Exploring Spatial Patterns of Uranium Distribution in Ukraine

Michael Govorov
Vancouver Island University
900 Fifth Street
Nanaimo, Canada
govorovm@viu.ca

Viktor Putrenko
Institute of Geography of the National
Academy of Sciences of Ukraine
44 Volodymyrska Street
Kiev, Ukraine
putrenko@rambler.ru

Gennady Gienko
University of Alaska Anchorage
3211 Providence Drive
Anchorage, USA
ggienko@uaa.alaska.edu

Abstract

Geovisualization and spatial statistical methods are widely used in environmental studies. The paper illustrates the use of several modelling techniques to reveal spatial patterns in distribution of uranium in groundwater in Ukraine. A comparative study of factor, correlation, and regression analyses, including their spatial implementations, has been carried out to describe the impact of several environmental variables on spatial distribution of uranium. Local factor analysis (or geographically weighted factor analysis, GWFA) was also proposed to supplement the study.

Keywords: Spatial statistics, geovisualization, geographically weighted analysis

1   Introduction

The problem of drinking water is relevant for many regions, including Ukraine. Quality of drinking water, determined by its chemical and biological content, depends on several factors, including the presence of radioactive elements such as uranium. Uranium concentration higher than 0.08 Mg/L is potentially dangerous to human health. Major environmental and natural indicators affecting concentration of uranium in surface waters can be grouped into four categories: geological structure and geomorphology, geochemical, climate, and mineralogical [1]. The goal of this study is to explore various geostatistical methods to model spatial dependences between several environmental variables and distribution of uranium in surface waters. The study is focused on comparative analysis of several techniques to identify the most robust method to describe spatial distribution of uranium in surface waters Ukraine. The analysis was implemented using tools available in SPSS and ArcGIS packages.

2   Methodology

The study is based on the results of geological surveys in Ukraine carried out by State Enterprise "Kirovgeologiya" [2, 3]. The database consists of 23 environmental indicators collected in 9353 sample points in Ukraine and neighbouring territories of Russia, Belarus, and Moldova. Not all environmental indicators are available for all neighbouring territories, limiting the study by using the complete dataset only available for Ukraine (6546 sample points, Table 1).

The following workflow was exploited to describe the impact of several environmental variables on spatial distribution of uranium. Note on terminology: the term 'natural variables', or 'environmental indicators', is widely used in this study to describe various natural and environmental factors, for example, mineralization, precipitation, relief, and many others. Some models use 'predictors' with the same meaning as 'variables'.

1.   Exploratory spatial data analysis (ESDA) is used for initial data analysis, such as check for statistical distribution, linearity, multicollinearity, and presence (or absence) of a pattern (both in spatial and non-spatial domains). Based on initial hypotheses and results of ESDA, further statistical data exploration was developed by building correlation, regression, and factor models.

The histogram of uranium concentration clearly shows that statistical distribution of the source data does not meet criteria for normality. However, logarithmic transformation brings the dataset closer to the normal.

The curve estimation procedure shows that the relationship between uranium and mineralization of water is more exponential than linear, which would require nonlinear regression modelling. However, implementation of multiple nonlinear regression for these variables can be problematic.

Spatial autocorrelation methods were used to identify patterns in spatial measurements of uranium concentration. According to Moran's I and Getis-Ord analysis [4], the distribution of uranium can be described as highly clustered with statistical significance. Moran's I index is 0.5475 (p-value = 0.0) and Observed General G = 0.00007 (p-value = 0.0). There is positive spatial autocorrelation as Moran's I is positive and greater than expected IndexE(I) = -0.00015. Thus, spatial patterns in observations of uranium should be taken into account in correlation, regression, and factors models to avoid bias due to over-counting [5].

2.   In the second step, quantitative measure of global correlation is used to confirm or reject several hypotheses of relationship between the dependent variable (uranium) and independent variables. The analysis should identify natural variables which define high concentration of uranium, taking into account multicollineaity of the data.

The classical global Pearson correlation coefficient r is used as a measure of global correlation. The coefficient is defined as $r = \dfrac{\sum_{i=1}^{n}(u_i - \bar{u})*(v_i - \bar{v})}{(n-1)*S_u*S_v}$ (1), where r is the correlation coefficient; u_i and v_i are the individual observations; $\bar{u}$ and $\bar{v}$ are the means of the two variables; n is the sample size; and S_u and S_v are the standard deviations of the two variables.

3.   In the next step, factor analysis is used to identify a smaller number of natural variables that define most of the

variance of uranium distribution. This study utilizes factor analysis based on the principal components using Varimax rotation with Kaiser normalization [6].

In the factor analysis model based on factor p is $\Lambda_p = \Omega_p \Gamma_p^{1/2}$ (2), where Λ_p is the matrix of factor loadings based on factor p; Ω_p is the diagonal matrix of ϖ_i that are the corresponding eigenvectors of $\boldsymbol{R}$, where $\boldsymbol{R}$ is the correlation $n \times n$ matrix or eigenvectors of Σ, where $\Sigma = [\sigma_{ij}]_{n \times n}$ (3) is the covariance $n \times n$ matrix of natural variables; and Γ is the matrix of eigenvalues. Here, n represents the number of variables and p is the number of factors.

After a varimax rotation, each original variable tends to be associated with one (or a small number) of components, and each component represents only a small number of variables. This simplifies interpretation of resultant factor and their associations with the variables. Varimax searches for a linear combination of the original factors such that the variance of the squared loadings is maximized, which amounts to maximizing [6].

4. The most significant natural variables, identified from factor and global correlation analysis, were used to demonstrate spatial nonstationarity of correlation between uranium and natural variables. This was done by implementing local correlation analysis (or geographically weighted correlation, GWC). In this study, a local form of bilinear regression with the optimized bandwidth was used to model spatially varying relationships between uranium and natural variables.

The geographically weighted local Pearson correlation coefficient $r_i(x_k, y_k)$ is used and defined as

$$r_i(x_k, y_k) = \frac{\sum_{j=1}^{n} \varpi_j * (u_j - \bar{u}) * (v_j - \bar{v})}{(n-1) * S_u * S_v} = \beta_1(x_k, y_k) \frac{S_u}{S_v} \quad (4)$$

where $r_i(x_{ki}, y_k)$ is the correlation coefficient, (x_k, y_k) is the location of observation i; u_i and v_i are the individual observations; $\bar{u}$ and $\bar{v}$ are the means of the two variables, ϖ_j is the weight assigned to each observation that based on a distance decay function centered on the observation i; n is the sample size; and S_u and S_v are the standard deviations of the two variables; $\beta_1(x_k, y_k)$ is the estimated parameter for the bivariate local least-squares regression on observation i.

5. Relationships between the dependent variable and all explanatory variables were modelled by using stepwise linear multivariate regression analysis (or ordinary least squares regression, OLSR). Then, several most significant explanatory variables identified from the stepwise multivariate regression analysis were used to build local linear multivariate regression models (or geographically weighted regression, GWR) with the optimized bandwidth [7].

Explanatory variables in spatial regression model have consistent relationship with the dependent variable both in geographic space and in data space. The requirements of linear multiple regression have to be addressed [8].

The OLS global regression formula applied to the explanatory independent variables x_p to best predict the dependent variable y and defined as $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_p * x_p + \varepsilon$ (5), where y is the dependent variable, x_p are the explanatory variables, β_p are global regression coefficients, p is number of variables, and ε is random error or residual.

A stepwise approach was used to build multivariate regression models. At each step, the independent variable not included in the equation had the smallest probability of F-statistic, if that probability is sufficiently small. The method terminates when no more variables are eligible for inclusion or removal. The AIC and *adjusted* R^2 are used to estimate the models' performance. The method also insures that VIF among explanatory variables are low to avoid multicollinearity among explanatory variables.

Explanatory variables in spatial regression model have consistent relationship with the dependent variable both in geographic space and in data space. The requirements of linear multiple regression have to be addressed [7]. There is a possibility to improve model results by applying local Geographically Weighted Regression (GWR), which takes into account effect of heteroskedasticity [7].

The geographically weighted local regression [7] defined as $y_i(x_k, y_k) = \beta_{0i}(x_k, y_k) + ... + \beta_{pi}(x_k, y_k) * v_{pi} + \varepsilon_i$ (6) on observation i, where $y_i(x_k, y_k)$ is the dependent variable estimated in location i; x_p are the explanatory variables; $\beta_p(x_k, y_k)$ are the local regression coefficients; p is the number of variables, and ε_i is the residual estimated in location i. Each local regression $y_i(x_k, y_k)$ equation is solved using a different weighting of the observations that based on a distance decay function cantered on the observation i.

GWR accounts for spatial nonstationarity in parameter estimates, but it does not directly address autocorrelation [9].

6. Finally, local factor analysis (or geographically weighted factor analysis, GWFA) with the optimized bandwidth [10], was proposed to generate linear multivariate regression models for the dependent variable (uranium) and six major factors produced in the global factor analysis.

The geographically weighted local factor/regression models defined as $y_i(x_k, y_k) = \beta_{0i}(x_k, y_k) + ... + \beta_{pi}(x_k, y_k) * F_{pi} + \varepsilon_i$ (7) on observation i, where $y_i(x_k, y_k)$ is the dependent variable estimated in location i; F_p are the factors from principal components extraction; $\beta_p(x_k, y_k)$ are the local regression coefficients; p is the number of variables, and ε_i is the residual estimated in location i. Each local regression $y_i(x_k, y_k)$ equation is solved using a different weighting of the observations.

3 Implemention

3.1 Global and Local Spatial Correlation Analysis

The most significant factors can be identified using global correlation of uranium with all indicators from the four

defined groups: geological, geochemical, climatic, and mineralogical (See *Global r* in Table 1).

The highest global correlation coefficients of uranium were obtained for humus (r=0.52), temperature (r=0.51),

3.2 Global Factor Analysis

Factor analysis has been used to find the input of a particular variable into distribution of uranium. Analysis of 23

Table 1: Global correlation and principal components of environmental indicators.

| | Global r | Component | | | | | |
		1	2	3	4	5	6
Explained variance (cumulative) %		30.2	44.2	54.1	62.3	68.0	72.5
SO_4	.448	0.952	0.139	-0.092	-0.045	0.1	0.091
Mineralization of water	.493	0.945	0.108	-0.159	-0.003	0.132	0.154
Hardness of water	.493	0.937	0.035	-0.173	0.01	0.062	0.172
Cl	.445	0.918	0.07	-0.147	0.016	0.148	0.128
NO_3	.269	0.629	-0.177	0.198	0.368	0.354	-0.061
Cu	-.067	0.027	0.886	0.101	0.074	0.043	0.038
Fe	-.148	0.012	0.851	0.286	-0.002	0.07	-0.01
Mn	.118	0.392	0.716	0.008	-0.091	0.088	-0.004
Zn	.007	-0.101	0.608	-0.417	0.103	-0.096	0.103
Precipitation	-.497	-0.328	0.243	0.762	0.16	-0.156	-0.276
Relief	-.244	-0.148	0.077	0.703	0.074	-0.014	-0.064
Slope	-.078	-0.011	-0.051	0.617	-0.139	0.235	0.091
Temperature	.507	0.434	-0.347	-0.521	-0.332	0.293	0.217
NH_4	-.117	-0.025	0.036	0.165	0.816	-0.162	-0.037
NO_2	.288	0.525	-0.009	-0.126	0.67	0.228	0.062
PO_4	.166	0.04	0.211	-0.327	0.627	0.331	0.199
Cr	-.390	-0.346	-0.228	0.25	0.589	-0.485	-0.215
Isopach	.023	0.13	0.083	0.287	0.071	0.704	-0.146
Humus	.521	0.372	0.058	-0.075	-0.021	0.648	0.306
Volume of natural groundwater resources	-.558	-0.431	0.26	0.384	0.264	-0.507	-0.29
HCO_3	.160	0.129	-0.017	0.077	0.038	-0.23	0.784
F	.353	0.207	0.048	-0.228	-0.039	0.128	0.546
As	.229	0.032	0.052	-0.044	0.023	0.298	0.473

precipitation (r=-0.50), and volume of natural groundwater resources (r=-0.56). Uranium also has significant correlation with the overall water mineralization (r=0.49) and its components: SO_4 (r=0.45), Cl (r=0.44), and hardness of water

environmental indicators revealed six principal components (Table 1).

The analysis shows that all six identified principal components largely coincide with the four groups of natural

Figure 1: Uranium (left), water mineralization (center), and distribution of local correlation (right). Global correlation r=0.49.
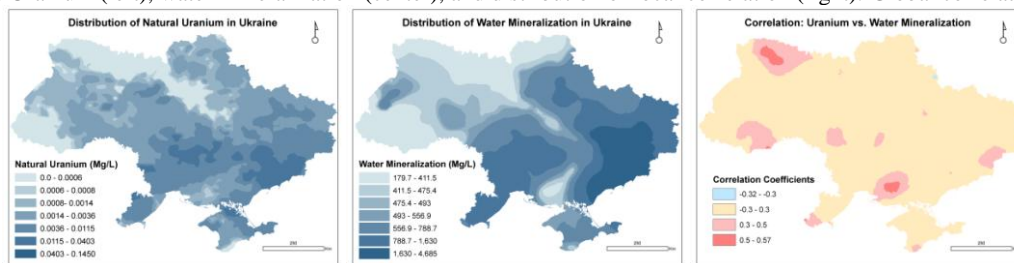
(r=0.49). These indicators are inter-dependent and highly correlated.

Local correlations for different indicators can form complex spatial patterns and anomalies. For example, Moran's *I* and Getis-Ord analyses indicate that the pattern of uranium is highly clustered. Thus, values of correlation coefficients inherit high nonstationarity and should be modelled by using local methods. For example, uranium and isopachs exhibit very low global correlation (only r=0.02), but coefficients of local correlation range from -0.71 up to 0.42 that shows very high associations between these two variables on the local level.

variables, outlined in the hypothesis of distribution of uranium in Ukraine (Table 2).

3.3 Local Spatial Correlation Using Principal Components

Local correlation analysis has been carried out for the most significant elements from each component (highlighted in Table 1). Figures below provide examples of local spatial correlation using principal components: mineralization of water from component 1 (Figure 1), and temperature and relief from component 3 (Figure 2).

Overall, the correlation analysis highlights significant global association between uranium and water mineralization with from -0.71 to 0.42 for isopachs (thickness of sedimentary rocks).

Figure 2: Uranium (left), relief (centre), and their correlation (right). Global correlation r=-0.24.

strong local correlation in certain areas in Ukraine. Strong relationship is defined mostly by climate parameters, geological formation, and distribution of large river basins.

3.4 Local Spatial Multiple Regression

Table 2: Environmental indicators (predictors), selected for multiple regression analysis.

Component	Natural Variables	Group	Predictors
1	Mineralization and hardness of water	Mineralogical	Hardness of water, mineralization of water
2	Metals dissolved in water	Geochemical	Cu, Fe, Cl, Zn
3	Climatic conditions of territory and formation of ground water	Climatic	Precipitation, temperature, humus
4	Organic compounds in water	Mineralogical	NO_3, NH_4, PO_4
5	Geomorphological characteristics	Geological structure and geomorphology	Relief, isopach
6	Mineral compounds and satellite elements of uranium	Geochemical	Bicarbonate, fluoride, arsenic

Among the natural variables, precipitation and temperature play important roles, but they are highly inter-correlated. Temperature shows the highest global correlation coefficient (r=0.51) with the highest values in local correlation (r=-0.40…0.60).

Geomorphological factors have low global correlation

Contribution of each factor in distribution of uranium can be better understood by carrying out multiple regression analysis. Eighteen multiple regression models were built incrementally using predictors outlined in Table 3. The first five most significant predictors (precipitation, humus, water hardness, F, and Fe) contribute 40.2% into the overall model. Local r-

Table 3: Multiple regression model summary.

| Model	Predictors	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	precipitation	0.51	0.26	0.26	1.1611
2	1 + humus	0.595	0.354	0.354	1.0849
3	2 + water hardness	0.623	0.388	0.387	1.0564
4	3 + F	0.629	0.396	0.395	1.0496
5	4 + Fe	0.634	0.402	0.402	1.0438
6	5 + As	0.637	0.405	0.405	1.0414
7	6 + SO₄	0.638	0.407	0.407	1.0395
8	7 − water hardness	0.638	0.407	0.407	1.0396
9	8 + isopach	0.641	0.411	0.41	1.0368
10	9 + NH₄	0.642	0.413	0.412	1.0351
11	10 + Cl	0.644	0.415	0.414	1.0332
12	11 + temperature	0.645	0.416	0.415	1.0325
13	12 + NO₃	0.645	0.417	0.416	1.0317
14	13 + HCO₃	0.646	0.418	0.416	1.0310
15	14 + Zn	0.647	0.418	0.417	1.0305
16	15 + Cu	0.647	0.419	0.418	1.0300
17	16 + PO₄	0.648	0.419	0.418	1.0296
18	17 + mineralization	0.648	0.42	0.419	1.0288

coefficients, but could be well used to predict the phenomenon at the local level. The local correlation coefficients range from -0.47 to 0.71 for relief (Figure 2), and

square coefficients for multiple regression model, built only on precipitation variable, are shown in Figure 3, left. Adding the humus component and then hardness of water improves the model.

Ftor (F), iron (Fe), and arsenium (As) add only 1.7% in variability of the data. None of the rest of environmental variables contributes more than 0.2% into the final model.

correlation for the model, based on the principal components, is only 0.48, while the same correlation for the latter model is 0.51. This indicates that the zones of high regression are

Figure 3: Local R^2-square coefficients for geographically weighted regression models: precipitation (left), precipitation + humus (centre), precipitation + humus + water hardness (right).
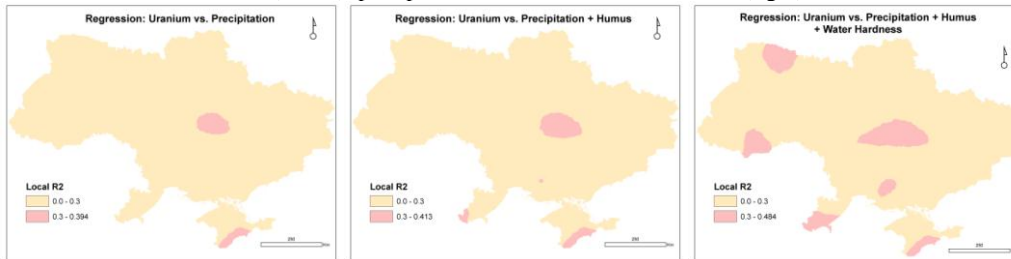

3.5 Multiple Regressions Based on Principal Components

An alternative approach for multiple regression is to build the models using principal components. All variables constituting

identified objectively, and the model of the five environmental variables shows stronger local associations comparing to the component model which takes into account all studied indicators.

Figure 4: Uranium (top), and six local principal components.
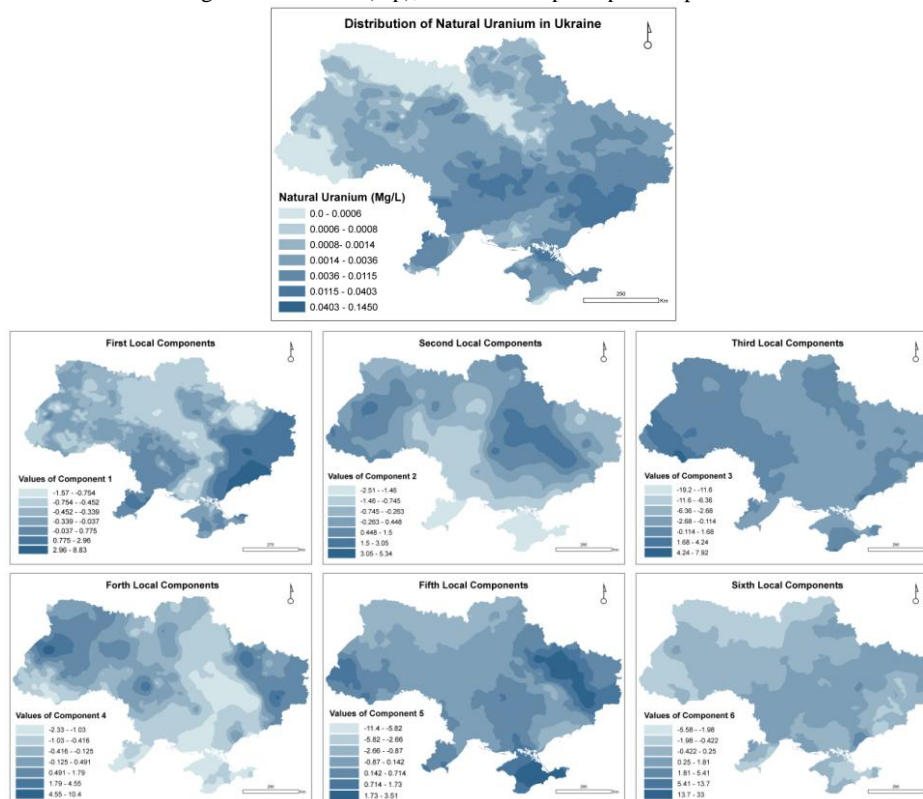
the first group of principle components are used to create the first multiple regression model. The model is further improved by incrementally adding elements from all other groups (Figure 4 and Figure 5).

Comparison of two multiple regression models based on the six principal components and the six individual environmental indicators shows that in general, both models indicate similar associations with the distribution of uranium in ground waters. However, the maximum coefficient of local

4   Conclusion

The paper demonstrates the use of a combination of different modelling techniques for better understanding of large-scale distributions of environmental indicators (such as uranium) by exploring spatial associations and patterns of associated natural variables and predictors. Spatial statistical modelling and local multiple regression can be successfully used for prediction of uranium concentration in surface waters

in different scenarios for changing climatic conditions in different parts of the territory of Ukraine. One intriguing association discovered from the modelling is the relationship between uranium and climatic variables. Concentration of

„Кіровгеологія". *Природні та антропогенні джерела формування радіоактивності природних вод України та радіаційний захист населення*, Киев, 2004.

Figure 5: Building consecutive multiple regression models using principal components.
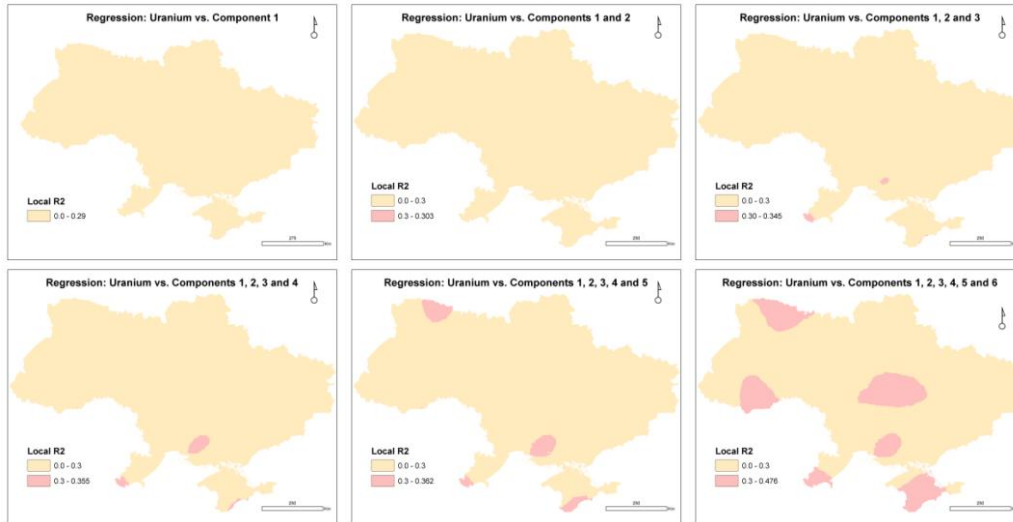
uranium has strong local correlation with precipitation, temperature, and humus. Precipitation and humus are the first two variables in the regression models for uranium. At the same time, precipitation has very high correlation with temperature (r=-0.825). Variation in temperature and precipitation due to the global climate change can alter their contribution in the uranium content. Those scenarios can be explored using the proposed spatial regression models.

The study confirms that the methods presented (global vs. local for regression and factor analysis) do not always provide overly objective ground for making conclusive inferences. Outcomes of different models sometimes do not support each other, e.g., some explanatory variables have low correlation with the dependent variable but at the same time have high percentage of explained variance in factor analysis. Global spatial regression modelling in a large-scale spatial analysis can be unsuitable for the local inference. The modelling results including their cartographic representations remain mainly descriptive and require interpretation by application experts.

Further research is envisioned in refining relationships between the environmental indicators and improving numerical forecasts by expanding the range of applied spatial statistical methods. The study is planned on exploring econometric models and spatial-clustering techniques to improve the robustness of the developed statistical model.

References

[1] Salih I., Pettersson H., Sivertun Å., Lund E. Spatial correlation between radon (222Rn) in groundwater and bedrock uranium (238U): GIS and geostatistical analyses. *Journal of Spatial Hydrology,* 2:2, 2002.

[2] Держкомприродресурсів України, Департамент геологічної служби, Казенне підприємство

[3] Макаренко ММ. Оцінка природних і техногенних факторів забруднень підземних і поверхневих вод природними радіонуклідами навколо уранових родовищ України. *Інформаційний бюлетень про стан геологічного середовища України*, Київ, 102-111, 2000.

[4] Getis A., Ord J.K. The Analysis of Spatial Association by Use of Distance Statistics, *Geographical Analysis*, 24:3, 1992.

[5] Anselin L. Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27(2), 93–115. 1995.

[6] Harman H.H. *Modern Factor Analysis*. 3rd ed., Chicago: University of Chicago Press; 1976.

[7] Fotheringham S.A., Brunsdon C., Charlton M. *Geographically Weighted Regression: the analysis of spatially varying relationships*. John Wiley & Sons, 2002.

[8] Tabachnick, B. G., Fidell, L. S. *Using Multivariate Statistics*. 6th ed., Needham Heights, MA: Allyn and Bacon, 2012.

[9] Jetz, W., Rahbek, C. and Lichstein, J.W. Local and global approaches to spatial data analysis in ecology. *Global Ecology & Biogeography*, 14: 97-98, 2005.

[10] Harris P., Brunsdon C., Charlton M.E. Geographically weighted principal components analysis. *International Journal of Geographical Information Science,* 25:1717-1736, 2011