Improving R and ArcGIS integration

K. Krivoruchko[1*] and D. Pavlushko[1]

[1] *Environmental Systems Research Institute, 380 New York St, Redlands, CA, USA, 92373; kkrivoruchko@esri.com, dpavlushko@esri.com*

[*] *Corresponding author*

Abstract. We discuss a new approach for integrating R with ArcGIS. Later this year Esri plans to release an open source R package that provides a solution *inside the application process* for passing data between ArcGIS and R. Using this new methodology, the researchers can easily build geoprocessing tools that wrap R scripts. This new methodology will potentially support a community of people who develop and share R-based geoprocessing tools for ArcGIS.

Keywords. *R integration; ArcGIS; Geoprocessing; Bayesian statistics; Thyroid cancer.*

1 Integrating R with ArcGIS

There are several variants of R scripts usage in ArcGIS applications, see for example [1,2]. Typically, a Python script is used for data transfer between ArcGIS and R. Executing the R script and rendering the results is performed using the ArcPy and other Python modules. Finally, the ArcGIS script tool allows the user to select the required data with which to run the tool and view the results in ArcGIS.
In this paper, we discuss a new approach for integrating R with ArcGIS. Esri plans to release R package that allows data to be passed between ArcGIS and R inside the application process. Using this new methodology, researchers can effortlessly build geoprocessing tools which wrap R scripts. The new approach that wraps R scripts will be free and open source. It works efficiently by minimizing library reloading, utilizing in-memory data access, and eliminating intermediate scratch files.

1.1 How the new approach works

The R script below shows how existing R code can be integrated into a geoprocessing tool. At the beginning of the script, the user initializes the *arcgisbinding* library, which can read and convert (potentially any) GIS data into an R data frame and, optionally, into the spatial data object. Then actual R script is running and its output is added to the existing data frame or a new data frame is created. Finally, the result of calculations is saved to ArcGIS dataset for further use in the geoprocessing.

```
#### The required libraries ####
library(arcgisbinding)
#### Read GIS Data Features ####
inputFC <- "C:\\Demo\\Some_Polygons.shp"
info <- arc.open(inputFC)
#### Create Data.Frame ####
df <- arc.select(info, c("FID", "CancerCases", "Population1985"))
#### Spatial Data Object####
spObject <- arc.data2sp(df)
#### Plot Spatial Data ####
```

```
spplot(spObject)
#### begin some R script, which calculates a new variable ####
####
#### end of some R script  ####
#### Add New Column/Field to the Data.Frame ####
df$new_field <- new_variable
#### Export Spatial Data to Feature Class ####
outputFC <- "C:\\Demo\\NewResult.shp"
arc.write(outputFC, df)
```

The procedure is so simple that a new geoprocessing tool can be created in a few minutes, providing that the R script was carefully tested and it works properly. Note that if the R script uses one of the plotting commands, the R pop up window will be displayed even though R environment is not running.

1.2 How the new approach improves performance of R/ArcGIS projects

New R integration approach improves performance of R/ArcGIS projects in various ways. In the list below, we start with features, which are more important for ArcGIS users, then we explain why this integration can be useful for R users and developers, and, finally, we highlight features that can be important for both GIS and R users.

- It will expand accessibility of R in the GIS community.
- The ArcGIS user can use statistical models created in R environment without even knowing Python and R languages, providing that she trusts the R script owner.
- The R usage experience is similar to other ArcGIS analysis tools usage.
- The approach to authoring and publishing analytic web services is the same for tools written in Python and R.
- It honors settings of the geoprocessing analysis environment.
  - It provides support for reading/writing of all feature and table formats available in ArcGIS.
  - It does not require R script developers to know Python.
  - While R data packages can handle relatively large datasets, they do not provide support for the traditional database management tasks. The ArcGIS platform has native support for personal, workgroup and enterprise level geodatabases. The R statistician can leverage the data management capabilities of ArcGIS then use the ArcGIS/R bridge to seamlessly bring the data into R for in-depth statistical analysis.
  - Near real-time or streaming sensor data (stock markets, weather, geolocation) is a valuable sources of information for the R statistician. The ArcGIS platform enables real-time event-based data streams to be integrated as data sources.
  - It expands the number of R libraries users (typically, thousands of researchers are visiting webpages, which provide additional data analysis functionality for ArcGIS).
    - It honors selected features and table records during data analysis.
    - It handles the reprojection of data as needed.
    - It integrates naturally with ArcGIS Python scripting environment and ModelBuilder so that the R scripts can be used together with standard Python scripts and third-party Python libraries, such as NumPy (the package for scientific computing with Python), SciPy (a Python-based open-source software for mathematics, science, and engineering), and ProBT (extended Bayesian networks framework) adding great flexibility to solving complex GIS problems.
    - The ArcGIS platform provides a powerful and convenient mechanism for sharing analysis workflows and data. ArcGIS can package geoprocessing tools and the data used by the tools into a single compressed file (.gpk). All resources (models, scripts, data, layers, and files) needed to reexecute the tools are included in the package. This means consumers of the package can rerun the tools to produce the exact same results.

We expect that new methodology will help to build a community of people who develop and share R-based geoprocessing tools.

2 Example: Bayesian analysis of thyroid cancer in children in Belarus

We illustrate the R/ArcGIS integration with regression modeling of thyroid cancer in children using data collected in Belarussian districts several years after the Chernobyl accident, from 1986 to 1994. We want to investigate the relationships between cancer rates and some environmental factors. The main reason for thyroid cancer epidemic was irradiation by short-lived iodine radionuclides, but iodine measurements collected immediately after the Chernobyl accident are scarce to reconstruct its spatial distribution. Therefore, following many other researchers, we will use the following explanatory variables: average value of ^{137}Cs soil contamination in the administrative districts and the distance from the districts to the Chernobyl nuclear power plant.

Epidemiological data, the number of cancer cases in children in each Belorussian district and the number of children in 1986 (the population under risk), are described and provided in [1]. Sufficiently complete (about 15,000 samples) of ^{137}Cs values collected in Belarus are provided in [1] and we will estimate the average ^{137}Cs values in the districts using Geostatistical Analyst's conditional Gaussian simulation (see description of the model in [1], chapter 10). We will use the average ^{137}Cs values as the exposure variable.

There are many ways to calculate distances between the administrative districts and the Chernobyl NPP. In this exercise, we use the following algorithm:

- Estimate children population density using Geostatistical Analyst's Areal Interpolation [3].
- Sample about 1000 points from that density using Spatially Balanced Design GP Tool [4].
- Use median distance between the sampled points inside each district and the Chernobyl location.

Figure 1 shows how the algorithm above works in a ModelBuilder, an ArcGIS application used to create, edit, and manage geoprocessing models. The model uses a spatial join to add district names to the sample points; then calculates the distance from each sample points to Chernobyl; and, finally, it uses a Python script to calculate the median for all points from the same district. It produces the output table with distances for each district.

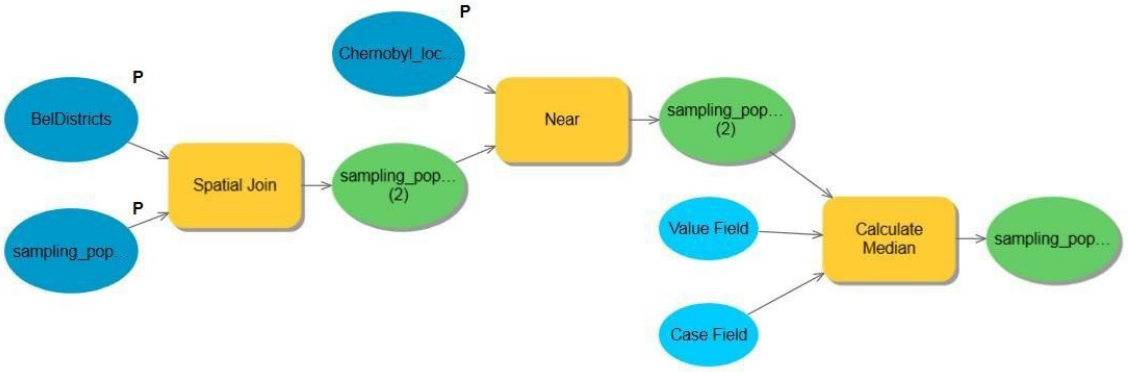

Figure 1. Distances between the administrative districts and the Chernobyl location are calculated in ArcGIS ModelBuilder.

For epidemiological regression, we use a conditionally specified prior spatial structure using a conditional autoregressive model. The model requires specification of the neighbors of each polygon and their weights. We create the neighbors list using the R *spdep* package [5].

Our model allows the regression coefficients to be spatially correlated and change locally. The R and WinBUGS [6] scripts are discussed in details in [1] (in appendix 3). A WinBUGS model is called and the inferences are summarized using R2WinBUGS R package [7]. We will provide the updated scripts and instructions for their usage in ArcGIS geoprocessing in the full paper.

Using this model, we can map the regression coefficients to see how much each covariate influences the value of the thyroid cancer risk locally. Figure 2 shows maps of the regression coefficients for ^{137}Cs soil contamination (left) and its standard error (right). We see that influence of the ^{137}Cs soil contamination covariate is gradually decreasing towards the northwest. However, the prediction standard error is very large, and we should be careful in relating the ^{137}Cs soil contamination to thyroid cancer risk.

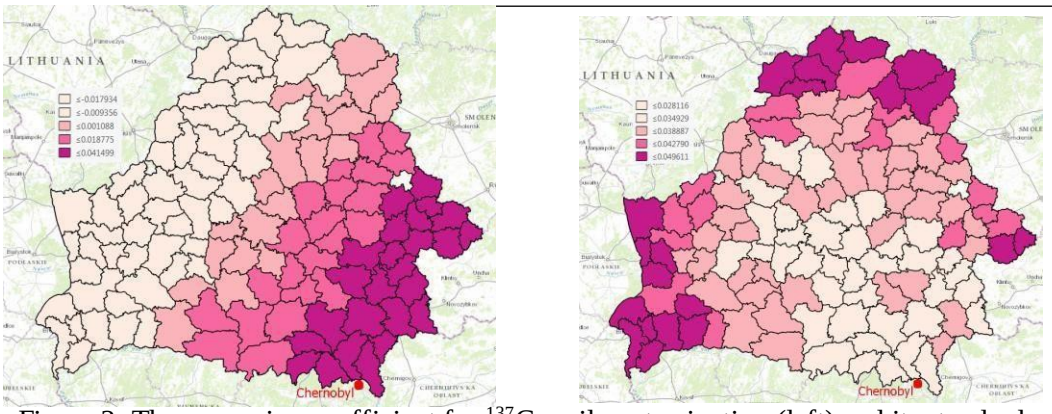

Figure 2: The regression coefficient for ^{137}Cs soil contamination (left) and its standard error (right).

Figure 3 shows the proportion of the environmental and spatially correlated components of the thyroid cancer risk for each administrative district. The geoprocessing tool, which runs the statistical model, is shown at right (note that it took 2 minutes and 17 seconds to add Bayesian statistical output to the shape file with the epidemiological data). We see that the average value of ^{137}Cs soil contamination and distance to Chernobyl do not play a significant role in the southern part of Belarus close to Chernobyl. This is because iodine ^{131}I deposition was very different from ^{137}Cs deposition since the latter radionuclide is much heavier and because the half-life of the former is very short.

It should be noted, however, that less sophisticated models found significant relationship between thyroid cancer in children and ^{137}Cs data, see [1].

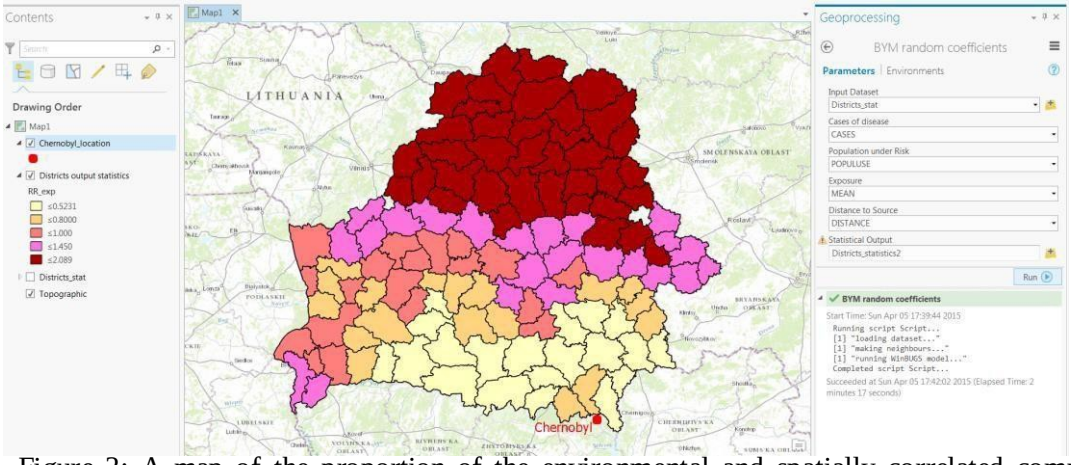

Figure 3: A map of the proportion of the environmental and spatially correlated components of the thyroid cancer risk and the geoprocessing tool, which runs the statistical model (at right.)

Note that the covariates used in this exercise are not precise and the analysis can be made more realistic by taking into account the covariates uncertainties.

References

[1] Krivoruchko K. (2011) *Spatial Statistical Data Analysis for GIS Users*. ESRI Press, 928 pp.
[2] Introduction to R scripting with ArcGIS, http://esri.ca/en/content/introduction-r-scripting-arcgis.
[3] Krivoruchko K., Gribov A. and Krause E. (2011) Multivariate Areal Interpolation for Continuous
[4] Krivoruchko K. and Butler K. (2013) Unequal Probability-Based Spatial Sampling. ArcUser Spring 2013, pp.10-17. Also available online at http://www.esri.com/esri-news/arcuser/spring-2013/unequal-probability-based-spatial-sampling
[5] Bivand R. and Piras G. (2015) Comparing Implementations of Estimation Methods for Spatial Econometrics. Journal of Statistical Software, 63(18), 1-36. URL http://www.jstatsoft.org/v63/i18/.
[6] The BUGS Project. http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs.
and Count Data. Procedia Environmental Sciences. Volume 3, 2011, Pages 14-19.
[7] Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R. Journal of Statistical Software, 12(3), 1-16.