Automatic detection and correction of heterogeneities between measurement networks

Jon O. Skøien[1], Olivier Baume[2], Edzer J. Pebesma[3], Gerard B.M. Heuvelink[2]

[1] Department of Physical Geography / Utrecht University
j.skoien@geo.uu.nl

[2] Land Dynamics Group, Wageningen University
Olivier.Baume@wur.nl
Gerard.Heuvelink@wur.nl

[3] Institut für Geoinformatik / Westfälische Wilhelms-Universität Münster
edzer.pebesma@uni-muenster.de

Abstract. As environmental variables do not respect country borders or administrative responsibilities, there is an increased interest in merging observations from different sources. However, different countries and different administrative bodies might have different measurement devices and different traditions for measuring. These differences can be seen as heterogeneities between different data sets, and will cause discontinuities if continuous maps are interpolated from the data. This paper presents methods to identify these heterogeneities as differences between networks, both when emerging from different national networks, and from different networks in a country. These differences can then be used for identification of biases for each single network. More homogeneous maps can be achieved by removing these biases.

1 INTRODUCTION

Environmental variables are correlated on large distances, independent on regional or national borders. Hence, there is an increased interest in merging the observations from local observation networks into international databases. Some examples can be found in forestry (*Köhl, et al.,* 2000), soil quality (*Wagner, et al.,* 2001) and nuclear radiation (*De Cort and De Vries*, 1997). One common problem in this procedure is that different network owners use different measurement devices, or treat the measurements differently before uploading the values to a central database.

If the merged data base is to be used as the source for creating maps of the environmental variable, these heterogeneities will lead to discontinuities in the map that are not a result of true variability of the variable. If it is not possible to explain the heterogeneities by deterministic factors, it is necessary to estimate the differences between different measurement networks with statistical methods, and correct for these differences. The aim of this paper is to present two such methods, one method for networks with overlapping boundaries (e.g. two or more networks within country) and one method for networks without overlapping boundaries (e.g. networks in neighbouring countries). This is done within the framework of the project INTAMAP (Interoperability and Automated Mapping, http://www.intamap.org), which aims to develop automatic interpolation methods for environmental variables.

2 Data

As an example, we use data from the EURDEP database of radioactive gamma dose rate measurements, collected from national gamma dose rate networks in more than 30 European countries (*De Cort and De Vries*, 1997). The monitoring networks in the different countries consist of everything from a few to more than 2000 monitoring stations (Germany). In most countries, there is only one monitoring network, whereas some countries have more than one network owner, each with their separate network. There are multiple sources of heterogeneity in the data that are uploaded to EURDEP. *Bossew et al.* (2008) analyzed the heterogeneities questioning all network owners about possible explanations for biases, but there were still heterogeneities not accounted for.

We tested the heterogeneity detection and correction methods using the gamma dose rate monthly averages from January 2006 of all stations in the EURDEP data base belonging to the European main land (excluding Iceland and the British Isles). Slovenia has observations from three different networks, and was used as an example of a country with multiple networks. The country boundaries were taken from the country shape-file of the European soil data base (http://eusoils.jrc.it). Each country border could then be found as a set of points defining the nodes of the border, in the range from 19 to 852 points. Most country borders are defined with more than 100 points.

3 Methods

We assume that the observed value y_{ij} at location s_{ij} in network i is the sum of three parts:

$$y_{ij} = \mu(s_{ij}) + e(s_{ij}) + b_i \tag{1}$$

where $\mu(s_{ij})$ is the (possibly spatially varying) mean, $e(s_{ij})$ is a zero-mean spatially correlated random variation and b_i is the network bias. The heterogeneities b_i between different networks were removed in a two-step procedure. First, the biases for each network within a country were identified and removed. Second, the biases between countries were identified and removed.

Geostatistics was used for estimating the differences between networks in both cases. For multiple networks within a country, observations in one network were used to predict the process at the locations of one of the other networks by ordinary kriging, given that the two networks share much of the same area. For all networks *p*, we can define a bias vector **B** and a difference vector **Q** as follows:

$$\mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} b_1 - b_2 \\ \vdots \\ b_1 - b_P \\ b_2 - b_3 \\ b_2 - b_4 \\ \vdots \\ b_{P-1} - b_P \end{bmatrix} \tag{2}$$

Thus, the elements of the vector Q were estimated using kriging as described above. If a network covers less than half of the region covered by another network, we did not estimate the difference $b_1 - b_2$. However, it might still be possible to estimate the difference $b_2 - b_1$. We therefore assume that there are more differences than networks. The underlying assumption for this method is that the observation locations in each network are independent of the intensity of the process (gamma radiation values in this case).

The differences between countries where found in a similar way, by predicting the process at the border between the two countries, using the data sets from each of the two countries separately. This is a form of stratified kriging. In this way, an estimate of vector **Q** is obtained.

In both cases, the relationship between vectors **Q** and **B** can be expressed through the use of a relationship matrix **D**, as:

$$\mathbf{Q} = \mathbf{DB} \tag{3}$$

where e.g. the first row of D has 1 and -1 in the two first columns, corresponding to the bias of the first and second network, respectively.

As there are more equations than biases, we can use Ordinary Least Squares to identify the biases. However, it is necessary to add at least one unbiasedness constraint, for identification of the interception. A logical choice is to choose that the sum of all biases is equal to zero. This will come in as a row of 1's in **D** and a zero in **Q**. We can then find an estimate for **B**:

$$\hat{\mathbf{B}} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{Q} \tag{4}$$

4 RESULTS AND DISCUSSION

4.1 Estimate of biases within one country

There are three networks in Slovenia. Figure 1 gives an indication of the locations of the stations. There are five stations in the first network (marked with small circles in Figure 1), 17 stations in the second network and 13 stations in the third network around the nuclear power plant Krško Nuclear Power Plant (marked with a larger circle). One of the stations in network 1 is reported to be located outside the borders of the country.

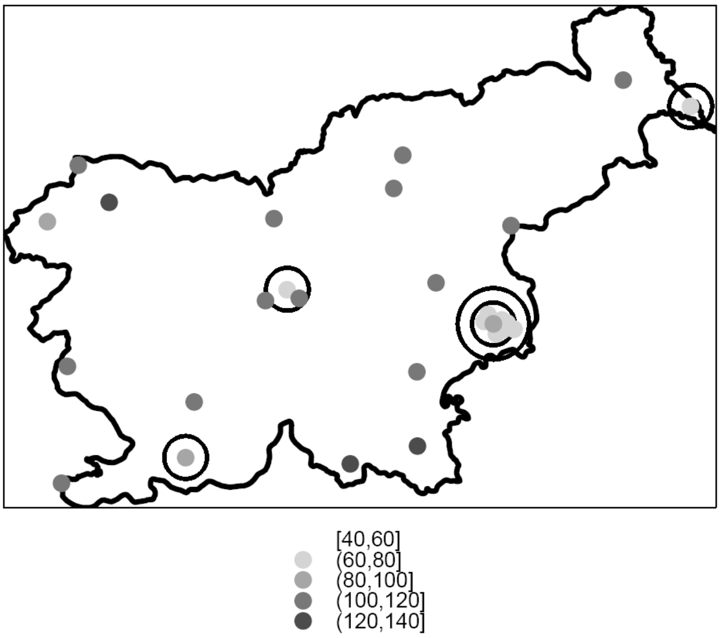

Figure 1: Map of Slovenia with observation stations. Small circles around observations in network 1, large circle around nuclear power plant. Units in nSv/h.

The estimated differences between the three networks (**Q**) are shown in Table 1.

Table 1: Estimated differences between Slovenian Gamma dose networks (matrix **Q**)

	Difference
b_1-b_2	-39.8
b_2-b_1	40.1
b_1-b_3	14.2
b_2-b_3	51.5

The matrix **D** could easily be found from Equation 3, and the biases could be estimated according to Equation 4. The results are in table 2. The results indicate that networks 1 and 3 measure too small values, whereas network 2 has a positive bias (compared to the average). The difference between the second and third network (about 50 nSv/h) is slightly greater than the information we later received from the Slovenian Nuclear Safety Administration (SNSA), which reported a difference of approximately 40 nSv/h. The estimation error of 10 nSv/h is acceptable for such a small number of stations.

Table 2: Estimated biases for Slovenian Gamma dose networks ($\hat{\mathbf{B}}$)

	Difference
b_1	-8.8
b_2	30.7
b_3	-22.0

Figure 2 shows the interpolated maps of gamma radiation in Slovenia before and after removal of heterogeneities. The map to the right shows that the spatial patterns are strongly influenced by the configuration of observation locations in network 1, whereas the second map does not indicate any spatial pattern due to the observation locations.

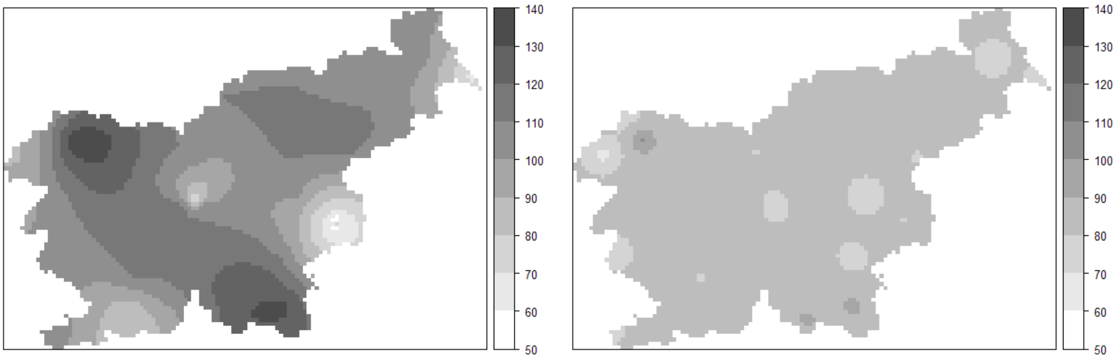

Figure 2: Interpolated maps of gamma radiation in Slovenia, before and after removal of network biases. Units in nSv/h.

4.2 Estimate of biases between countries

Applying the method from Section 3 to the mainland of Europe, we estimated 45 differences between 28 European countries. Using the OLS method as presented in Equation 4, we estimated the biases for each country. The results are presented in Figure 3, as a bar chart. The method identified both heterogeneities explained in the AIRDOS report (*Bossew, et al.*, 2008) and heterogeneities described but not explained in the report. The largest (absolute) biases were estimated for Greece, Spain and Romania.

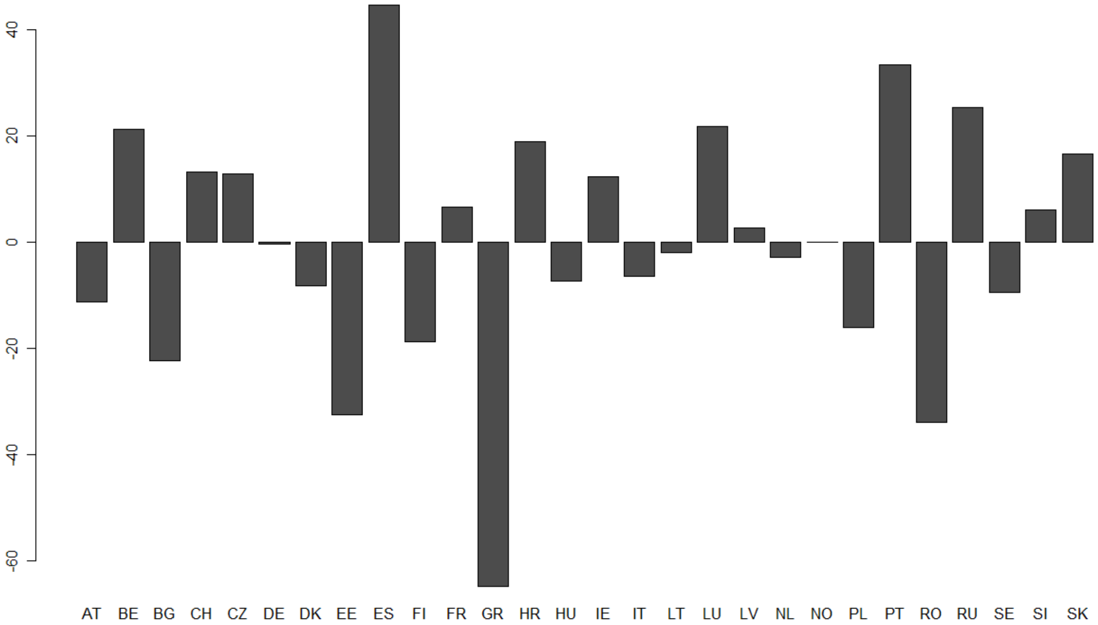

Table 3: Biases of national Gamma radiation networks. Units in nSv/h.

5 Discussion and Conclusions

Results indicate that the heterogeneity detection and correction methods give reliable estimates of network and country biases. However, after closer examination of the estimated biases, we have limited confidence in the estimated bias for Greece. This seems, among other possibilities, to be a result of an unfortunate configuration of sampling points and trends in the data due to different soil. The underlying assumption behind the method is that there are no trends in the data leading to discontinuities at the borders, which seems to be the case here. As the method is supposed to be used as a background operation without human interaction, we are still working on how to better deal with such problems. Possible solutions include identification and explanation of these trends.

The methods, as presented in this paper, do not use the prediction uncertainty that is a natural result from kriging methods. In the near future we plan to extend the method to include the prediction uncertainties for weighted least squares estimation of the biases, and also to estimate the significance of the bias estimates.

Acknowledgements

This work was funded by the European Commission, under the Sixth Framework Programme, by the Contract N. 033811 with the DG INFSO, action Line IST-2005-2.5.12 ICT for Environmental Risk Management. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

6 References

Bossew, P., M. De Cort, G. Dubois, U. Stöhlker, T. Tollefsen, and U. Wätjen (2008) AIRDOS, Evaluation of existing standards of measurement of ambient dose rate; and of sampling, sample preparation and measurement for estimating radioactivity levels in air., AA N° TREN/NUCL/S12.378241, JRC ref. N°21894-2001-04 A1CO ISP BE.

De Cort, M., and G. De Vries (1997) The European Union radiological data exchange platform (EURDEP): "Two years of international data exchange experience." Radiation Protection Dosimetry 73 (1-4): 17-20.

Köhl, M., B. Traub, and R. Päivinen (2000) "Harmonization and standardization in multinational environmental statistics - mission impossible?" Environmental Monitoring and Assessment 63: 361-380.

Wagner, G., A. Desaules, H. Muntau, S. P. Theocharopoulos, and P. Quevauviller (2001) "Harmonisation and quality assurance in pre-analytical steps of soil contamination studies - conclusions and recommendations of CEEM Soil project." The Science of The Total Environment 264: 103-117.