Multiphase sampling using expected value of information

Sytze de Bruin, Daniela Ballari & Arnold K. Bregt

Wageningen University, Laboratory of Geo-Information Science and Remote Sensing, P.O. Box 47, 6700 AA, The Netherlands
Sytze.deBruin@wur.nl, Daniela.Ballari@wur.nl, Arnold.Bregt@ wur.nl.

Abstract

This paper explores multiphase or infill sampling to reduce uncertainty after an initial sample has been taken and analysed to produce a map of the probability of some hazard. New observations are iteratively added by maximising the global expected value of information of the points. This is equivalent to minimisation of global misclassification costs. The method accounts for measurement error and different costs of type I and type II errors. Constraints imposed by a mobile sensor web can be accommodated using cost distances rather than Euclidean distances to decide which sensor moves to the next sample location. Calculations become demanding when multiple sensors move simultaneously. In that case, a genetic algorithm can be used to find sets of suitable new measurement locations. The method was implemented using R software for statistical computing and contributed libraries and it is demonstrated using a synthetic data set.

Keywords: Iterative sampling, adaptive sampling, infill sampling, decision analysis, mobile sensors.

1 Introduction

After a major incident such as the recent fire in a chemical factory in Moerdijk (January 5, 2011), The Netherlands, authorities have to decide whether or not food produced in the vicinity of the accident is suitable for human consumption. Such decision making typically relies on information obtained from a small sample, but it may improve when non-covered regions are "filled in" by additional sampling (Johnson, 1996; Cox et al., 1997) by mobile sensors. The costs of misclassification in cases such as depicted above are often unequal for type I and type II errors, with the costs of false negatives or "safe" decisions being higher than those of false positives. Selection of new sample locations should therefore account for this difference. At the same time, the costs for visiting new sites may differ between mobile sensors located within the area. For example, sensors situated near a new sample location need less travelling.

The method described in this paper involves optimising new sample locations based on information obtained from the previous sample. The phenomenon to be mapped is considered static within the time frame of the analysis (e.g. surface contamination after an incident). Expected value of information (EVOI) is used for

quantifying the suitability of the sample. EVOI expresses the benefit expected from data collection prior to actually doing the measurements (De Bruin et al., 2001; de Bruin and Hunter, 2003; Back et al., 2007). In contrast to kriging variance (Baume et al.) and entropy based methods (Zidek et al., 2000), EVOI is data dependent and it can incorporate different misclassification costs for false positives and false negatives. Heuvelink et al. (2010) used a stochastic model of the environmental phenomenon and also accounted for differences between misclassification costs, but here a direct Bayesian approach is used that is potentially faster when few samples are added per iteration.

Our aim is to demonstrate and discuss some strategies for using EVOI to add observations to a previous sample while accounting for constraints imposed by a sensor network.

2 Methods

2.1 Expected value of information

EVOI is estimated as the difference between expected 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. For example, if a sensor is placed, measurement with it may indicate presence (*signal*) or absence ($\overline{signal}$) of the phenomenon. The probability of obtaining a sensor signal at some location, Pr(*signal*) can be computed from sensor properties and the prior probability of presence, Pr(*present*), as follows (1):

$$\begin{aligned}\Pr(signal) = \Pr(signal \mid present) \times \Pr(present) + \\ \Pr(signal \mid absent) \times \Pr(absent)\end{aligned} \quad (1)$$

where $\Pr(signal \mid present)$ is the probability that a warning is issued if the phenomenon is present and $\Pr(\overline{signal} \mid absent) = 1 - \Pr(signal \mid absent)$ is the probability that the sensor correctly gives no signal. These probabilities are given in the sensor specifications (i.e. sensitivity and specificity).

Decision making is assumed to be based on Bayes actions, i.e. minimising expected loss. Accordingly, placing a sensor is sensible if the expected loss of the upper branch of Figure 1 is lower than the expected loss of the lower branch. If only misclassifications involve costs, the latter is calculated as (2):

$$E(cost_{lower}) = \min(cost_{false\_positive} \times \Pr(absent);\ cost_{false\_negative} \times \Pr(present)) \quad (2)$$

where min(.) is a function returning the minimum of its arguments and $cost_{false\_negative}$ and $cost_{false\_positive}$ are costs of misclassification. The conditional probabilities shown in Figure 1 are calculated with Bayes' rule, e.g (3):

$$\Pr(absent \mid signal) = \frac{\Pr(signal \mid absent) \times \Pr(absent)}{\Pr(signal)} \quad (3)$$

Hence, the expected cost of the upper branch is calculated by (4):

$$\begin{aligned} E(cost_{upper}) = \Pr(signal) \times \min(&cost_{false\_positive} \times \Pr(absent \mid signal), \\ &cost_{false\_negative} \times \Pr(present \mid signal)) + \\ \Pr(\overline{signal}) \times \min(&cost_{false\_positive} \times \Pr(absent \mid \overline{signal}), \\ &cost_{false\_negative} \times \Pr(present \mid \overline{signal})) \end{aligned} \quad (4)$$

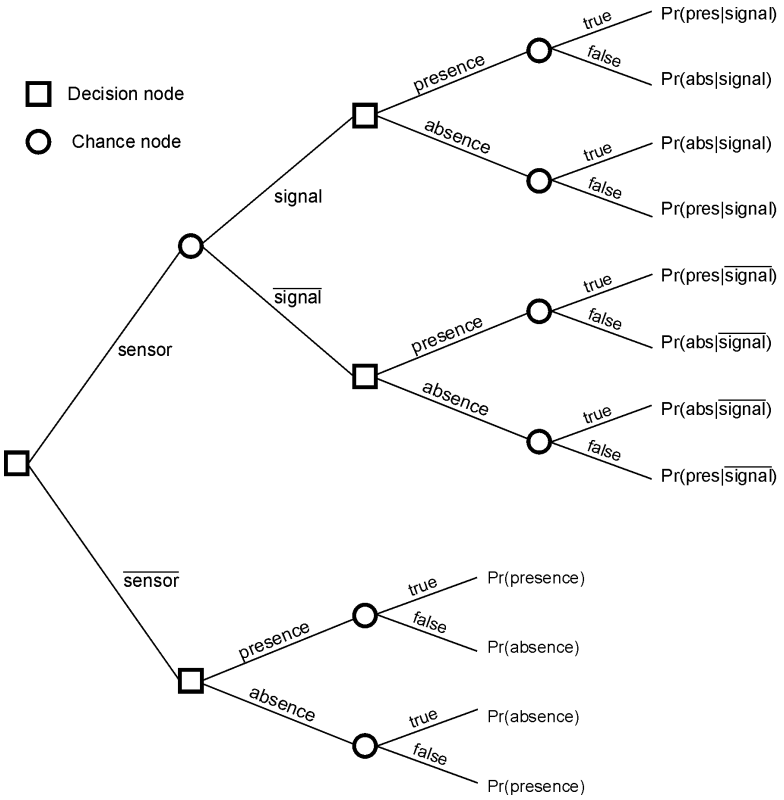

Figure 1. Decision tree showing decisions to place a sensor or not and to map presence or absence of a phenomenon (e.g. hazard).

EVOI is the difference between $E(cost_{lower})$ and $E(cost_{upper})$, where *lower* refers to the lower branch of the decision tree and *upper* to the upper branch. 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 are 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 substantially. 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. This situation was handled using a genetic algorithm.

2.2 Case study

A case study was conducted using a synthetic data set constructed by applying a threshold at 20 to a Gaussian random field of 100 x 100 grid cells of unit size with mean 20 nugget 1 and a spherical structural spatial correlation component with range 40 and a partial sill (semivariance) 16. Sensor data were obtained by sampling the synthetic data and adding random measurement error. The initial sample consisted of 16 points on a regular grid. Sensor data were interpolated using indicator kriging. Computations were done in R (Venables et al., 2010) using the geostatistical package *gstat* (Pebesma, 2004) and the genetic algorithm implemented in the package *genalg*.

Three approaches were considered for adding new sample locations to the original sample: (1) add single location at a time, move sensor with lowest cost (in this case Euclidean distance); (2) add two locations simultaneously and scan only the area that can be reached by the sensors within one time step; (3) add two sample locations simultaneously, scan the whole area, and move the sensors with lowest cost. The costs of misclassification were arbitrarily set at 2 and 3 (no unit) for false positives and false negative, respectively.

3 Results

Figure 2a shows probabilities of occurrence interpolated from the initial sample of 16 sites. Figure 2b shows the map of global EVOI, i.e., EVOI computed after aggregating expected misclassification costs for observations made at each grid location, separately. The best location thus corresponds to the highest global EVOI. Not surprisingly, this occurs between observations differing in value (indicated by arrow).

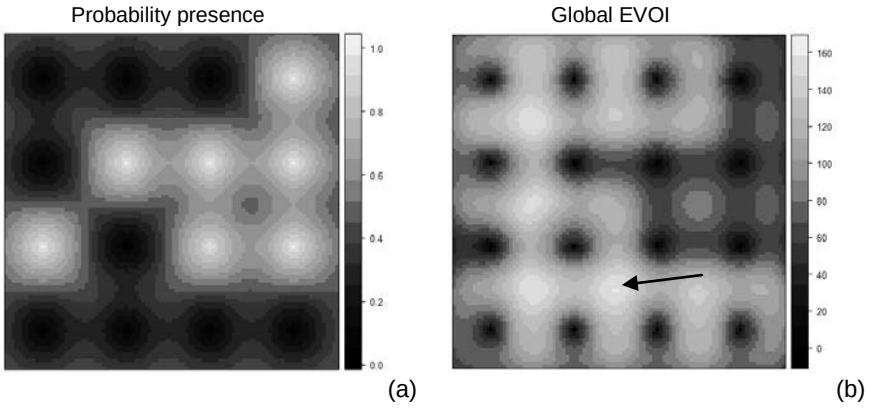

Figure 2. Probability map (a) and global EVOI (b) computed from the initial sample of 16 regularly spaced points. The arrow points to the location having highest global EVOI.

Figure 3 shows an example of an optimised sensor configuration after the 17th observation has been made (16 initial and 1 infill measurements) on a backdrop of the probability of presence of the phenomenon (cf. Figure 2). Euclidean distance was used for deciding which sensor to move to the next location, but another cost criterion could have been used with only minor modification of the algorithm.

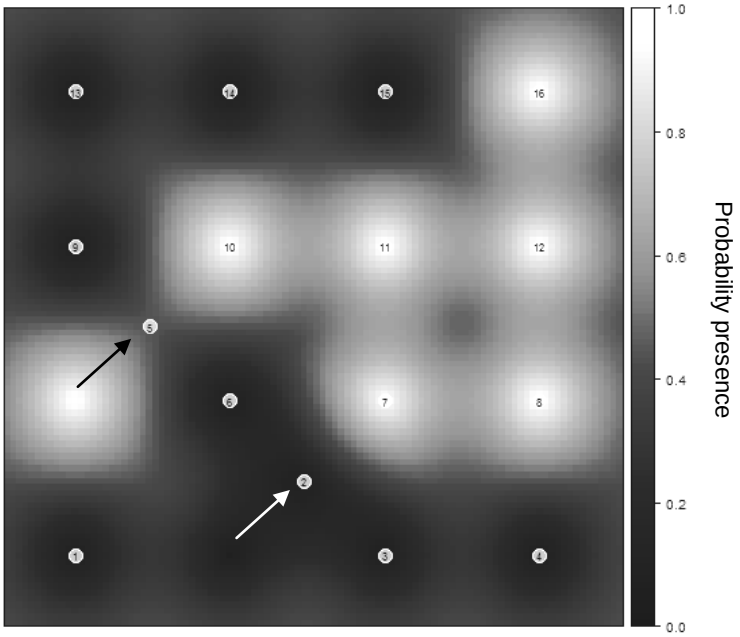

Figure 3. Configuration of initially regularly spaced sensors after two iterations with a single observation per step (approach 1). First sensor 2 moved (white arrow) and a measurement was taken, next sensor 5 moved (black arrow), but the measurement has not yet been taken.

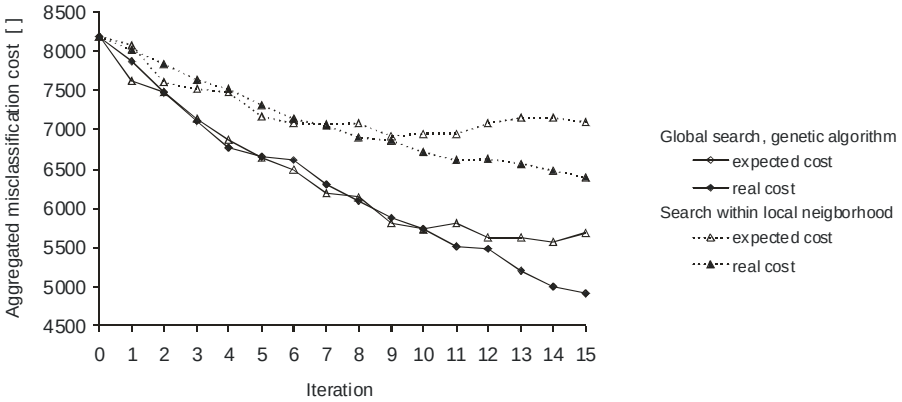

Figure 4. Effect of the way sensor constraints are taken into account on aggregated misclassification costs with two simultaneously moving sensors (approaches 2 and 3).

Figure 4 shows the effect of the two approaches to account for sensor constraints described in section 2.2, with two simultaneously moving sensors. Not surprisingly,

both expected and real misclassification costs were much lower when the full study area was scanned in search of the best sample locations. Of course, in the case of local sensor neighbourhood scanning, results depend on the start locations chosen. Large differences between real costs (normally not known) and expected costs are indicative of misspecification of the geostatistical model used for interpolating the probability map.

4 Conclusions

The Expected value of information (EVOI) approach puts new observations at locations that intuitively make sense and it can help deciding when to stop a survey. The method accounts for specified misclassification costs; these can be dissimilar for different kinds of errors (e.g. false positives or false negatives). Constraining potential sample locations to the space that can be travelled by a small set of mobile sensors is a bad idea since the sensors may get trapped in some area and may thus fail to visit potentially interesting spots. Rather, cost distances can be used for deciding which sensors to move to next globally optimal locations. Genetic algorithms may be useful for optimising the sample locations for multiple sensors moving simultaneously.

References

Back, P.E., Rosén, L., Norberg, T. (2007), "Value of information analysis in remedial investigations". *Ambio,* Vol.36: 486-493.

Baume, O.P., Gebhardt, A., Gebhardt, C., Heuvelink, G.B.M., Pilz, J. (2011), "Network optimization algorithms and scenarios in the context of automatic mapping", *Computers & Geosciences*, In Press, Corrected Proof.

Cox, D.D., Cox, L.H., Ensor, K.B. (1997), "Spatial sampling and the environment: some issues and directions". *Environmental and Ecological Statistics*, Vol. 4: 219-233.

De Bruin, S., Bregt, A., Van de Ven, M., (2001), "Assessing fitness for use: the expected value of spatial data sets", *International Journal of Geographical Information Science*, Vol.15: 457-471.

De Bruin, S., Hunter, G.J. (2003), "Making the trade-off between decision quality and information cost". *Photogrammetric Engineering and Remote Sensing* Vol. 69: 91-98.

Heuvelink, G.B.M., Jiang, Z., De Bruin, S., Twenhofel, C.J.W. (2010), "Optimization of mobile radioactivity monitoring networks". *International Journal of Geographical Information Science*, Vol. 24: 365-382.

Johnson, R.L. (1996), "A Bayesian/geostatistical approach to the design of adaptive sampling programs". *In:* Rouhani, S., Srivastava, R.M., Desbarats, A.J., Cromer, M.V., Johnson, A.I. (eds.). *Geostatistics for Environmental and Geotechnical Applications*, American Society for Testing and Materials**,** pp. 102-116.

Pebesma, E.J. (2004), "Multivariable geostatistics in S: the gstat package". *Computers & Geosciences*, Vol. 30: 683-691.

Venables, W.N., Smith, D.M., R Development Core Team (2010), *An Introduction to R*, The R Foundation for Statistical Computing, Vienna, Austria, 101p.

Zidek, J.V., Sun, W.M. Le, N.D. (2000), "Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields". *Journal of the Royal Statistical Society Series C-Applied Statistics*, Vol 49: 63-79.