Live Mapping, Data Acquisition and Decision Making

S. de Bruin[1], A. K. Bregt[2]

Wageningen University, P.O. Box 47, 6700AA Wageningen, The Netherlands
Email: {sytze.debruin[1]; arnold.bregt[2]}@wur.nl

1. Introduction

Recent disasters such as the Fukushima nuclear accident have shown that spatially distributed sensor data are an important source of information for citizens (Hoetzlein 2012) and professional decision makers. Often, the information content of sensor data changes with time (Heuvelink and Griffith 2010) and it depends on the observed quantities (Bhattacharjya et al. 2010) and hence on the locations where measurements are made. *Human decision making* about where and when to measure and which places to avoid is expected to benefit from real-time mapping and live feedback of up-to-date information (Hiemstra et al. 2009). On the other hand, the Expected Value Of Information (EVOI) has been proposed as a tool for *automated* selection of new measurement locations (de Bruin et al. 2011). Both approaches require timely integration of available data and prompt feedback when new information becomes available.

The purpose of this paper is to discuss needs for methods that support automated and human decision making about sequentially added new sample locations based on information obtained from the previous samples. We use findings from students mapping (1) an invasive species in a natural park, (2) a fictitious moving toxic plume over Wageningen campus and (3) a simulated example of automated mobile sensors exploring a contaminated environment.

2. Cases

2.1 Invasive species

Students used smartphones for mapping *Molinia caerulea* in a park in the east of the Netherlands It was assumed that the user of the map wanted a product in which misclassification errors are minimised. False negatives were assumed twice as expensive as false positives. The data acquired were instantaneously interpolated by indicator kriging using a postulated spherical variogram. Computations at server side were implemented in Python and R (R Development Core Team 2011) using the gstat library (Pebesma 2004); the client software was a standard web browser supporting the W3C Geolocation API. All student groups had immediate access to the most recent map. They used the map for deciding where to observe next, given the objective of the map.

2.2 Toxic plume

A toxic plume rotating around its origin with an amplitude of 0.1π radians, a period of 240 minutes and perturbed by an additive spatio-temporally correlated Gaussian random field was simulated and stored as a spatio-temporal grid on a server. Information about the plume was not disclosed to students. They were instructed to use smartphones for collecting data upon which decisions would be based to evacuate Wageningen campus. During the fieldwork, toxin concentrations were extracted from the grid using the location of the observer. Results were returned to the observer, both as an alpha-numeric value and on a map showing up to 30 recent observations using colour for the concentration and symbol size for the age of the observation. The map was accessible to all students. Live mapping involved no interpolation; students interpolated the data on the second day of the practical as an exercise.

2.3 Automated site selection

The automated site selection case is based on de Bruin et al. (2011). We assume spatially correlated positive/negative data, with the costs of false negatives or "safe" decisions being higher than those of false positives. EVOI is estimated as the difference between expected misclassification costs at the present stage of knowledge and expected costs when new information becomes available. Figure 1 shows a tree with square nodes indicating decisions to place a sensor for measuring the phenomenon at some location and decisions about mapping presence or absence of the phenomenon using the information at hand. Chance nodes (circles) indicate the outcome of random events once a decision has been taken. The chance nodes at the at the second level of the tree (signal / no signal) indicate that the method can deal with measurement error.

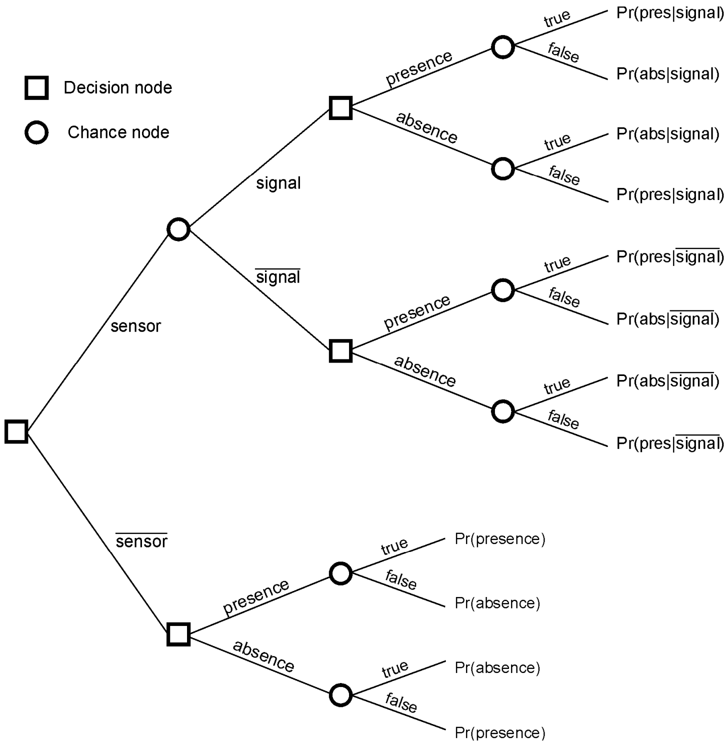

Figure 1. Decision tree for sensor placement and mapping presence or absence of a phenomenon.

Assuming rational decision making, a measurement makes sense if the difference between the expected loss of the lower branch and that of the upper branch of Figure 1 is larger than the cost of the measurement. In case correct classifications involve no costs or benefits, the expectation of the lower branch (1) is calculated as:

$$E(cost_{lower}) = \min(cost_{false\_positive} \times \Pr(absent), cost_{false\_negative} \times \Pr(present)) \quad (1)$$

where the function min(.) returns the minimum of its arguments, $\Pr(absent)$ and $\Pr(present)$ are the prior probabilities of absence and presence, respectively and $cost_{false\_negative}$ and $cost_{false\_positive}$ are the costs of misclassification. The expected cost of the upper branch (2) is calculated by:

$$\begin{aligned} E(cost_{upper}) = {} & \Pr(signal) \times \min(cost_{false\_positive} \times \Pr(absent \mid signal), cost_{false\_negative} \times \Pr(present \mid signal)) \; + \\ & \Pr(\overline{signal}) \times \min(cost_{false\_positive} \times \Pr(absent \mid \overline{signal}), cost_{false\_negative} \times \Pr(present \mid \overline{signal})) \end{aligned} \quad (2)$$

where $\Pr(signal)$ is the prior probability that a warning is issued, $\Pr(absent \mid signal)$ is the probability that the phenomenon is present given a warning, etc. The probabilities are computed from prior data (we used a sample of 16 sites regularly spread over the area) and sensor specifications.

We consider the *aggregated* expected costs of misclassification over the study area and find a single optimal sample location as the one that maximises EVOI and thus minimises $E(cost_{upper})$. The aggregated costs of misclassification were computed by creating maps for both a signal and no signal obtained at the sensor location and multiplying the expected costs for these situations with the probability of their occurrence. If the locations of two or more observations are to be simultaneously optimised, complexity of the computations increases, since nearby observations are typically conditionally dependent. At the same time the size of the solution space increases. For example, with two simultaneous observations, four expected cost maps and their probabilities need to be computed for each pair of locations while solution space increases by a factor (n-1), with n being the number of potential sample locations. We handled the latter situation using a genetic algorithm, while spatial correlation was modelled by a given variogram.

3. Results and discussion

3.1 Invasive species

Apart from some technical issues owing to poor GPS receipt, the field work was successful. Students appreciated the instantaneous integration of survey data and they were able to identify locations where additional data should be collected. The interpolation procedure caused some problems, however. The stationarity assumption of the interpolator was violated since sub-regions differed in the degree and variability of infestation. This was not shown on the maps. Also the unknown mean posed difficulties. Incorporation of prior knowledge into the interpolator (e.g. regression kriging) could have circumvented this, but the interpolation procedure could not be adapted on the fly. Other points of attention are the problem of obtaining the variogram and the lack of a theoretical foundation for indicator kriging.

3.2 Toxic plume

There were no technical problems during data acquisition. Almost immediately one group encountered the source of the plume and —informed by the local peak— several groups succeeded in delineating the dynamic plume along the intervention level (see Figure 2). In doing so they were somewhat disappointed by the fact that the plume did not respond to temporary changes in wind conditions.

Spatio-temporal interpolation on the second day of the exercise employed a simple interpolator, since students lacked knowledge of spatio-temporal geostatistics. Time was accounted for by interpolating over time intervals, which slowed down the *perceived* dynamics of the plume. If spatio-temporal geostatistics were applied there would (again) have been a problem of modelling the variogram.

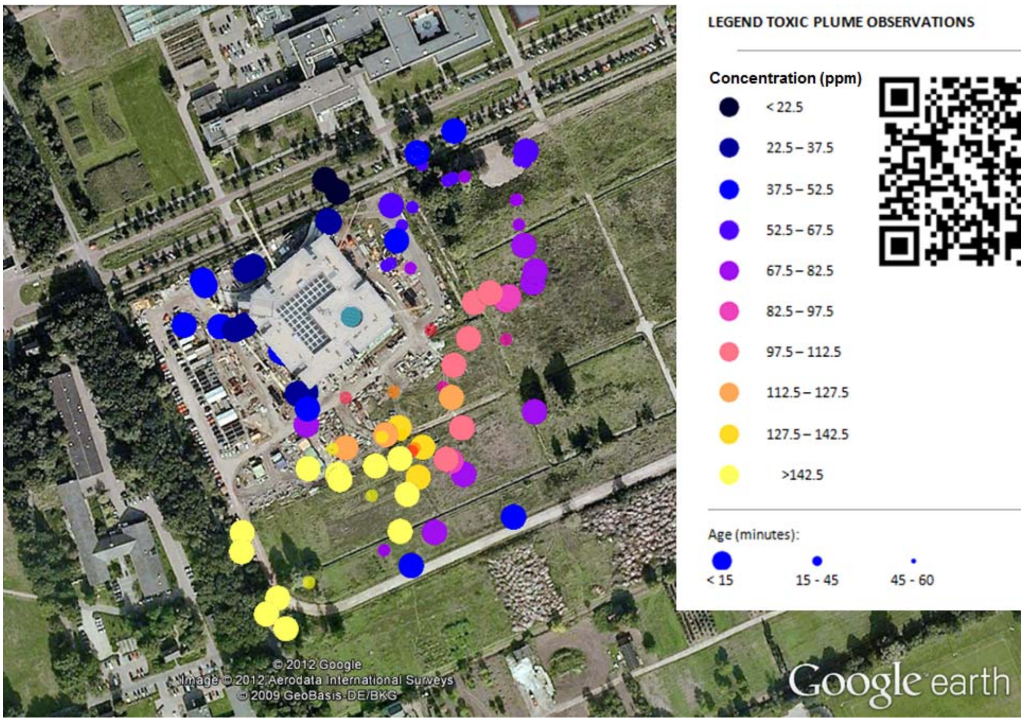

Figure 2. Hundred most recent measurements at 15.00, February 24, 2012. Background: “Wageningen Campus.” 51°59'06.06" N, 5°39'47.85" E. **Google Earth**. January 1, 2005.

3.3 Automated site selection

The Expected Value Of Information approach placed new observations at locations that intuitively make sense. Constraining potential sample locations to the space that could be travelled by a small set of mobile sensors caused the sensors to get locally trapped so that they failed to visit more informative spots. This was avoided by first determining globally optimal locations and next deciding which sensors to move. Particularly the latter problem is computationally demanding and requires some heuristic optimiser such as genetic algorithms. Furthermore, the method requires spatio-temporal interpolation. Since data transition zones are preferentially sampled, variogram modelling may be challenging.

4. Conclusions

Today, enabling technology such as smartphones, network infrastructures and protocols can be considered commonplace. Yet, fast processing of spatial data using suitable methodology continues to be a challenge, even more so when used within the context of adaptive data acquisition and spatio-temporal interpolation (Pebesma et al. 2011). Automated adaptive data acquisition requires global rather than local data, otherwise sensors get trapped in local optima.

References

Bhattacharjya D, Eidsvik J and Mukerji T, 2010. The value of information in spatial decision making. *Mathematical Geosciences* 42(2):141-163.

De Bruin S, Ballari D and Bregt AK, 2011, Multiphase sampling using expected value of information. In: Fonte CC, Goncalves L and Goncalves G (eds), *7th International Symposium on Spatial Data Quality, Raising awareness of Spatial Data Quality*. Coimbra, Portugal, 65-70.

Heuvelink, GBM and Griffith DA, 2010, Space-time geostatistics for geography: A case study of radiation monitoring across parts of Germany. *Geographical Analysis*, 42 (2):161-179.

Hiemstra PH, Pebesma EJ, Twenhöfel, CJW and Heuvelink GBM, 2009, Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Computers and Geosciences*, 35(8):1711-1721.

Hoetzlein RC, 2012, Visual communication in times of crisis: The Fukushima nuclear accident. *Leonardo* 45(2):113-118.

Pebesma EJ, 2004, Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30(7):683-691.

Pebesma EJ, Cornford D, Dubois G, Heuvelink GBM, Hristopulos D, Pilz J, Stöhlker U, Morin G and Skøien JO, 2011, INTAMAP: The design and implementation of an interoperable automated interpolation web service. *Computers and Geosciences*, 37(3):343-352.

R Development Core Team, 2011, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria (http://www.R-project.org).