Exploring spatial scale by interactive map for geographically weighted correlation

Joseph PERCIVAL, Narumasa TSUTSUMIDA, Daisuke MURAKAMI, Takahiro YOSHIDA, and Tomoki NAKAYA

Abstract: We introduce *gwpcorMapper*: an interactive web-based mapping application software for visualizing and analyzing geographically weighted correlation and partial correlation statistics. The significance of *gwpcorMapper* is that it allows geospatial analysts to easily explore the influence of spatial scale (localness) on correlation analyses. We demonstrate the application with over 200 variables using the small-area national census statistics of the year 2005 in Tokyo.

Keywords: Geographically weighted statistics, census data, correlation

1. Introduction

With the increase of data availability, the need to be able to work with large datasets is becoming ubiquitous. In the field of Geospatial Science, researchers often have to deal with very large data sets, where it is not uncommon to deal with upwards of hundreds of variables. Specialized statistical tools are required to inspect such large and complex data.

Geographically weighted (GW) models are a set of statistical tools that quantify relationships between data in terms of both spatial scale and the geographic proximity between objects. The GW approach uses a moving-window kernel across geographical space and calculates a statistical model or a summary statistic with distance-decayed spatial weights. This approach is useful for examining the underlying spatial structure of data, particularly for investigating spatial non-stationarity. If GW outputs vary across space, it can be said that spatial non-stationarity exists in the data. Popular methods in the GW toolkit include GW regression (Brundson et al., 1996), GW correlation (Brundson et al., 2002) and GW partial correlation methods (Percival and Tsutsumida, 2017).

An essential aspect of the GW approach is selecting an appropriate bandwidth and type of the moving window kernel. The bandwidth determines the effective spatial scale under which the statistical model is employed, while the kernel type defines the weighting scheme of the data.

Spatial scale and bandwidth size are often of highest concern. While a small bandwidth may lead to erratic results due to an insufficient number of data points within a kernel to perform local statistical analyses, a large bandwidth may not capture any meaningful geographical variations in the underlying spatial process. Bandwidth sizes can be defined by either fixed (e.g. 1 km within a data object) or adaptive (e.g. the nearest 100 objects). Multiple bandwidth optimization methods have been proposed and implemented in popular GW toolsets, including selection by cross-validation (CV) (Fotheringham et al., 1998) or Akaike's information criterion (AIC) (Fotheringham et al., 2003). However, it has been observed that such selection methods may not

Joseph Percival
Georepublic
ipercival@gmail.com

always find the most sensible bandwidth in terms of interpretation (Tsutsumida et al., 2019). In this respect, bandwidth selection should be made only after inspecting results of the full bandwidth function (Gollini et al., 2015). While tools exist for determining the bandwidth size statistically, a visualization tool of resulting maps under varying bandwidth sizes is lacking.

This research introduces the development of a new software named *gwpcorMapper*, an R-based web application for visualizing maps of GW correlation and partial correlation statistics. This application allows exploratory correlation analysis by interactively changing bandwidth sizes and kernel types under different statistics. *gwpcorMapper* compliments existing toolsets by making exploratory GW correlation analyses simpler and less time consuming compared to existing approaches that require command line operations to analyze data and create individual maps. In the case of analyzing geospatial datasets with hundreds of variables across the area of interest, the conventional tools produce a very large number of maps to be printed. Our *gwpcorMapper* simplifies this process by enabling geospatial analysts to quickly map results of GW correlation with user-selected variables in a web browser and easily allows the inspection of spatial non-stationarity at various spatial scales. We demonstrate the application of *gwpcorMapper* by using it to map geographically weighted correlation and partial correlation coefficients between 204 variables using the small-area national census statistics of the year 2005 in Tokyo.

2. Software Description

gwpcorMapper is built in R and runs in a web browser powered by R Shiny: an R package for building interactive web applications using the R programming language (Chang et al., 2017). R is one of the most popular and active software environments amongst geospatial analysts and features a wide range of user-driven geospatial statistical packages (e.g. Gollini et al., 2015). *gwpcorMapper* extends and includes many of these tools, notably the *GWpcor* package (Percival and Tsutsumida, 2017).

The *gwpcorMapper* interface incorporates a minimalistic and modern design, with two basic user interface components: (i) a sidebar for variable selection and parameter tuning of GW correlation and GW partial correlation analyses, and (ii) a graphics panel consisting of a map and scatter plots of results (Figure 1). The map panel makes use of Mapbox base maps and the scatter plots are created using the Plotly graphics library (Sievert 2018). The application features callback functions between the control panel, map, and scatter plots that allow dynamic interaction for users. Users can click on map features to highlight local results.

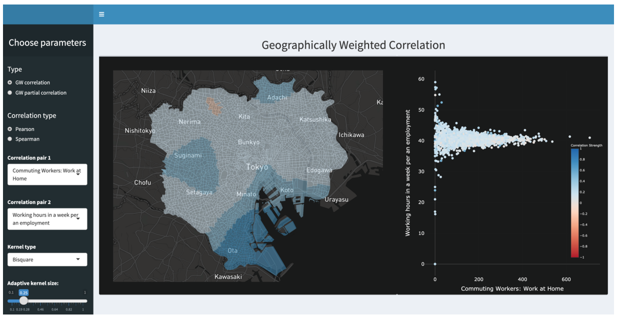

Figure 1. The user interface of *gwpcorMapper*.

2.1 Parameter selection

The sidebar component contains all user controls for analyzing and mapping GW correlation coefficients. The current version includes the following features: (i) GW tool selection (either GW correlation or GW partial correlation), (ii) correlation type (either parametric, Pearson's r or non-parametric, Spearman's ρ), (iii) drop-down panels for variable selection, (iv) kernel type (one of either gaussian, exponential, bi-square, tri-cube, or box-car kernel functions), and a slider for selective the adaptive kernel bandwidth size. An additional function to import data formatted by comma separated text files

(.csv) or geopackage (.gpkg) will be also featured. Users can select and upload data and choose between GW correlations and associated parameters and have the output maps be generated on the fly.

2.2 Results visualization

Selected parameters are used to produce an interactive spatial surface of the local correlation coefficients over a Mapbox base map in the graphics panel. A scatter plot showing the global relationship between the selected two variables is also plotted as a reference. These two figures interact dynamically both with each other and the user. For example, when a user of the software clicks on an area in the map, the corresponding point in the scatter plot is highlighted and its statistical attributes are shown both on the clicked feature of the map and the highlighted point in the plot.

3. Application Example

To demostrate *gwpcorMapper*, we use the small-area national census statistics of the year 2005 in Tokyo, Japan. This dataset has 204 variables that describe the socioeconomic and demographic structure of 3,134 administrative units called *chocho-aza* in the 23 special wards of Tokyo. This data is loaded into the software as a geopackage.

This dataset will produce a total of 41,412 bivariate variable pairs for correlation. The ease of being able to map local correlation coefficients of such pairs in this application is evident as users can simply search and select variables to plot from a drop-down menu. As an example of exploratory GW correlation analysis, we select two variables (the *total number of commuting workers* and the *number of working hours in a week per employment*) to see how the output can interact with the bandwidth size of the moving window kernel. To emconpass the importance of the explanatory process, we compare with the result of bandwidth selection via a CV selection method (Gollini et al., 2015), which resulted in 3,132 as the optimal adaptive bandwidth; viz. that 3,132 out of 3,134 administrative units should be contained in the moving window kernel. This represents 99% of the data and is very near the global model. Plotting this on a map reveals little spatial heterogeinty in local correlation coeffients between the two variables (Figure 3d).

However, in examining the CV scores across the full bandwidth spectrum, we find that the CV approach does not perform well as its scores do not indicate a clear global minimum (Figure 2). As expected in a previous study (Tsutsumida et al. 2019), there are cases that CV scores appear to decay exponentially along increasing adaptive bandwidth sizes. In this situation, automated bandwidth selection by CV cannot show any possible geographic variations in the local statistics under investigation. Instead, it may be more meaningful to visually inspect local correlation coefficients across varying bandwidth sizes using *gwpcorMapper* to subjectively choose a bandwidth size based on intuitive interpretation (Figure 3). By doing so we can see how the degree of scale (bandwidth size) can influence local statistics, and in the case of our two selected variables, we can identify strong localized correlative relationships in the areas surrounding Ota, Suginami, Itabashi, Adachi, and Chuo wards.

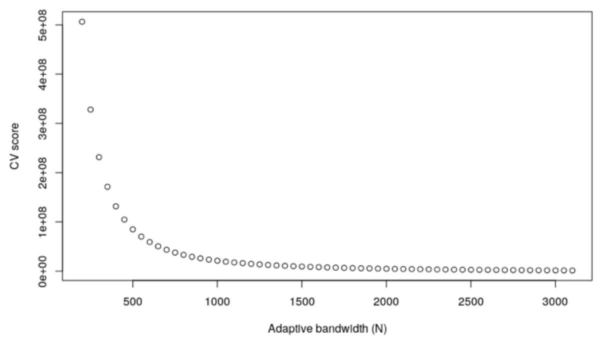

Figure 2 CV scores across various adaptive bandwidth sizes.

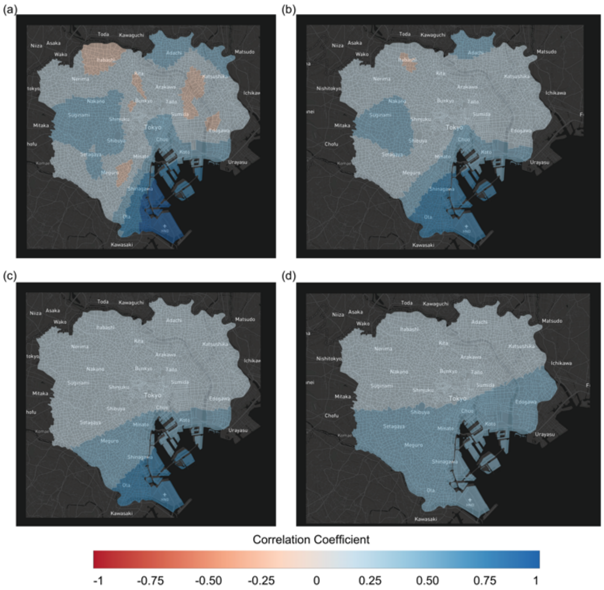

Figure 3. Localized correlation coeffients between the *total number of commuting workers* and the *number of working hours in a week per employment* with adaptive bandwidth sizes of (a) 10%, (b) 25%, (c) 50%, and (d) 99% of total.

4. Conclusion

Beyond providing the visual tools to subjectively select parameters, *gwpcorMapper* offers geospatial analysts and researchers a simple interactive tool to explore GW correlations on a map without typing R commands into a console. Its interactive features enable users to achieve quick visualization of geographically localized correlative relationships amongst multivariate data which can be selected and changed easily from searchable drop-down lists.

Finally, it should be noted that *gwpcorMapper* is open source, and its source code, along with a full feature list and manual are located in a GitHub repository at https://github.com/naru-T/gwpcormapper. Currently, *gwpcorMapper* can be downloaded and run locally in a docker container.

Acknowledgement

This work is supported by ROIS-DS-JOINT (006RP2018 & 004RP2019).

References

Brunsdon, C., Fotheringham, A.S. and Charlton, M.E., 1996. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**, 281-298.

Brunsdon, C., Fotheringham, A.S. and Charlton, M., 2002. Geographically weighted summary statistics—a framework for localised exploratory data analysis. *Computers, Environment and Urban Systems*, **26**, 501-524.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. 2017. shiny: Web Application Framework for R. *R package version 1.0.5.* https://CRAN.R-project.org/package=shiny

Fotheringham, A.S., Charlton, M.E. and Brunsdon, C., 1998. Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, **30**, 1905-1927.

Fotheringham, A.S., Brunsdon, C. and Charlton, M., 2003. *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., and Harris, P. 2015. GWmodel: An R Package for Exploring Spatial Heterogeneity Using Geographically Weighted Models. *Journal of Statistical Software*, **63**, 1-50.

Sievert, C. 2018. *plotly for R*. https://plotly-r.com

Percival, J. and Tsutsumida, N., 2017. Geographically weighted partial correlation for spatial analysis, *GI_Forum*, **5**, 36-43.

Tsutsumida, N., Rodríguez-Veiga, P., Harris, P., Balzter, H. and Comber, A., 2019. Investigating spatial error structures in continuous raster data. *International Journal of Applied Earth Observation and Geoinformation*, **74**, 259-268.