European conference of the Czech Presidency of the Council of the EU
TOWARDS eENVIRONMENT
Opportunites of SEIS and SISE: Integrating Environmental Knowledge in Europe
http:/www.e-envi2009.org/proceedings/
J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pilmmann, I.Holoubek, T. Bandholz (eds.)
Masaryk University, 2009

Data harmonization of environmental variables: from simple to general solutions

Olivier P. Baume[a], Jon O. Skøien[b] , Florence Carré[c], Gerard B.M. Heuvelink[a] and Edzer J. Pebesma[d]

[a] *Department of Environmental Sciences, Wageningen University, the Netherlands (olivier.baume@wur.nl; gerard.heuvelink@wur.nl)*

[b] *Department of Physical Geography, University of Utrecht, the Netherlands (j.skoien@geo.uu.nl)*

[c] *Land Management & Natural Hazards Unit, DG JRC, Ispra, Italy (florence.carre@jrc.it)*

[d] *Institute for Geoinformatics, University of Münster, Germany (edzer.pebesma@uni-muenster.de)*

Abstract: European environmental databases can be merged if regional networks or experimental campaigns target the same variable. In practice, measurements contain biases especially when compared from one network to another, or from one experiment campaign to another. Therefore merging is sensible only when data are harmonized. Harmonization of European environmental databases is often meaningful for decision making, and sometimes it is crucial. In this paper, a solution for data harmonization, developed under the INTAMAP FP6 project, is presented. The INTAMAP FP6 project is currently developing an interoperable framework for real time automatic mapping of critical environmental variables by extending spatial statistical methods. The harmonization procedure takes only additive biases into account. The bias is estimated and removed before final interpolation. We describe two applications of harmonization at a European level. The original motivation for INTAMAP came from the field of radiology, where a European real-time database has been active for all member states over a decade. This so-called EURDEP database contains heterogeneities because national networks measure and report data following different devices and processes. As an example, monthly averaged gamma dose measurements across eight European countries are harmonized. The second application deals with the soil Carbon/Nitrogen (CN) ratio in Europe. The CN ratio of the forest soil cover is one of the best predictors for evaluating soil functions, such as required in climate change issues. Although samples were analyzed according to a common laboratory method, preliminary statistical analyses showed that laboratories introduced errors in the measurements and should be taken into account. Results from both applications are discussed to continue the work of harmonization.

Keywords: Measurement error; Sensor calibration; Geostatistics, Uncertainty, Environmental model.

1. BACKGROUND

A problem with international environmental databases is that values reported on a specific target variable may not have followed the same data process from one network to another, or from one campaign to another. Examples include the use of different measuring devices, the use of different standards, or the analysis of samples in different laboratories. Such discrepancies in a database may lead to wrong decisions in policy making: mapping

environmental variables across borders may lead to wrong detections of threshold values at which remediation actions are required; mapping environmental variables at a European scale may lead to wrong analysis and model building of the target variable.

In environmental assessment, while standardization implies the definition of a common way to measure a specific target variable, harmonization is a bottom-up approach for gathering different standards. Standardization is generally accepted as an obvious solution (see for instance Wagner et al. [2001], Schröder at al. [2006]), although standardization can become very costly when national databases already exist for many years. In the extreme case of several historical campaigns, standardization is not an option for analyzing legacy data. As a more flexible alternative to standardization, several authors from the field of geostatistics suggest harmonization procedures to handle heterogeneous databases.

Harmonization modelling makes a clear distinction between the target variable that one wants to map, and the measured variable that is in fact reported from a particular network or measurement campaign. Fassó et al. [2007] applied a space-time kriging model for air quality mapping in the region of Lombardy, which includes additive and multiplicative biases between two measurement networks of PM10. Brenning et al. [2008] analysed the discontinuity errors between field measurements to map soil conductivity from field scale to landscape scale. Skøien et al. [2009] used a line kriging method to identify discontinuities at country borders among 30 networks of the European radioactivity exchange database. Baume et al. [2007] applied a universal kriging model to filter additive biases between national networks. The latter was extended to a Bayesian setting in order to include prior knowledge on the biases between networks (Baume et al. [2008]).

All harmonization methods have been applied to specific examples. In this paper we consider the method developed by Baume et al. [2008], which may be applied with or without prior information on the biases. We apply the method to two different examples. First an application on radioactivity exposure aims at removing biases between national networks in order to process automatic mapping of harmonized data. The example further uses the network gaps estimated by Skøien et al. [2009] as prior information in the model. The second example targets soil quality data of the soil Carbon/Nitrogen (CN) ratio from Europe. Harmonization of several national campaigns may lead to a better understanding of the influence of tree type on CN ratio values.

2. METHODS

The main principle of harmonization is to distinguish a common target variable Z defined for the whole database, from the measured variables Y_i, which are indexed for each specific network or measurement campaign. A way to denote the relationship is to use some general calibration functions F_i so that

$$Z(\mathbf{S}_i) = F_i^{-1}(Y_i(\mathbf{S}_i)), \tag{1}$$

where the $\{\mathbf{S}_i\}$ are the measurement locations of the specific network or measurement campaign i. The calibration functions F_i are device-dependent. For simplicity we assume that one additive bias b_i per network or campaign i provides sufficient harmonization. Thus, equation (1) reduces to

$$Z(\mathbf{S}_i) = Y_i(\mathbf{S}_i) - b_i. \tag{2}$$

2.1 Geostatistics and harmonization

There are two main methods to estimate the biases b_i. The first method is to indirectly consider the local differences between measurements from neighbouring networks or campaigns, and to transform the estimated differences into biases (see Brenning et al. [2007], Skøien et al. [2009]). The second method is to directly consider the presence of

biases in an overall linear model by means of a universal kriging model (Baume et al. [2007]). In this paper we consider the second method.

Universal kriging is a geostatistical method for mapping environmental variables that may include some explanatory variables $\{X_k\}$ as well. For example, the background radioactivity level (gamma dose for global radioactivity exposure) is influenced by the altitude of the measurement and by soil type. In the general case, we relate the target variable Z with K explanatory variables X_k through K coefficients A_k to be estimated. Writing the relationships in terms of random variables, and introducing residuals R_Y and R_Z, the overall linear model of universal kriging for harmonization writes:

$$\begin{cases} Y_i(\mathbf{S}_i) = & Z(\mathbf{S}_i) + b_i + R_{Y_i}(\mathbf{S}_i) \\ Z(\mathbf{S}_i) = & \sum_{k=1}^{K} X_k(\mathbf{S}_i) \cdot A_k + R_Z(\mathbf{S}_i) \end{cases} \tag{3}$$

.

We assume that the stochastic residuals R_Y and R_Z have zero-mean and that residuals of the true state variable R_Z have a constant covariance structure in the whole domain of study.

2.2 A harmonized database for mapping and modelling

We solve the system (3) by means of the least squared difference method (see Baume et al. [2007] for a detailed description). In a first step, we obtain an estimate of the biases $\{b_i\}$, an estimate of the linear relationships A_k between the target variable and its explanatory covariates, and an estimate of the covariance of the true process R_Z. We use a simple iterative process from ordinary least squares estimation to general least squares, until convergence. In a second step, as a result of universal kriging, it is possible to directly compute an optimized interpolation where the biases are removed.

We call this interpolation procedure harmonized kriging. Optionally some information can be added on the model prior to the estimation procedure. The inclusion of prior information on the biases has been reported by Baume et al. [2008] in the context of radioactivity exposure. In harmonized kriging, the inclusion of prior information is guided by the fact that in linear model (3), multicollinearity among covariates and biases is a source of estimation impairment and misinterpretation.

In a simple framework that only includes additive biases b_i in model (3), one problem that remains with harmonized kriging is the choice of a reference. By default, the biases are computed such that they sum up to 0, but alternatively, if the measurement standard of one of the networks or campaigns can be elected as a reference, its bias may be set to 0.

3. APPLICATIONS

Harmonization is anticipated to give more accurate decisions because the harmonization model (3) makes an explicit distinction between the measured variables and the true state variable. Kriging interpolation maps provide a probability distribution of the target variable at each interpolated location that eventually lead to the probabilistic computation of threshold detections. Without a harmonization procedure in the analysis of a database crossing borders, the risk of false decision is greater.

3.1 Radioactivity exposure and risk assessment

We take the example of bias estimation in gamma dose rate data (physical unit nSv/h) from the European Data Exchange Platform (EURDEP – http://eurdep.jrc.it/). This database is used for radioactivity background level estimation and for early warning situations. Monthly averages of December, 2006 were extracted for eight countries – Austria (AT),

Belgium (BE), Switzerland (CH), Czech Republic (CZ), Germany (DE), Italy (IT), Luxemburg (LU) and Netherlands (NL). Within each of these countries the data are assumed to belong to one network only. Different networks may use different probe types and data processing before the upload to the EURDEP database. For instance, Germany subtracts the self-effect of the probes whereas not all other countries do. Despite an attempt to collect such information from each country (Bossew et al. [2007]), there are still differences between national networks that are not accounted for. As a consequence, the harmonization can only be based on the national membership of the measurements.

Following model (3), we assumed one bias per country and included two covariates in the linear model of the target variable *Z*. The first covariate is elevation because elevation influences the amount of terrestrial radiation detected: secondary cosmic radiation raises exponentially with altitude hence augmenting background gamma dose rate level (Wissmann et al. [2008]). As most measurement sites are located at low altitudes, non-symmetric distributions appear for mountainous countries. The second covariate is soil type. For instance, radioactivity in volcanic regions is generally higher. Lowest background radioactivity can be found mostly in alluvial plains.

To improve harmonization, the harmonization model (3) was completed with prior information on the biases (see Baume et al. [2008]). These prior values correspond to local estimation of biases along borders. Local estimates of the biases along borders - and respective standard deviations – were obtained with the method developed by Skøien et al. [2009]. As these local estimates are less sensitive to covariates, they can be considered sensible prior information. The prior values of biases are given in Table 1.

3.2 Carbon/Nitrogen ratio for evaluating soil functions

As an indicator of soil mineralization processes, the CN ratio of forest soils is one of the best predictors for evaluating soil functions such as biomass production and carbon storage capacity of forest soils. When integrated to risk assessment, these functions can serve for modelling scenarios of soil sustainability with climate change issues (e.g. gas fluxes emissions, biofuel production). For instance, for a soil having a relative high CN ratio, the mineralization process tends to be slower and the weak leaching of nitrogen results in a weak quantity of N gas fluxes emission.

The CN ratio is strongly dependent on the forest species and management, and on environmental factors (Burke et al. [1989]) such as climate, relief, soil type and parent material. Furthermore, since the forest management is done locally, the CN ratio variability has to be analysed locally. After a preliminary data analysis (Carré et al. [2008]), data harmonization appears an important step to level out main gaps between local sample analysis campaigns. Analysis of soil cover samples were analysed by several laboratories, each having somewhat different procedures and circumstances. Therefore laboratory origin is considered as the origin of bias in the dataset. Most countries centralised the sample analysis in one laboratory except for Germany, where analysis was centralised per region. In the CN ratio application, the model of the target variable *Z* included two covariates. First, each measurement sample was related to the percentage of two complementary tree cover types: coniferous cover and broadleaves cover. Second, the model included the pH as a continuous covariate. No prior information on the biases was available in this application.

4. RESULTS AND DISCUSSION

4.1 Influence of prior information on bias estimation in Radioactivity exposure

A comparison of bias estimates in the radioactivity case given in Table 1 shows that both the local method (“Prior values”, Skøien et al. [2009]) and harmonized kriging (“Without prior”) yield similar results. Table 2 gives the “Posterior with prior” estimates of the country biases. Standard deviation values associated to the prior estimates (Table 1, column 4) correspond to the inverse of the weight given to the prior values to estimate the posterior

values. In this example, prior values have larger standard deviations than harmonized kriging estimates because local estimation of biases rely only on the measurements close to the borders. In the extreme case of Italy where the network is sparse, the standard deviation is the largest.

Table 1. Radioactivity exposure – gamma dose: network bias estimates and associated standard deviations for harmonization with and without prior information (*nSv/h*).

Country	without prior	*Standard deviation*	Prior values	*Standard deviation*	Posterior with prior	*Standard deviation*
Austria (AT)	-18.31	*2.04*	-15.00	*3.24*	-18.29	*1.96*
Belgium (BE)	10.76	*2.66*	14.15	*2.95*	12.25	*1.93*
Switzerland (CH)	10.33	*2.98*	11.23	*2.55*	10.36	*1.93*
Czech Republic (CZ)	-0.63	*2.87*	2.18	*3.95*	0.10	*2.27*
Germany (DE)	-5.88	*1.43*	-3.52	*2.60*	-5.20	*1.28*
Italy (IT)	-2.81	*4.41*	-18.97	*12.14*	-6.78	*3.76*
Luxembourg (LU)	16.02	*4.08*	17.96	*5.49*	16.18	*3.15*
Netherlands (NL)	-9.49	*2.09*	-8.05	*2.85*	-8.62	*1.64*

4.2 Interpretation of harmonized kriging in the Carbon / Nitrogen ratio mapping

As no prior information was included in the CN ratio case study, we plotted the histogram of the correlation level between natural covariates (cover type and pH) and the membership of the measurements (related to countries and regions) to evaluate the risk of impairment in the estimation process (Figure 1). Correlation values are low with absolute values lower than 0.4. The majority of correlation absolute values in the model are lower than 0.1. Table 2 presents the bias estimates and their standard deviation. Figure 2 compares the interpolation map without correction of the biases (mentioned as “ordinary kriging”) and the interpolation with biases removed (using “harmonized kriging”).

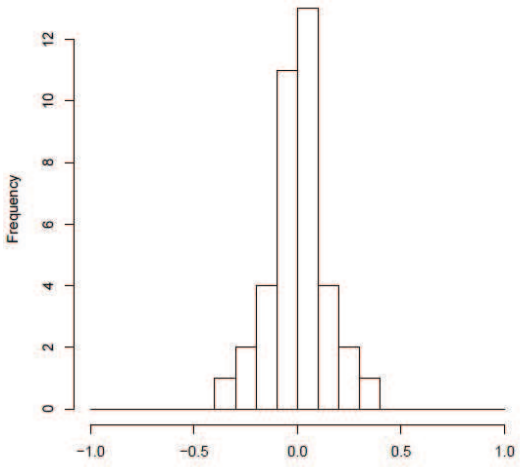


Figure 1. Carbon/Nitrogen ratio: histogram of correlation coefficients between natural covariates and country / region membership.

CN ratios are overestimated in Belgium, Finland and Latvia, whereas Great-Britain and Poland for example were underestimating the CN values (Table 2). Lowest standard deviations are found in largest regional datasets. Harmonization decreases differences between countries. However, the harmonized map still shows hotspot values in France, Latvia, Germany and Poland (see Figure 2(b)). Hotspots are due to the presence of

coniferous forests and low pH (acid soils like podzols), for which the decomposition of organic matter is slow (the CN ratio is then high). pH and forest species allowed to take the soil environment into account for correcting the CN ratios.

Table 2. Carbon / Nitrogen ratio: country/laboratory biases. Data were analysed by one laboratory in each country, except for Germany.

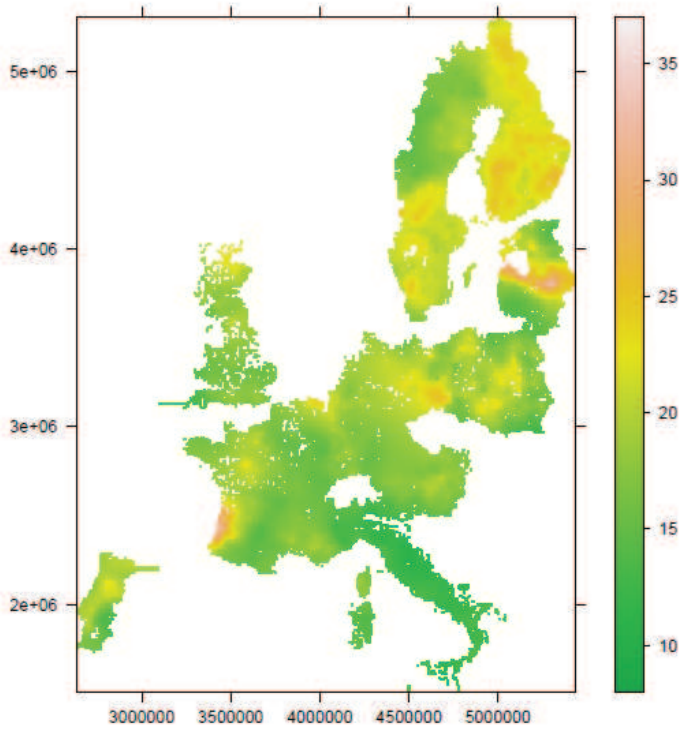
Country (lab. code)	Bias estimate	*Standard deviation*
Austria	-1.2	*0.8*
Belgium	2.6	*2.0*
Estonia	-1.5	*0.9*
Finland	2.4	*0.5*
France	0.9	*0.5*
Germany (401-Lab01)	-2.1	*1.2*
Germany (404-FHE)	0.7	*1.2*
Germany (408-lab_0)	-1.2	*1.6*
Germany (410-lab_g)	0.5	*1.2*
Germany (411-1)	-0.7	*1.6*
Germany (412-Lab_4)	1.5	*2.3*
Italy	-2.0	*0.6*
Latvia	7.5	*1*
Lithuania	-2.8	*1.1*
Poland	-1.7	*0.5*
Portugal	1.3	*0.8*
Slovenia	-0.7	*1.5*
Sweden	-1.0	*0.5*
Great-Britain	-2.4	*0.6*

The harmonized map also shows spatial gradients in the values, notably in England (from North to South), France (from West to East) and Scandinavia (from North to South). These gradients can mainly be explained by bioclimatic gradients which have an impact on the decomposition rate of the organic matter (humidity and high temperature increase the rate of decomposition or decrease the CN ratio). The harmonized kriging map has a stronger correlation than the ordinary kriging map with variables that have not been taken into account in the harmonization process.

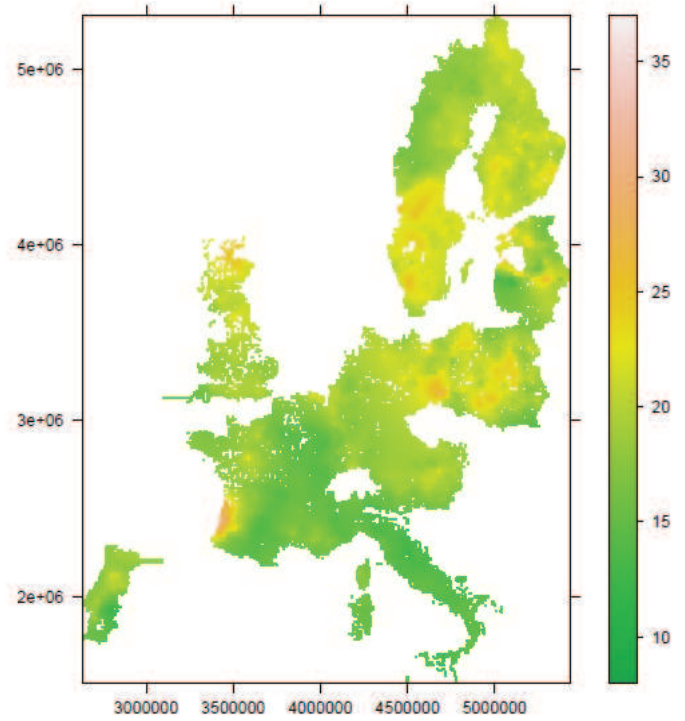
5. CONCLUSION

We proposed a geostatistical modelling solution to the harmonization problem of environmental databases with additive biases. We applied the harmonization methodology to two different datasets. The first example was taken from the problem of radioactivity exposure assessment and shows that the estimation of additive biases in a linear model may yield results similar to a more local procedure, such as the method developed by Skøien et al. [2009]. The second example, dealing with the Carbon/Nitrogen ratio of forest soils indicates that the harmonized map is more correlated to important related factors such as bioclimatic synthetic variables.

As a conclusion, the simple solution that we propose seems to work in different contexts, especially when the number of networks or laboratories is large. The model can be improved with prior information, which gives a guarantee to avoid estimation impairment due to multicollinearity in the linear model. The method will be validated for the Carbon/Nitrogen ratio case study in July 2009 after a second analysis of the measured data by one central laboratory (the French laboratory that had already analyzed the French data).



(a) Ordinary kriging



(b) Harmonized kriging

Figure 1. Carbon/Nitrogen ratio maps of forest soils for parts of Europe.

ACKNOWLEDGMENTS

This work is funded by the European Commission, under the Sixth Framework Program, 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.

REFERENCES

Baume, O., Skøien, J.O., Heuvelink, G.B.M. and Pebesma, E.J., Geostatistical approach to data harmonization. Presented at: *statGIS2007*, Klagenfurt, Austria, 2007.

Baume, O., Skøien, J.O., Heuvelink, G.B.M., and Pebesma, E.J., Data harmonization with geostatistical tools: a Bayesian extension. Presented at: *Geostats2008*, Santiago, Chile, 2008.

Bossew, P., De Cort, M., Dubois, G., Stöhlker, U., Tollefsen, T. and Wätjen, U., AIRDOS, Evaluation of existing standards of measurement of ambient dose rate; and of sampling, sample preparation and measurement for estimating radioactivity levels in air. TREN/NUCL/S12.378241, JRC ref. 21894-2004-04 A1CO ISP BE, 2007.

Brenning, A., Koszinski, S., and Sommer, M., Geostatistical homogenization of soil conductivity across field boundaries, *Geoderma*, 143 (3-4): 254-260, 2008.

Burke, I.C., Yonker, C.M., Parton, W.J., Cole, C.V., Flach, K.D., and Schimel, S., Texture, climate, and cultivation effects on soil organic matter content in U.S. grassland soils. *Soil Sci. Soc. Am. J.*, 53:800-805, 1989.

Carré, F., Jeannée, N., Casalegno, S., Lemarchand, O., Reuter, H.I., and Montanarella, L., Mapping the CN ratio of the forest litters in Europe – lessons for global DSM**.** Presented at: *Logan2008*, Salt Lake City, United States, 2008.

Fassó, A., Cameletti, M., and Nicolis, O., Air quality monitoring using heterogeneous networks, Environmetrics 18: 245-264, 2007.

Schröder, W., Pesch, R., and Schmidt, G., Identifying and closing gaps in environmental monitoring by means of metadata, ecological regionalization and geostatistics using the UNESCO biosphere reserve Rhoen (Germany) as an example, *Environmental Monitoring and Assessment*, 114: 461-488, 2006.

Skøien, J.O., Baume, O., Pebesma, E.J., and Heuvelink, G.B.M., Identifying and removing heterogeneities between monitoring networks. *Environmetrics*, accepted, 2009.

Wagner, G., Desaules, A., Muntau, H., Theocharopoulos, S., and Quevauviller, P., Harmonization 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, 2001.

Wissmann, F., Rupp, A., and Stöhlker, U., *Characterization of Dose Rate*, Zeitschrift Kerntechnik, in Press, 2008.