Appendix 6

Spatial interpolation of settlement-average thyroid doses due to ^{131}I after the Chernobyl accident: 2. Joint modeling of the data in Belarus and Ukraine

Spatial interpolation of settlement-average thyroid doses due to ^{131}I after the Chernobyl accident: 2. Joint modeling of the data in Belarus and Ukraine

A. ULANOVSKY[1,*,✉], R. MECKBACH[1], P. JACOB[1],

S. SHINKAREV[2], I. LIKHTAREV[3], L. KOVGAN[3]

[1]GSF — National Research Center for Environment and Health,
Institute of Radiation Protection, 85764, Neuherberg, Germany
[2]State Research Center — Institute of Biophysics of the Ministry of Health,
48 Zhivopisnaya Str., 123182, Moscow, Russia
[3]Radiation Protection Institute, Ukrainian Academy of Technological Sciences,
53 Melnikova Str., 04050, Kiev, Ukraine

Abstract

A spatial analysis has been undertaken to investigate settlement-average integrated thyroidal activities of ^{131}I in Belarus and Ukraine. The data for the 903 settlements in the both countries are processed, harmonized, and analyzed. The spatial interpolation procedure has been developed and validated through a feasibility study with data on ^{137}Cs ground deposition densities. It is shown that the developed procedure can successfully compensate for a preferential sampling bias evident in the Belarusian part of the data. A spatial trend is estimated by a local spatial regression technique and residuals are interpolated using ordinary kriging procedure. The developed technique is applied to 1247 settlements in Belarus and Ukraine, which have no measurement data on ^{131}I thyroidal activity, and estimations are made for the means and standard deviations of the settlement-average ^{131}I integrated activity. The developed prediction technique has a potential to increase a number of settlements included in a population-based studies of childhood thyroid cancer after the Chernobyl accident.

[*] On leave from: Joint Institute of Power and Nuclear Research – "Sosny", National Academy of Sciences of Belarus, 220109, Minsk, Belarus

✉ Corresponding author.

Tel.: +49-89-3187-2789. Fax: +49-89-3187-3363. E-mail: ulanovsky@gsf.de

INTRODUCTION

The present paper continues an analysis of statistical properties of a spatial distribution of the settlement-average thyroid doses to the members of public in Belarus and Ukraine after the Chernobyl accident. The purpose of the spatial interpolation of these thyroid doses (or relevant dosimetric quantities) is to expand a scope of the population-based study of the childhood thyroid cancer in Belarus and Ukraine after the Chernobyl accident.

The best available estimates of the thyroid doses are based on results of direct measurements of ^{131}I thyroidal content, which took place in May-June 1986 in both countries. Such dose estimates are referred in the paper as "measured" ones. The population-based risk assessment study deals with "measured" thyroid doses, which are known only in a limited set of settlements. Then, an application of the spatial statistics techniques to interpolate between the "measured" data has a potential to provide estimates of the average thyroid doses in the "non-measured" settlements. The companion paper [1] provides a motivation of the study and discusses results of a feasibility study for Belarusian data.

The feasibility study was undertaken to investigate capabilities of the geostatistical techniques for interpolation of the thyroid doses based on direct measurements of ^{131}I thyroidal content. Namely, the feasibility study dealt with data ^{137}Cs ground deposition instead of the thyroid doses to ^{131}I. The study included the Belarusian data as these are known to be measured mostly in highly contaminated settlements, thus resulting in a biased "measured" sample and leading to a danger of systematic overestimation bias in the interpolated data.

The feasibility study [1] proved to be successful for Belarusian settlements located around the Chernobyl power plant and it revealed a deficiency of the predictions for settlement located in Mogilev oblast and in the northern part of Gomel oblast. Consequently, only the former, together with the Ukrainian data, are analyzed in the present paper.

MATERIALS AND METHODS

Description of the data

Belarus

Based on the results of the feasibility study [1], the settlements from the South-Eastern part of the Gomel oblast are included in the current study. Defined as "measured" are 308 settlements with more than 10 individual thyroid measurements.

Interpolation has been performed not with the average thyroid doses, which are age-dependent, but rather with a derived quantity – weighted average integrated thyroidal activity of ^{131}I – *g*. This quantity is defined through the following factorization [2] of the settlement-average thyroid dose for the age a :

$$D(a) = C_D f(a) g \qquad (1)$$

where

C_D = dose-conversion coefficient, Gy kBq^{-1} d^{-1};

g = weighted average integrated thyroidal activity of ^{131}I, kBq d;

$f(a)$ = dimensionless average age-dependence of the integrated activity of ^{131}I.

The function $f(a)$ is normalized to sum up to one for ages up to 18 years.

The average age dependence, $f(a)$, is assessed for the Belarus using data on the individual measurements of ^{131}I in the thyroid. Typically, the Belarusian individual measurements data bear no gender information. For some individuals such information can be derived or implied from the last name, however, for a significant part of the measured individuals no clue exists to help in determining the individual's gender. Therefore, the function $f(a)$ has been derived as a gender-average for urban and rural settlements, separately [2]. The arithmetic mean (AM) and standard deviations (SD) of the g along with coordinates of the settlements have created a data set for spatial interpolation.

As target settlements, i.e. those "non-measured" where prediction are to be made, 654 settlements have been selected, which located within 30-*km* range from any sample (i.e. "measured") settlement.

Ukraine

Ukrainian individual data were resulted from the factorization procedure [3] similar to one mentioned above for Belarus. However, some assessment details are

different than those for Belarusian data. First, the Ukrainian data are gender-dependent. Then, the meaning of the g is a geometric mean (GM) of the integrated thyroidal activity of ^{131}I for the reference age-group, which spans from 12 to 14 years. The corresponding uncertainty is represented as a geometric standard deviation (GSD). Some settlements count less than 10 individual measurements if the measurement results were regarded as high-quality and reliable.

To use the data for Belarus in Ukraine simultaneously in a spatial analysis it is necessary to harmonize them. For this, the Ukrainian data have been recalculated from GM–GSD to AM–SD values. Then, the gender-dependent AM and SD values have been averaged between genders accounting for the sample sizes of the gender subgroups:

$$\bar{g} = \frac{n_m \bar{g}_m + n_f \bar{g}_f}{n}$$
$$\sigma_g^2 = \frac{1}{n(n-1)}\left(n_m \bar{g}_m^2 + n_f \bar{g}_f^2 - n\bar{g}^2 + (n_m - 1)\sigma_m^2 + (n_f - 1)\sigma_f^2\right) \qquad (2)$$

where

$\bar{g}$ and σ_g^2 = gender-independent mean and variance for the total population,

$\bar{g}_m$ and σ_m^2 = mean and variance of the male sub-group,

$\bar{g}_f$ and σ_f^2 = mean and variance of the female sub-group,

n_m, n_f, n = size of male, female, and total population in a settlement.

Those settlements, where the total number of measured individuals of both genders exceeded 10, have created Ukrainian part of the sample data set. There are 593 of them. Targets settlements in Ukraine were selected on the basis of administrative sub-division. This resulted in 563 target settlements in the Ukraine.

Finally, the Belarusian data are re-scaled to fit the definition of the *g*-values in Ukrainian data. Namely, the age-dependence function in Belarus has been re-normalized to be equal to one for the reference age-group. Then, the *g*-values being multiplied by the renormalization factor resulted in a data set compatible with the Ukrainian data.

Spatial distributions of the data points

Spatial distribution of the sample settlements is presented in Fig. 1, which shows spatial locations and values of the settlement-average ^{131}I thyroidal activity for the reference age group — *g*. Points in the figure are drawn in rectangular

coordinates obtained from the geographical ones by special transform based on Transverse Mercator projection.

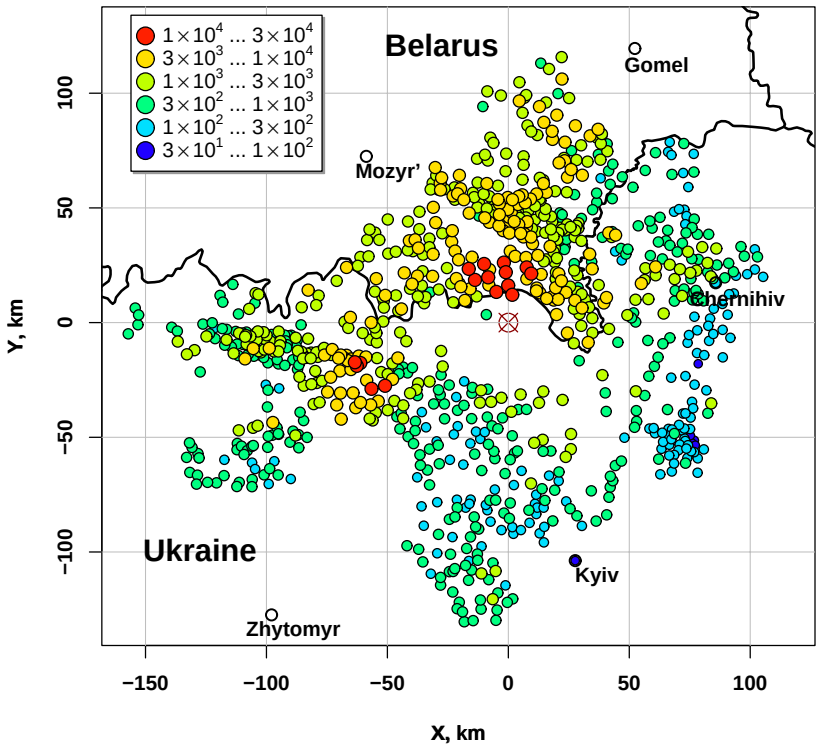

Fig. 1. Sample location map. Distribution of settlement-average ^{131}I thyroidal activity for the reference age group in sample settlements in Belarus and Ukraine.

From the Fig. 1 one can see that the *g*-values for Belarusian settlements are practically missing low values (light blue and blue colors in the picture). Indeed, this reflects the preferential sampling specific for Belarusian data, i.e. the fact that thyroid measurement campaign had been conducted in the most contaminated places based, mainly, on an exposure rate in air. This resulted in apparently biased sample in Belarus. Ukrainian data do not show evidence of such preferential sampling (cf. Fig. 2). In the Fig. 2, the probability plots are given for *g*-values in both countries separately and in the combined sample.

To derive a tendency from the sample data, a non-standard, combined procedure for spatial interpolation is developed. At the first stage, an attempt is made to reconstruct a spatial trend using a technique of local spatial regression on the sample points. Then, the regression residuals are analyzed and found to be spatially correlated, and the classical kriging methods are used to predict unknown values.

The feasibility study with ^{137}Cs data [1] has shown that the combined interpolation procedure might work for ^{131}I integrated activity data in Belarus. Results of the feasibility study demonstrated satisfactory quality of prediction for the settlements in South-Eastern part of the country. The bias due to preferential sampling is compensated by the selected trend model. Predicted uncertainties are validated through extensive cross-validations and found to adequately represent uncertainties derived from a comparison of the predicted and true values.

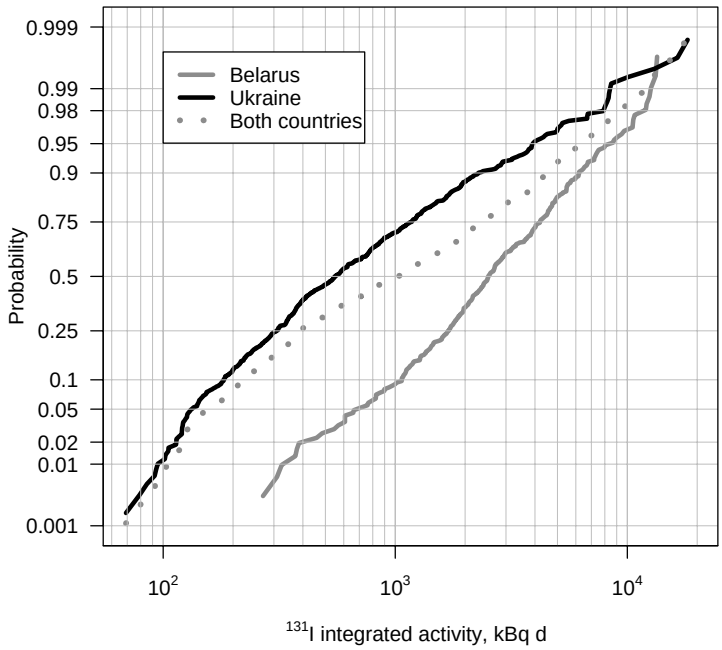

Fig. 2. Probability distributions of settlement-average ^{131}I thyroidal activity for the reference age group in sample settlements in Belarus and Ukraine.

REPRESENTATION OF THE DATA

The sample data, $\{g_i : i = 1,\ldots,n\}$, are considered as realizations of a spatial random process, $G(x)$, in the sample points, x_i, located in a two-dimensional spatial domain. That is, x_i is a set of vectors. Because of apparent log-normality, transformed data, $\tilde{g} = \ln(g)$, are considered as realizations of a Gaussian spatial process $\tilde{G} = \ln(G)$. The sample data demonstrate both systematic behavior and random fluctuations, thus the following model of the random process is assumed

$$\tilde{G}(x) = m(x) + Z(x) = m(x) + Y(x) + \varepsilon, \qquad (3)$$

where $m(x)$ represents non-stochastic spatial component of the random process $\tilde{G}(x)$ and called hereafter "trend"; $Z(x)$ is a stochastic part of the process, which can be separated into correlated and non-correlated components, $Y(x)$ and ε, respectively. Variance of the non-correlated component, $\mathrm{var}(\varepsilon) = \tau^2$, is called *nugget* in the geostatistical literature and can be interpreted as a combined result of micro-scale variations and a measurement error.

Trend modeling

Trend modeling has been performed using LOESS method [4]. Two most important parameters are the span, α, and a degree of regressing polynomials, L. There exist no unique quantitative recommendations on appropriate selection of these parameters ([5], p. 437). More details on the trend modeling technique used in the present study can be found in the companion paper [1].

The Fig. 3 presents the trend model as a contour plot. Presented in the figure are sample data, also. The trend model is estimated in the areas where the selected target settlements are located. This is done to avoid any meaningless extrapolation of the locally defined trend.

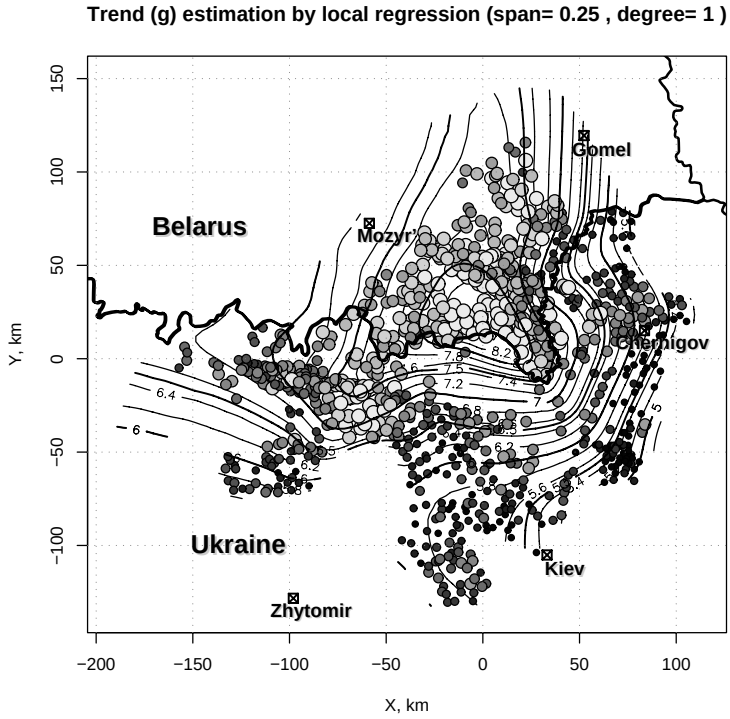

Fig. 3. Estimation of trend by local regression. Points are sample settlements. Color and size of points represent deciles of sample distribution.

Modeling of the residuals

Residuals after fitting the sample data by the local trend model are analyzed using classical geostatistical approach — ordinary kriging (see e.g. [6–11]). The software used for this purpose are GEOR library [12] and R [13]. The modeling details can be found in the companion paper [1].

Presented in the Fig. 4 are variograms of the residuals corresponding to the different trend models. The trend models have different degree of the 'locality', i.e. they differ by the *span* value. Reduction and stabilization of the variogram *sill* demonstrates that trend models spatial variations and leaves spatially correlated residuals.

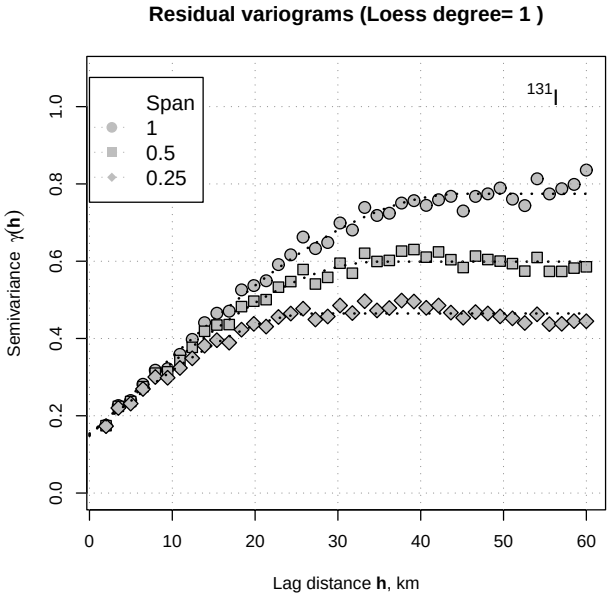

Fig. 4. Comparison of the variograms obtained for the trend models of the first order varying by the span value. Effect of the variogram sill reduction and stabilization is seen as span decreases.

The residuals are spatially correlated and distributed normally. These authorize them for the ordinary kriging procedures.

Prediction by kriging

Prediction of the *g*-values in the target settlements have been done using the local regression to derive the spatial trend and subsequent ordinary kriging of the residuals. Details of this technique can be found in [1].

Distributions of the predicted and the sample values are given in Fig. 5. From the figure one can see that range of the predicted values is less than that of the sample ones for both countries. This is mainly due to the fact that the local regression and the kriging are smoothing procedures and they filter out pure stochastic, non-correlated part of the data variance. This stochastic part is expressed by the variogram *nugget* and corresponds to GSD being equal to approximately 1.5. The nugget is commonly interpreted as a combined effect of micro-scale spatial variations and measurement errors.

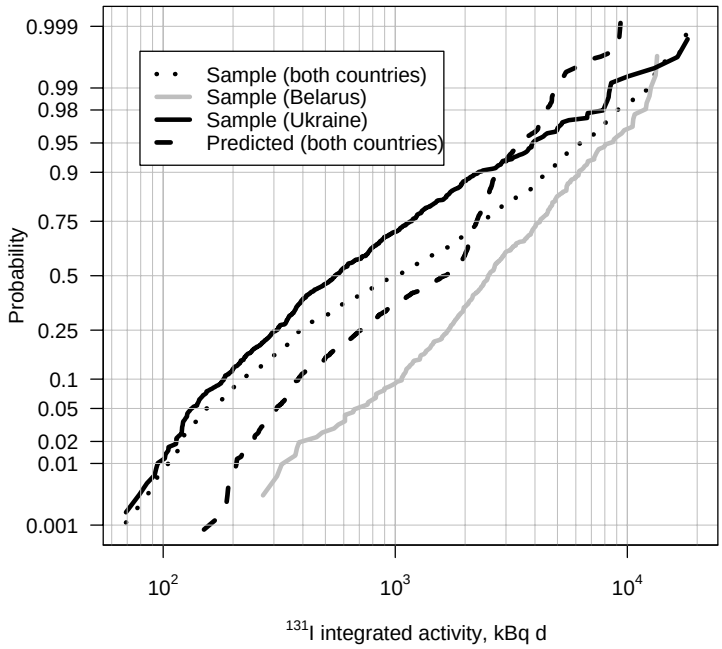

Fig. 5. Distributions of the predicted and sample values.

In the Fig. 6 one can see a location map of the sample and the predicted values. Comparison of the map with the sample location map (Fig. 1) reveals that spatial distributions of the target settlements in Belarus and Ukraine are different. Namely, a majority of the Belarusian target settlements located outside the area covered by sample points, i.e. spatial predictions in the Belarus fall mostly in an extrapolation case. Unlike Belarus, in Ukraine the target points are distributed more homogeneous between the sample ones, thus these are mostly interpolation cases.

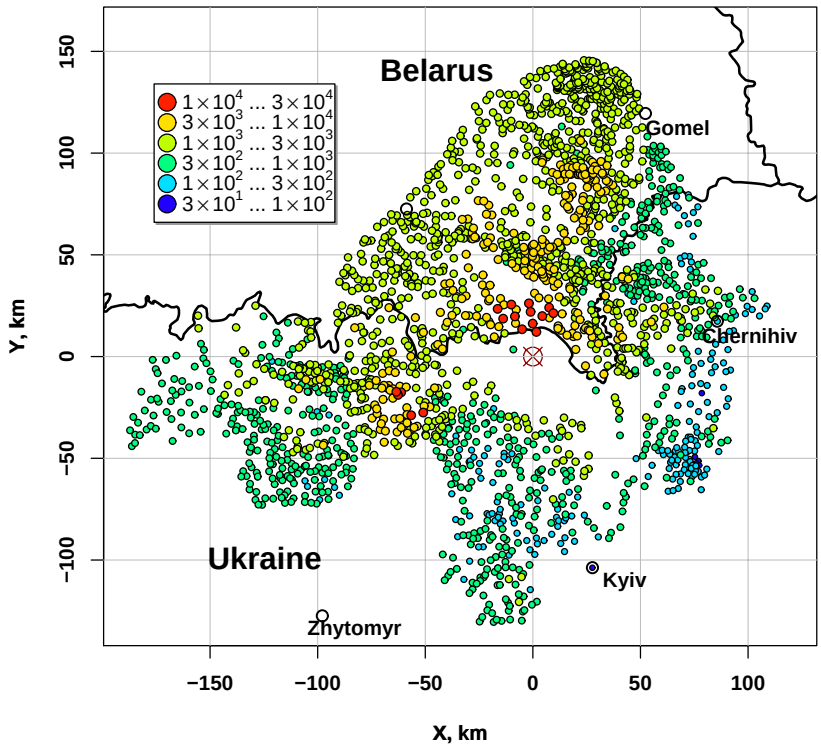

Fig. 6. Map of predicted and sample data

Uncertainty of the predicted values is expressed by kriging variance. The kriging variance in a prediction point is bounded between the nugget and the sill. Therefore, the real uncertainty of the prediction coincides with the kriging variance if the trend model adequately reproduces local mean value in the given prediction point. Otherwise, inadequate trend model would introduce unknown systematic error. Due to this, the whole spatial interpolation technique is adequate within a certain spatial area, which can be characterized by the range of the variogram. Correspondingly, kriging variance is also an adequate measure of the prediction uncertainty within the variogram's range.

Cross-validation

Cross-validation has been performed using the sample data and a "leaving-one-out" resampling technique. For this, a point is removed from the sample dataset and a prediction is made for the removed point based on the rest of the sample. This procedure is repeated for every sample point and then one can compare the distributions of the true sample and the predicted sample values.

For comparison purposes, the sample points, where the cross-validation is taking place, are characterized by minimum distance to any other sample point, $H_{\min}$. This criteria is not sufficient to unambiguously distinguish between inter-

and extrapolation cases, however it still provides some clue for this. Namely, one can say that the target points with higher $H_{\min}$ are more likely to fall in extrapolation case and *vice versa*.

The true sample and predicted sample data are compared in the Fig. 7. One can see fairly good agreement between these, however the scatterplot implies somewhat larger predicted values for the low sample ones, particularly in the range of the true values less than 3×10^3 kBq d.

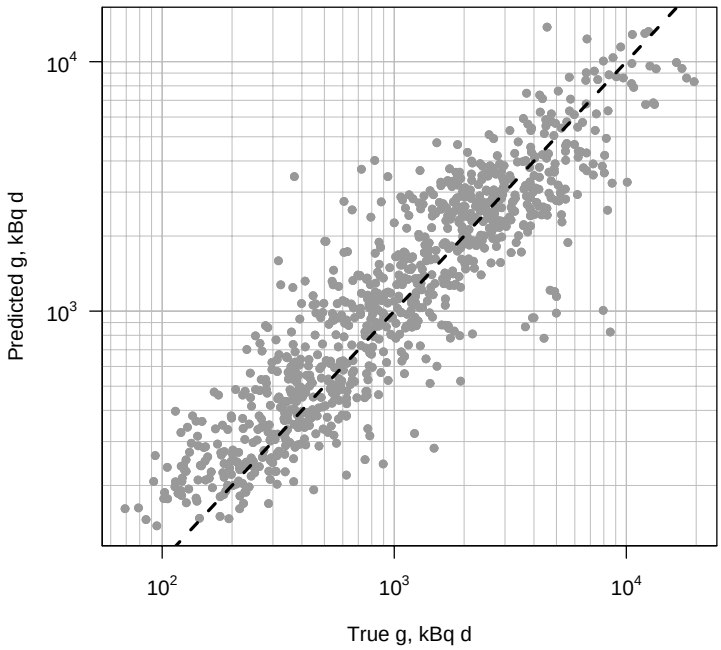

Fig. 7. Cross-validation results. Comparison of the predicted and the true sample values. The dashed line represent a case of equality of the true and the predicted values.

To characterize the cross-validation results in a more quantitative mode the *standardized error* is used (see e.g. [11, 12]). The standardized error

$$\delta_{std} = \frac{q_{pred} - q_{true}}{\sigma}$$

assumes a normal distribution for a prediction around the true value, i.e. $\delta_{std} \propto N(0,1)$. It is shown above (see e.g. Fig. 2) that the sample data are apparently log-normal. That means that another quantity

$$\delta^*_{std} = \frac{m^* + z_K - \ln g_{true}}{\sigma_K} = \ln\frac{g_{pred}}{g_{true}} - \frac{1}{2}\ln\left(\frac{\sigma^2}{g^2_{pred}} + 1\right) \qquad (4)$$

has the standard normal distribution (see notation details in [1]). The latter quantity is plotted against g_{true} and $H_{\min}$ in the Figs. 8 and 9. The values of standardized error (4) are plotted as a point cloud in these figures. The solid line represents local estimate of the mean, while dashed lines correspond to local mean ± local standard deviation. Deviation of the local mean values from the zero line signals an existence of a bias in the predictions and serves as an indicator of a systematic error. If the local standard deviations exceed one this means the estimated kriging error does not represent the real error of the prediction.

The standardized error plotted against the true sample values in the Fig. 8. This figure illustrates the smoothing effect of the spatial interpolation techniques mentioned above. Namely, the predicted values apparently underestimate in the region of the true values more than 3×10^3 kBq d. On the other hand, the values less than 3×10^2 kBq d are overestimated.

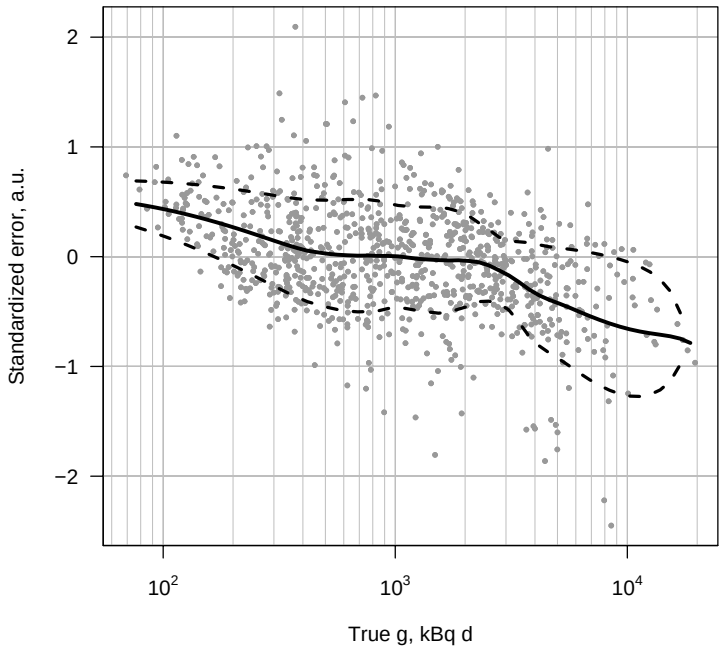

Fig. 8. Cross-validation results. Standardized errors vs. the true sample values.

One can not reveal any meaningful tendency while comparing the standardized errors as a function of $H_{\min}$ (see Fig. 9). This could be a result of dense location of the sample points and correspondingly low values of the $H_{\min}$.

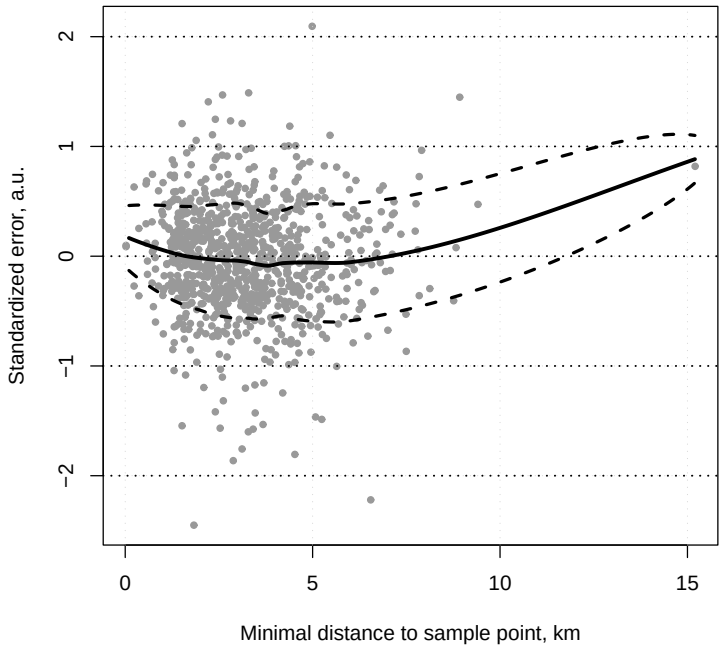

Fig. 9. Cross-validation results. The standardized errors vs. $H_{\min}$.

CONCLUSIONS

The spatial analysis has been performed on the ^{131}I integrated thyroidal activity data in the Ukrainian and Belarusian settlements. The analysis has shown that a special, non-standard technique is needed to interpolate between the data and to account for apparent preferential sampling bias in the Belarusian data.

The data for the both countries are harmonized and found to be consistent to each other despite of the significant differences in measurement and dose assessment techniques.

The developed technique is applied to the data, which are based on direct historical thyroid measurements in 903 settlements, and resulted in the predictions for the 1247 target settlements. The spatial interpolation procedure allowed to assess mean values as well as standard deviation of the means.

The uncertainty of the predicted values is bounded between the nugget and the sill of the residual variogram, therefore uncertainty estimates are safe within only limited spatial scale, typically within the range of correlation of the sample data (i.e. range of the residual variogram). Care must be taken while selecting the trend model to assure: (1) normality of the residuals; (2) stability of the trend; (3) realism of a prediction of the systematic features seen in the data pattern; (4) meaningfulness of the prediction uncertainty estimates; (5) reasonable range of the prediction extrapolations.

Acknowledgment

This work has been supported by the German Federal Ministry of Environment, Nature Preservation, and Reactor Safety and the German Federal Office of Radiation Protection under the contract No. StSch 4240.

REFERENCES

1. Ulanovsky A, Meckbach R, Jacob P, and Shinkarev S, Spatial interpolation of settlement-average thyroid doses due to ^{131}I after the Chernobyl accident: 1. Feasibility study with ^{137}Cs deposition data in Belarus. — This report. Appendix A5.
2. Post-Chernobyl thyroid doses in Belarus based on measurements of the ^{131}I activity in the human thyroid and on a factorization method. – This report. Appendix A3.
3. Likhtarov I, Kovgan L, Vavilov S, et al. Post-Chernobyl thyroid doses in Ukraine. — This report. Appendix A1.
4. Cleveland W, Grosse E (1991) Computational methods for local regression. Statistics and Computing 1: 47–62
5. Venables WN, Ripley BD (2000) Modern Applied Statistics with S-Plus. Third edition. Springer, New York
6. Journel A, Huijbregts C (1978) Mining Geostatistics. Academic Press, London
7. Isaaks E, Srivastava R (1989) An Introduction to Applied Geostatistics. Oxford University Press, Oxford
8. Cressie N (1993) Statistics for spatial data, Revised Edition. Wiley, New York
9. Wackernagel H (1995) Multivariate Geostatistics. An Introduction with Applications. Springer, Berlin
10. Webster R, Oliver M. (2001) Geostatistics for environmental scientists. Wiley, Chichester
11. Chilès JP, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York
12. Ribeiro Jr PJ, Diggle PJ (2001) geoR: A package for geostatistical analysis. R-News 1 No. 2: 15–18 (ISSN 1609-3631)
13. Ihaka R, Gentleman R (1996) R: A Language for Data Analysis and Graphics, Journal of Computational and Graphical Statistics 5: 299–314