Exploring Spatial Patterns of Uranium Distribution in Ukraine

Michael Govorov  
Vancouver Island University  
900 Fifth Street  
Nanaimo, Canada  
goverovm@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 IndexEl(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 multicollinearity of the data.

The classical global Pearson correlation coefficient \( r \) is used as a measure of global correlation. The coefficient is defined as

\[
 r = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{(n-1)S_u S_v}
\]  

(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 \( \Lambda_p \) is the matrix of factor loadings based on factor \( p \); \( \Omega_p \) is the diagonal matrix of \( \sigma_i \) that are the corresponding eigenvectors of \( R \), where \( R \) is the correlation matrix or eigenvectors of \( \Sigma \), where \( \Sigma = [\sigma_{ij}]_{n \times n} \) (3) is the covariance matrix of natural variables; and \( \Gamma \) 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 regression, GWR). 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

\[
 r_i(x_k, y_k) = \frac{\sum_{j=1}^{n} \sigma_j \cdot \left( u_j - \bar{u} \right) \cdot \left( v_j - \bar{v} \right)}{(n-1) \cdot S_u \cdot S_v} = \beta_i(x_k, y_k) \frac{S_u}{S_v} \tag{4}
\]

where \( r_i(x_k, y_k) \) is the correlation coefficient, \( (x_k, y_k) \) is the location of observation \( i \); \( u_j \) and \( v_j \) are the individual observations; \( \bar{u} \) and \( \bar{v} \) are the means of the two variables, \( \sigma_j \) 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; \( \beta_i(x_k, y_k) \) is the estimated parameter for the bivariate local-leastsquares 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].

Exploratory 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

\[
y_i(x_k, y_k) = \beta_{0i} + \beta_{1i}x_{1i} + \beta_{2i}x_{2i} + \ldots + \beta_{pi}x_{pi} + \varepsilon_i \tag{6}
\]

on observation \( i \), where \( y_i(x_k, y_k) \) is the dependent variable estimated in location \( i \); \( x_p \) are the exploratory variables; \( \beta_{pi}(x_k, y_k) \) are the local regression coefficients; \( p \) is the number of variables, and \( \varepsilon_i \) 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 centered 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 factors/regression models defined as

\[
y_i(x_k, y_k) = \beta_{0i} + \beta_{1i}y_{1i} + \ldots + \beta_{pi}y_{pi} + \varepsilon_i \tag{7}
\]

on observation \( i \), where \( y_{ki}(x_k, y_k) \) is the dependent variable estimated in location \( i \); \( F_p \) are the factors from principal components extraction; \( \beta_{pi}(x_k, y_k) \) are the local regression coefficients; \( p \) is the number of variables, and \( \varepsilon_i \) 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 Implementation

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), 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 (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.

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

### 3.4 Local Spatial Multiple Regression

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.

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

<table>
<thead>
<tr>
<th>Component</th>
<th>Natural Variables</th>
<th>Group</th>
<th>Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mineralization and hardness of water</td>
<td>Mineralogical</td>
<td>Hardness of water, mineralization of water</td>
</tr>
<tr>
<td>2</td>
<td>Metals dissolved in water</td>
<td>Geochemical</td>
<td>Cu, Fe, Cl, Zn</td>
</tr>
<tr>
<td>3</td>
<td>Climatic conditions of territory and formation of ground water</td>
<td>Climatic</td>
<td>Precipitation, temperature, humus</td>
</tr>
<tr>
<td>4</td>
<td>Organic compounds in water</td>
<td>Mineralogical</td>
<td>NO\textsubscript{3}, NH\textsubscript{4}, PO\textsubscript{4}</td>
</tr>
<tr>
<td>5</td>
<td>Geomorphological characteristics</td>
<td>Geological structure and geomorphology</td>
<td>Relief, isopach</td>
</tr>
<tr>
<td>6</td>
<td>Mineral compounds and satellite elements of uranium</td>
<td>Geochemical</td>
<td>Bicarbonate, fluoride, arsenic</td>
</tr>
</tbody>
</table>

#### Table 3: Multiple regression model summary.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>precipitation</td>
<td>0.51</td>
<td>0.26</td>
<td>0.26</td>
<td>1.1611</td>
</tr>
<tr>
<td>2</td>
<td>1 + humus</td>
<td>0.595</td>
<td>0.354</td>
<td>0.354</td>
<td>1.0849</td>
</tr>
<tr>
<td>3</td>
<td>2 + water hardness</td>
<td>0.623</td>
<td>0.388</td>
<td>0.387</td>
<td>1.0564</td>
</tr>
<tr>
<td>4</td>
<td>3 + F</td>
<td>0.629</td>
<td>0.396</td>
<td>0.395</td>
<td>1.0496</td>
</tr>
<tr>
<td>5</td>
<td>4 + Fe</td>
<td>0.634</td>
<td>0.402</td>
<td>0.402</td>
<td>1.0438</td>
</tr>
<tr>
<td>6</td>
<td>5 + As</td>
<td>0.637</td>
<td>0.405</td>
<td>0.405</td>
<td>1.0414</td>
</tr>
<tr>
<td>7</td>
<td>6 + SO\textsubscript{4}</td>
<td>0.638</td>
<td>0.407</td>
<td>0.407</td>
<td>1.0395</td>
</tr>
<tr>
<td>8</td>
<td>7 − water hardness</td>
<td>0.638</td>
<td>0.407</td>
<td>0.407</td>
<td>1.0396</td>
</tr>
<tr>
<td>9</td>
<td>8 + isopach</td>
<td>0.641</td>
<td>0.411</td>
<td>0.41</td>
<td>1.0368</td>
</tr>
<tr>
<td>10</td>
<td>9 + NH\textsubscript{4}</td>
<td>0.642</td>
<td>0.413</td>
<td>0.412</td>
<td>1.0351</td>
</tr>
<tr>
<td>11</td>
<td>10 + Cl</td>
<td>0.644</td>
<td>0.415</td>
<td>0.414</td>
<td>1.0332</td>
</tr>
<tr>
<td>12</td>
<td>11 + temperature</td>
<td>0.645</td>
<td>0.416</td>
<td>0.415</td>
<td>1.0325</td>
</tr>
<tr>
<td>13</td>
<td>12 + NO\textsubscript{3}</td>
<td>0.645</td>
<td>0.417</td>
<td>0.416</td>
<td>1.0317</td>
</tr>
<tr>
<td>14</td>
<td>13 + HCO\textsubscript{3}</td>
<td>0.646</td>
<td>0.418</td>
<td>0.416</td>
<td>1.0310</td>
</tr>
<tr>
<td>15</td>
<td>14 + Zn</td>
<td>0.647</td>
<td>0.418</td>
<td>0.417</td>
<td>1.0305</td>
</tr>
<tr>
<td>16</td>
<td>15 + Cu</td>
<td>0.647</td>
<td>0.419</td>
<td>0.418</td>
<td>1.0300</td>
</tr>
<tr>
<td>17</td>
<td>16 + PO\textsubscript{4}</td>
<td>0.648</td>
<td>0.419</td>
<td>0.418</td>
<td>1.0296</td>
</tr>
<tr>
<td>18</td>
<td>17 + mineralization</td>
<td>0.648</td>
<td>0.42</td>
<td>0.419</td>
<td>1.0288</td>
</tr>
</tbody>
</table>
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.

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 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 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 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.

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 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


