# Conservative updating of sampling designs

**Kristina B. Helle*, Edzer Pebesma**
Institute for Geoinformatics
University of Münster
Münster, Germany
*kristina.helle@uni-muenster.de

*Abstract*—**When improving existing monitoring networks, to adapt to changed requirements, keeping as many stations as possible is cheapest and therefore often preferred over a completely new setup. Here, the sampling design for ambient gamma dose monitoring in Norway is optimised. We consider two goals: equal spread of stations, and detection of plumes that affect densely populated areas. For optimisation, we compare and improve algorithms that replace the existing stations one by one: a greedy algorithm replaces the most unimportant station by the best candidate station; random replacement keeps all random improvements. A new approach is random replacement that rejects all sampling designs with too many stations moved. We add a penalty term to the cost function to search sampling designs with few station moves. This combines the advantages of the two previous approaches: The greedy algorithm replaces the most unimportant stations only, therefore as many stations as possible are kept. Random search can consider more candidates and often is faster. Random replacement with penalty is faster than the greedy algorithms, whereas for detection, the resulting sampling designs were of the same quality: moving a station pays off with a similar improvement in cost.**

***Keywords: update cost, spatial sampling design, space coverage, plume detection, greedy algorithm***

## I. INTRODUCTION

Moving environmental sensors is often expensive and difficult due to the lack of suitable locations. Therefore, when improving existing monitoring networks, it is desired to keep as many stations as possible. Most European countries run a network of sensors for ambient gamma dose rates. They serve for the observation of background values as well as for early warning and coordination of countermeasures in case of nuclear emergencies like accidental releases from nuclear power plants (NPPs). Many of these networks were set up after the Chernobyl accident in 1986. The aims of gamma dose rate monitoring have changed over the last 20 years, and adaption to new threats like terroristic attacks, and international harmonisation are an issue.

We optimise sampling designs for two different goals: equal spread of the stations over the whole area, and detection of plumes that affect many people. The fitness of a sampling design to either of these goals is quantified by cost functions, see section II.A. The optimal designs for these aims are opposed, as stations for detection will be clustered in regions of dense population and high risk, see fig. 2. Therefore they provide good test cases for optimisation algorithms.

Optimisation should minimise the cost functions for sampling designs with few station moves. Therefore an algorithm is considered good, if it finds the sampling design with lowest cost, given the number of station moves. Besides, algorithms should be fast, i.e. find good sampling designs by few evaluations of the cost function. Speed up is important for real-time applications or if the effort for the evaluation of the cost function is high, like in Beekhuizen (2008). We compare two known algorithms and develop a new approach to fit our requirements. Random replacement of stations, keeping all improvements, is a very simple, and often fast, algorithm. Greedy algorithms outperformed other methods in the optimisation of the mean average kriging variance (Baume et al. 2009) and are likely to keep many stations. The new algorithm is random search with penalty on station moves.

The optimisation algorithms considered here move stations one by one and keep the number of stations constant. The greedy algorithm consecutively deletes and adds stations, always selecting the best of all given possibilities. This ensures that few stations are moved, but for each improvement all candidate stations must be browsed, thus the number of candidates is limited. The random algorithm moves stations by chance and keeps all improvements. The number of candidate stations can be very high, but this algorithm tends to move more stations than the greedy algorithm for the same improvement of the cost function. The new approach avoids this: we add a penalty term that is proportional to the number of moved stations, to the cost function. A sampling design is rejected, if the high number of moved stations is not outweighed by the improvement in the original cost function.

## II. METHODS

Monitoring of ambient gamma dose rates in Norway is chosen as the use case, because its existing monitoring network consists of only 27 stations at the main land and that makes optimisation by greedy algorithms feasible. The area is discretised to 5 km x 5 km grid cells $x \in X$, and among these, a sampling design $S = \{x_1, \cdots, x_{27}\}$ is chosen.

### A. Cost Functions

The two aims of the optimisation are translated to cost functions to be optimised by the algorithms. **Equal spread** of the stations within the whole area of interest may be a good choice if measurements shall be interpolated. We define the cost function as the sum of the Euclidian distance from all grid cells to the closest station

$$c_1(S) = \sum_{x \in X} \min_{x_j \in S} \delta(x, x_j)$$

This is a common criterion for space coverage, e.g. used by van Groenigen (1997). It was computed using the spatstat R package (Baddeley and Turner 2005). The cost of the existing sampling design was 7.33E+8.

The **detection** cost function is based on simulated plumes, see section II.B. Let $x \in X$ denote the grid cells, $t \in T$ the time steps. Then $r_i(x,t)$ is the dose rate of plume $i$ at location $x$ and time $t$. It is defined as the hourly average of the dose rate increase above background. We consider a plume to be detected, if the dose rate increase at any time at any sensor location exceeds 50 nSv/h. This is the legal detection limit for ambient gamma dose rate sensors in Germany [AVV-IMIS 2006, Tab. 8.1 1a]. Smaller increases are hard to distinguish from natural variation. Thus we define an indicator function

$$I(i, x_j) = \begin{cases} 1 & : \text{ if } \forall t \in T : r_i(x_j, t) \leq 50\text{nSv/h} \\ 0 & : \text{ else} \end{cases}$$

and use it to define an indicator for the detection of a plume by any sensor of a sampling design

$$\prod_{x_j \in S} I(i, x_j) = \begin{cases} 1 & : \text{ if } \forall x_j \in S : I(i, x_j) = 1 \\ 0 & : \text{ else} \end{cases}$$

The detection cost function is the weighted number of non-detected plumes

$$c_2(S) = \sum_{\text{i plume}} w_i \prod_{x_j \in S} I(i, x_j)$$

The weights $w_i$ are proportional to the risk of the plume

$$w_i = (\Sigma_{x \in X} p(x) \ \Sigma_{t \in T} r_i(x,t))^q$$

where $p(x)$ is the population in the respective grid cell. The exponent is $q = 0.05$, to attenuate the effect of inhomogeneous population density such that weights finally differ by about a factor of 10. The cost of the existing sampling design was 45.2.

*B. Datasets*

The detection cost function is based on simulations of plumes, generated by the dispersion model NPK-puff (Twenhöfel et al. 2007). The weather data is taken from simulations and measurements for the full year 2005 at De Bilt, Netherlands. Simulations run 48 h from randomly chosen start dates and yield hourly average dose rate increases for each grid cell. Two types of outbreaks are considered, similar to Melles et al. (2009), see tab. I, release duration is always 1 h.

TABLE I. SOURCE TERMS OF THE SIMULATED PLUMES

| | Nuclide | Radioactivity | Height | Heat cont. |
|---|---|---|---|---|
| "NPP accident" | Kr 88 | 1E+16 Bq | 18 m | 1 MW |
| "Terroristic attack" | Cs 137 | 1E+13 Bq | 3 m | 100 kW |

"Nuclear power plant accident" outbreaks were started at eight NPPs in neighbour countries of Norway, following Lauritzen et al. (2005). Fifty plumes started at each of the NPPs. Of these 400 plumes, 30 touched Norway and were taken into account. At each of the five biggest cities in Norway, ten "terroristic attack" outbreaks were started. 45 of these 50 plumes touched Norway. Altogether 75 plumes were considered, see fig. 1.

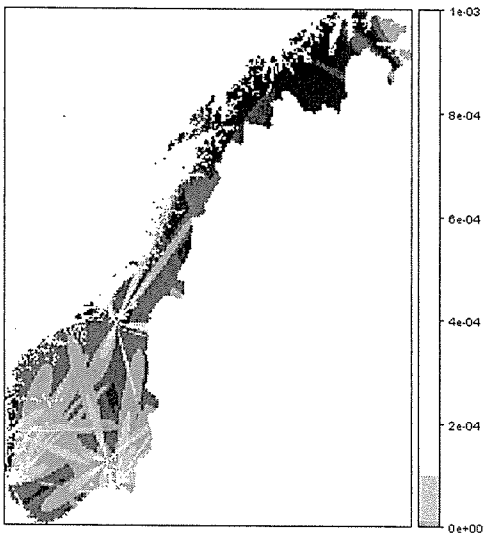


Figure 1. Sum of dose of all plumes

Population per grid cell is derived from a map of all settlements with more than 200 inhabitants (Statistisk Sentralbyrå 2010), assuming equal distribution of the remaining 21% of the population. All optimisations started with the current locations of gamma dose sensors at the Norwegian main land.

*C. Optimisation Algorithms*

Four algorithms were run on each of the two cost functions: a greedy algorithm with full and reduced set of candidates, random search, and random search with penalty on station moves.

The **greedy algorithm** first checks all sampling designs with one of the original stations deleted and considers the one with the best cost for further improvement. Next, the candidate stations are added one at a time and the best one is kept. Then, the next station is deleted, and so on, until no further improvements are found. Computational effort to replace one sampling location is proportional to the number of existing and candidate sampling locations, therefore a subset of candidates must be chosen from the set of all possible locations. We choose sets of 400 candidates to limit the maximal number of iterations needed to move all 27 stations to 10 000. These candidate sets should allow near-optimal solutions, therefore different candidate sets are used for the two cost functions. For space coverage a randomly placed, regular cell spacing is used. For the detection cost function, we use settlements with more than 850 inhabitants.

For comparison we also run the greedy algorithm with all grid cells as candidates.

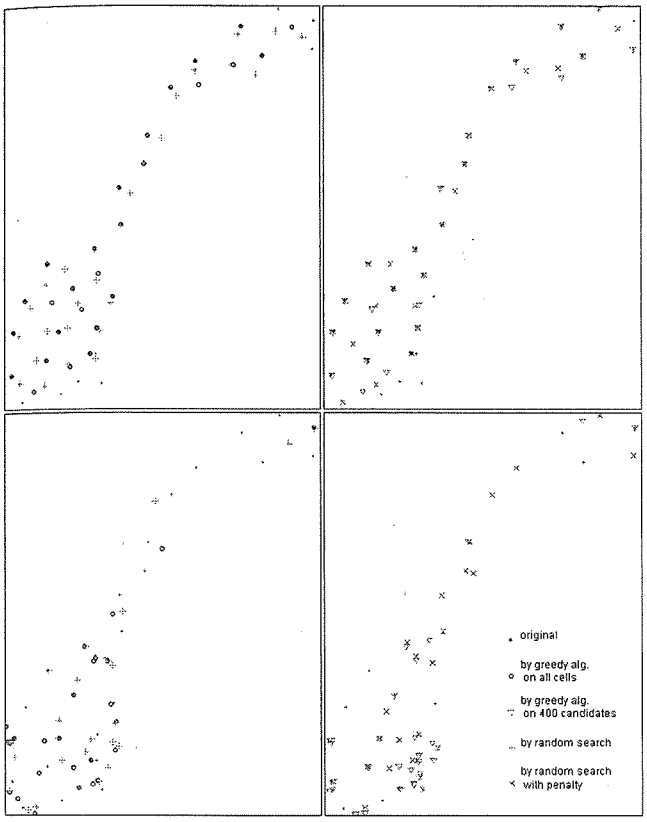


Figure 2. Optimised sampling designs for equal spread (upper) and detection (lower)

The **random algorithm** was run on the set of all 13 000 grid cells. It randomly replaces one of the existing stations and keeps all improvements. In contrast, Spatial Simulated Annealing (SSA, van Groenigen 1997) decreases the size of the moves during the run and accepts deterioration with a low probability. SSA did not improve the results and was therefore abandoned. Optimisation is terminated after 1000 iterations with no improvement.

The new approach is random search with modified cost functions. A **penalty term** $C_k(S)$, $k = 1,2$ is added to the cost functions to focus the random algorithm on sampling designs that keep many of the existing stations. It is proportional to the number of moved stations

$$n(S) = |S_{orig} - S|$$

to reject sampling designs where this number is high and not compensated by an else very low cost function. Thus the cost functions change to

$$c_k(S) + C_k(S) = c_k(S) + n(S) \cdot a_k \qquad k = 1,2$$

where $a_k$ is an adjustment factor. It is set proportional to the difference of the cost function for the existing sampling design $S_{orig}$ and the one of a sampling design with one optimal station $\tilde{x}$ added, and multiplied with a calibration factor of 0.1

$$a_k = 0.1 \cdot |c_k(S_{orig}) - c_k(S_{orig} \cup \{\tilde{x}\})|$$

The computation was run in the R environment for statistical computing (R Development Core Team 2009) using the package sp (Pebesma and Bivand 2005).

## III. RESULTS

Greedy algorithms stopped, when deleting and adding one station not further improved the cost. Random algorithms terminated after 1000 iterations with no improvement. For equal spread the sampling design with lowest cost was found by the random algorithm, but only with many stations moved, see tab. II, fig. 3. For detection, the greedy algorithm on all cells outperformed the other algorithms. The disadvantage however was long computation time of about 3 days, which is about 100 times more than for the other algorithms, see tab. III, fig. 4. The data for the random algorithms are averages of 10 runs.

TABLE II. FINAL OPTIMIZED SAMPLING DESIGNS FOR EQUAL SPREAD

| Algorithm | Cost | Station moves | Iterations total |
|---|---|---|---|
| greedy on all cells | 64.8e+7 | 8 | 11799 |
| greedy on 400 candidates | 66.6e+7 | 7 | 3385 |
| random search | 61.0e+7 | 24 | 7322 |
| random search with penalty | 62.9e+7 | 14 | 5062 |

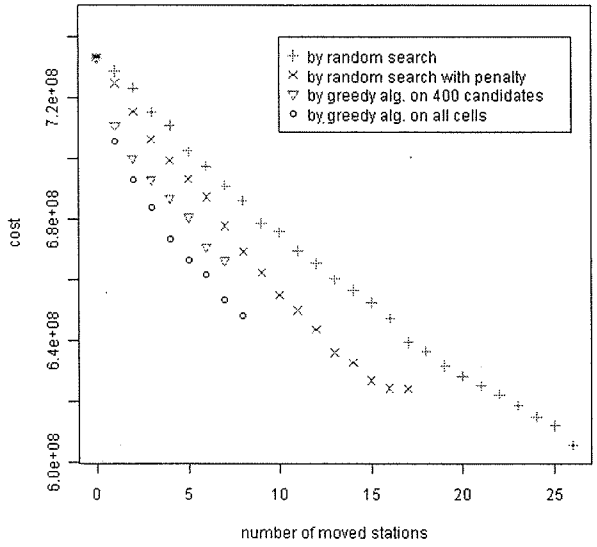


Figure 3. Optimisation runs for equal spread. Cost for the optimal sampling designs with given number of station moves.

Each algorithm found sampling designs with different numbers of station moves. Fig. 3 shows the results of the optimization of the equal spread cost function. It shows the cost of sampling designs with the indicated number of stations moved. The greedy algorithms performed best and random search with penalty is better than without. For simplicity, the number of iterations needed is not considered

in this comparison. This is done for the detection cost function below.

TABLE III. FINAL OPTIMISED SAMPLING DESIGNS FOR DETECTION

| Algorithm | Cost | Station moves | Iterations total |
|---|---|---|---|
| greedy on all cells | 3.0 | 21 | 301393 |
| greedy on 400 candidates | 16.1 | 20 | 8884 |
| random search | 8.9 | 22 | 3615 |
| random search with penalty | 16.0 | 12 | 2934 |

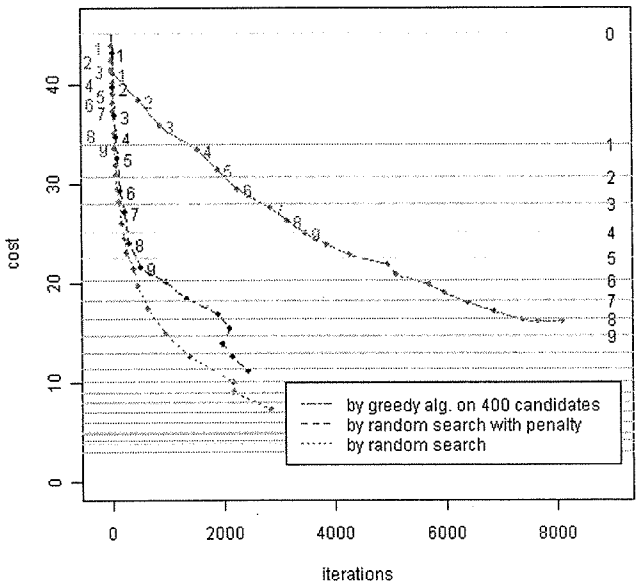


Figure 4. Optimisation runs for detection. Cost and number of station moves for the sampling designs found.

Fig. 4 shows the optimisation runs for the detection cost function. The lines go through the sampling designs found during these runs, showing cost and number of iterations needed to find this sampling design. The numbers indicate the number of stations moved in the respective sampling design. The horizontal lines refer to the results of the greedy algorithm on all cells. This shows for example, that the random algorithm reached low cost very fast. However, it moved 9 stations to obtain a cost that was similar to the cost of a sampling design with only one station moved, found by the greedy algorithm on all cells. It is obvious that the performance of the greedy algorithm on 400 candidates and random search with penalty is between these extremes and that the latter is faster. The numbers are at similar cost levels for both algorithms, this means, they find sampling designs with same number of stations moved and similar cost.

## IV. CONCLUSION AND DISCUSSION

Greedy algorithms as well as random algorithms with penalty were able to determine sampling designs with few station moves and low cost.

The greedy algorithm on all cells as candidates yielded very good results for the detection cost function. However, if computation time is an issue, random search with penalty on station moves is faster than and as good as the greedy algorithm on a reduced set of candidates. For the equal spread cost function, random search outperformed greedy algorithms in improving the cost function, if number of station moves is not limited. For keeping stations under the optimization, random search with penalty was better than without and faster but worse than the greedy algorithms.

The cost functions used here were simplistic and should be improved to find optimal sampling designs for Norway. A major improvement would be the use of more realistic weather data for the simulation of the plumes.

## ACKNOWLEDGEMENTS

The authors are grateful to Arjan van Dijk (RIVM, NL) who provided the NPK-Puff simulation program, the weather data, and considerable support for its use and to Jan Erik Dyve (NRPA, NO) for data about the existing GDR network in Norway. This work was funded by the European Commission under the Seventh Framework Program, by the Contract N. 232662. The views expressed herein are those of the authors and not necessarily those of the European Commission.

## REFERENCES

Allgemeine *Verwaltungsvorschrift zum Integrierten Mess- und Informationssystem zur Überwachung der Radioaktivität in der Umwelt (IMIS) nach dem Strahlenschutzvorsorgegesetz (AVV-IMIS)*, 13 Dec 2006.

Baddeley, A., and Turner, R. (2005). spatstat: an r package for analyzing spatial point patterns. *Journal of Statistical Software*. 12 (6).

Baume, O., Gebhardt, A., Gebhardt, C., Heuvelink, G., and Pilz, J. (2009). Network optimization algorithms and scenarios in the context of automatic mapping. In: *StatGIS 2009*.

Beekhuizen, J. (2008). *Dealing with uncertainty in determining the optimal locations of mobile measuring devices*. Wageningen University.

van Groenigen, J.W. (1997). Spatial simulated annealing for optimizing sampling. In: *GeoENV I Geostatistics for environmental applications*. (pp.351–361).

Lauritzen, B., Jensen, P.H., and Nielsen, F. (2005). Requirements to a Norwegian National Automatic Gamma Monitoring System. *Risø National Laboratory, Roskilde*.

Melles, S.J., Heuvelink, G.B.M., Twenhöfel, C.J.W., van Dijk, A., Heimstra, P., Baume, O., and Stöhlker, U. (2009). Optimization for the design of environmental monitoring networks in routine and emergency settings. In: *StatGIS 2009*.

Pebesma, E.J., and Bivand, R.S. (2005). Classes and methods for spatial data in r. *R News*. 5 (2), 9–13.

Statistisk Sentralbyrå (2010). *Kart over tettsteder 2009*, [Online], available: http://www.ssb.no/beftett/ [14 Jan 2010]

R Development Core Team (2009). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Twenhöfel, C.J.W., van Troost, M.M., and Baader, S. (2007). Uncertainty analysis and parameter optimisation in early phase nuclear emergency mangement. Bilthoven: RIVM.