this is the first implementation of an event bush mechanism where we actually learn probabilities in the Bayesian network from data, rather than assign them on the basis of logical considerations.

Obviously, open questions remain. The current implementation of the event bush technology leaves much for the user to specify concerning temporal and spatial labels. First, the spatial issues still remain largely uncovered by the current formalism. Space-bound data once again brings us to learning issues: an event bush should be able to automatically benefit from spatial data just like it does from time series. Hopefully, some of the learning could follow the lines of this paper, if an event bush is meant to be used for spatial and temporal modeling simultaneously, and we are provided with space-related time series. Other types of data (especially data of qualitative nature), however, may require additional insights.

In general, we plan to further automatize the creation of event bushes and, ultimately, embed a working prototype of the formalism into some existing GIS framework. This will bring us to real-life applications of the event bush technology.

Acknowledgements

The research has been funded by Project V4 of Istituto Nazionale di Geofisica e Vulcanologia, Italy, “Development of quantitative tools to evaluate volcanic hazard”, a Human Capital Foundation grant, Information Technologies cluster project, 2006, and NATO-Russia collaborative linkage grant 982484.

The authors thank Oksana Kanzheleva for fruitful discussions and the anonymous referees for numerous useful suggestions.

References

Carniel, R., O. Jaquet, and M. Tárraga. Perspectives on the application of the geostatistical approach to volcano forecasting at different time scales. In *Caldera volcanism: Analysis, modelling and response, Developments in Volcanology*. Elsevier, 2008.

Jaquet, O., S. Low, B. Martinelli, V. Dietrich, and D. Gilby. Estimation of volcanic hazards based on cox stochastic processes. *Physics and Chemistry of the Earth, Part A: Solid Earth and Geodesy*, 25(6–7):571–579, 2000.

Jaquet, O. and R. Carniel. Stochastic modelling at stromboli: a volcano with remarkable memory. *Journal of Volcanology and Geothermal Research*, 105:249–262, 2001.

Jaquet, O. and R. Carniel. Multivariate stochastic modelling: towards forecasts of paroxysmal phases at stromboli. *Journal of Volcanology and Geothermal Research*, 128(1–3):261–271, 2003.

Jensen, F. V. *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag, 2002.

Matheron, G. *Traité de Géostatistique appliqueé. Tome* 1. Technip, Paris, 1962.

Neapolitan, R. E. *Learning Bayesian Networks*. Pearson Prentice Hall, 2004.

Pshenichny, C. A. A draft for complex formal approach in geoscience: modeling geohazards. In *Proceedings of IAMG ’03*, 2003.

Pshenichny, C. A. and S. I. Nikolenko. Temporal assessment by means of an event bush. *Geophysical Research Abstracts*, 9(00497), 2007.

Pshenichny, C. A., S. I. Nikolenko, R. Carniel, A. L. Sobissevitch, P. A. Vaganov, Z. V. Khrabrykh, V. P. Moukhachov, V. L. Shterkhun, A. A. Rezyapkin, A. V. Yakovlev, R. A. Fedukov, and E. A. Gusev. The event bush as a potential complex methodology of conceptual modelling in the geosciences. In *Proceedings of the* 4*th Biennial Meeting of iEMSs*, 2007a.

Pshenichny, C. A., S. I. Nikolenko, R. Carniel, P. A. Vaganov, Z. V. Khrabrykh, V. P. Moukhachov, V. L. Akimova-Shterkhun, and A. A. Rezyapkin. The event bush as a semantic-based numerical approach to natural hazard assessment (exemplified by volcanology). *Computers and Geosciences*, to appear, 2008.

Pshenichny, C. A., S. I. Nikolenko, A. L. Sobissevitch, and A. V. Yakovlev. Spatial volcanic hazard assessment by the event bush method. In *Proceedings of the XXIV IUGG General Assembly*, 2007b.

Tulupyev, A. L., S. I. Nikolenko, and A. V. Sirotkin. *Bayesian Networks: A Probabilistic Logic Approach*. St. Petersburg, Nauka, 2006.