HHS Public Access

Author manuscript

Trans GIS. Author manuscript; available in PMC 2019 June 01.

Published in final edited form as:

Trans GIS. 2018 June ; 22(3): 721–736. doi:10.1111/tgis.12452.

Integrating spatial data analysis functionalities in a GIS environment:

Spatial Analysis using ArcGIS Engine and R (SAAR)

Hyeongmo Koo[1], **Yongwan Chun**[1], and **Daniel Griffith**[1]

[1]The University of Texas at Dallas, School of Economic, Political and Policy Sciences, 800 West Campbell Road, Richardson, United States.

Abstract

Spatial data analysis (SDA) tools to efficiently handle and explore spatial data have become readily available. Although these SDA tools have their own strengths and purposes, they suffer from limited support in terms of a development environment offering easy customization and high extensibility, a strength of open source software. This paper presents a stand-alone software package for SDA in a geographic information systems (GIS) environment, called Spatial Analysis using ArcGIS Engine and R (SAAR), which provides an integrated GIS and SDA environment. A set of SDA tools in SAAR utilize functions in R using R.NET, while other tools were developed in .NET independent of R. SAAR provides an efficient working environment for both general and advanced GIS users. For general GIS users with limited programming skills, SAAR furnishes advanced SDA tools in a popular ArcGIS environment with graphical user interfaces. For advanced GIS users, SAAR offers an extensible GIS platform to help them customize and implement SDA functions with relatively little development effort. This paper demonstrates some functionalities of SAAR using census data for Texas counties.

Keywords

spatial data analysis; GIS; R-GIS integration; .NET

1. Introduction

A synergy arising from integrating spatial data analysis (SDA) techniques with Geographic Information Systems (GIS) has been discussed in the literature (e.g., Goodchild et al. 1992), revealing that an integration of these two components provides capabilities to efficiently handle spatial data for GIS, and to effectively visualize and explore data for SDA. Some well-known outcomes within this context include GeoDa and OpenGeoDa (Anselin et al. 2006; Anselin and McCann 2009), CommonGIS (Andrienko et al. 2002, 2003), STARS (Rey and Janikas 2006), and GeoSurveillance (Yamada et al. 2009). These packages offer a wide array of SDA tools and GIS functions, each offering slightly different strengths according to its designed purpose. Regardless, these packages also have potential

Correspondence: **Mr.Hyeongmo Koo**, The University of Texas at Dallas, School of Economic, Political and Policy Sciences, 800 West Campbell Road, Richardson, Richardson, United States. Hyeongmo.Koo@utdallas.edu.

weaknesses. First, they do not support a seamless working environment between SDA and GIS. Analysts may need to move back and forth between an SDA tool and GIS software, for example, to create a high quality final map with an SDA result. This undesirable feature can be a hurdle for GIS users who are not familiar with a particular SDA tool. Second, many of the packages either do not support, or are limited in their support of, a development environment for customization and/or extensibility. For example, spatial regression capabilities, such as those based upon the simultaneous autoregressive or conditional autoregressive model, are lacking in these packages. Although some packages support an open source environment, extensive programming is necessary to implement SDA functions. In contrast, one strength of R, an open source software environment for statistical computing and graphics, is its high extensibility and easy customization; that is, users freely can customize and extend its functions. However, R requires a very steep learning curve due to its command-line interface, and still is limited in its abilities and user-friendliness for visualizing and exploring spatial data.

This paper presents an SDA software package in a popular ArcGIS environment, called Spatial Analysis using ArcGIS Engine and R (SAAR). This application provides an environment integrating GIS and SDA using ArcGIS Engine and R. Specifically, using ArcGIS Engine components, SAAR can efficiently and effectively visualize and explore spatial data with a familiar and interactive graphical user interface (GUI) in a popular GIS environment, and also can support a lightweight and specialized stand-alone application environment rather than a complete and general purpose GIS application. SDA tools in SAAR are implemented programmatically with R functions based on an interoperability bridge program R.NET (R.NET 2015). This seamless interoperation between ArcGIS Engine and R is hidden from end-users, and does not require a user to open and install R and other additional components. In this development environment, SAAR furnishes various advanced SDA tools with relatively little development effort. In addition, it can provide an efficient development framework with high extensibility and easy customization because it involves an open-source environment.

This paper provides a detail introduction to SAAR. The following section describes the motivation for its development. The next section presents the main functionality of SAAR with its distinctive development configuration. Then detail capabilities of the functions are illustrated with median house values in Texas counties. Finally, this paper concludes with a discussion about future directions for the continued development of SAAR.

2. Motivation

SDA and GIS have been incorporated in two general approaches (Goodchild et al. 1992). One modifies or extends statistical software packages for spatial data handling (e.g., LeSage 1999; Bivand 2000; Bivand and Gebhardt 2000). Although this approach provides an easy implementation of SDA techniques for GIS, it suffers from two major drawbacks. First, because of its aspatial statistical software environment, this approach has limited visualizing methods for spatial data. Hence, this incorporation is limited in exploratory SDA (ESDA) and model validation checking, which is inconsistent with the trend of modern software development (Goodchild et al. 1992). Second, general statistical software packages (e.g., R

and SAS) often require a steep learning curve (Delmelle et al. 2011), because of their command-line interfaces and programming skill requirements.

Another approach is integrating SDA tools into a GIS application. Broadly speaking, this approach can be further classified into either loose or tight coupling of an existing GIS application with a statistical application, or a stand-alone application (Goodchild et al. 1992). In loose coupling, GIS and statistical applications operate independently and transfer input and output files through common data formats: e.g., CrimeStat (Levine 2010). A manual communication between input and output files in loose coupling causes inefficiencies in workflows. In contrast, tight coupling modifies a GIS application to operate statistical software packages within the GIS environment (e.g., Symanzik et al. 2000; Rura and Griffith 2010; Delmelle et al. 2011). Thus, tight coupling can allow communications between GIS and statistical applications through a GIS application interface, and hide a process for data exchange from end-users (Delmelle et al. 2011). Also, tight coupling generally supports an interactive exploration of spatial data through linked windows and brushing techniques in GIS applications (e.g., Unwin and Hofmann 1997; Brunsdon et al. 1998; Dykes 1998; Haining et al. 1998). However, tight coupling potentially can suffer from application version upgrades because it primarily relies on main GIS software. That is, a tight coupling application requires installation of particular GIS software, and needs constant maintenance to match its version to that of the employed GIS software, specifically when new version releases of the main GIS software occur.

For integrating SDA with GIS, stand-alone software packages have been developed to provide a more efficient installation and a more stable maintenance environment than coupling implementations (e.g., Andrienko et al. 2002, 2003; Anselin et al. 2006; Rey and Janikas 2006; Anselin and McCann 2009; Yamada et al. 2009, Jacquez et al. 2014). However, these stand-alone software packages often require extensive efforts for development because SDA techniques may need to be developed from scratch. Furthermore, basic GIS functions are not readily available and need to be programmatically implemented (Vandergast et al. 2011). Even if extensive effort is expended to develop a stand-alone software package, an update of the software package (e.g., an implementation of new SDA techniques) still requires huge development efforts due to a limitation of the software package for customization and/or extensibility. For example, Open GeoDa (Anselin and McCann 2009) and spatial statistics tools in ArcGIS Desktop (ESRI 2016), which are popular SDA applications, still have limited spatial regression capabilities. In addition, although SpaceStat 4.0 (Jacquez et al. 2014) furnishes various aspatial regression procedures (e.g., Poisson and logistic models) and mixed model regression, it also contains only spatial lag and error models.

SAAR was developed to overcome these limitations by integrating ArcGIS Engine and R. Specifically, ArcGIS Engine, which supports building a customized stand-alone GIS application without requiring the complete ArcGIS Desktop environment, helps SAAR efficiently visualize and explore spatial data with familiar GUIs in a popular GIS environment. In addition, it provides stable maintenance in a stand-alone application environment, with a smaller amount of required disk space than that of a complete GIS application. SAAR implements powerful SDA functions relatively easily because it utilizes

various SDA functions in R, benefitting from R's open source environment (Bivand and Gebhardt 2000; Bivand et al. 2013b). Furthermore, SAAR can achieve high extensibility and easy customization because of a relatively simple development process for integration between ArcGIS Engine and R.

This preceding integration is consistent with efforts to integrate R functionalities in popular GIS applications: R and ArcGIS. However, because the R processing configuration in QGIS still has a command-line based interface to execute R functions from a GIS environment (i.e., not a GUI), GIS end-users still find these modules difficult to use without acquiring R programming skills. The R-ArcGIS Bridge can implement R functions in a geoprocessing tool in an ArcGIS environment, and help to visualize geoprocessing results with a map. However, because the current R-ArcGIS Bridge supports limited output types in an ArcGIS dataset (e.g., a shapefile and geodatabase), it cannot produce various types of SDA results (e.g., graphs and summary tables).

3. The development architecture

SAAR is a stand-alone application including two main groups of functionalities: GIS and SDA (Figure 1). Each group consists of two components: a visualization, and an analytic engine. The visualization engines for both groups use .NET Framework for displaying general graphics windows and controls. ArcGIS Engine[1] is used for map display and navigation for GIS functionality. Note that a valid ESRI license (i.e., an ArcGIS Engine Runtime or an ArcGIS Desktop license) is required to run SAAR. Meanwhile, SAAR has separate analytic engines for GIS and SDA functionalities. Specifically, the GIS analytic engine utilizes ArcObjects to manipulate and retrieve spatial data, and the analytic engine for SDA techniques utilizes R functions.

The distinctive feature in the development environment of SAAR is the integration of these two different GIS and SDA analytic engines using an interoperability bridge program R.NET (R.NET 2015). Because the GIS functionality of SAAR is developed in .NET Framework, operating R functions within .NET Framework through R.NET is the first integration step. Specifically, R.NET enables R functions to operate in the native R shared libraries within .NET Framework. In SAAR, all necessary R libraries and developed internal SDA functions are included in the installation package of SAAR so that end-users do not need to install any additional programs or libraries.

The interoperation between .NET Framework and R is developed with a relatively simple process through R.NET (Figure 2). First, the instance of an R.NET object needs to be retrieved and initialized in order to use R in .NET Framework. An initialized R.NET object can create various types of bridge classes that can transfer input and output between R and .NET Framework. For example, a double array class in .NET is compatible with the bridge class of a numeric vector in R.NET, and the numeric vector in R.NET interacts as a real vector in R using the method *SetSymbol* of R.NET. After transferring an input from .NET Framework to R, an R function can be executed from .NET Framework using the

[1]GIS software developed with ArcGIS Engine is allowed to be legally distributed.

Evaluate method of R.NET. The outcome of the R function also can be retrieved through a bridge class, or directly with a .NET class (i.e., arrays). Following these relatively simple steps, this seamless interoperation between R and .NET Framework is achieved in SAAR, and SDA functions in R are fully integrated into a GIS environment without the need to develop programming code from scratch for functions. Furthermore, advanced users should be able to easily customize and extend SAAR by bringing more functionalities from R while having relatively little experience working with .NET programming.

4. Functions in SAAR

This section describes a list of the currently implemented functions in SAAR, and their detail specifications. Figure 3 presents functions available in SAAR, which are classified into the following three broad categories: basic GIS and geovisualization, ESDA, and confirmatory spatial data analysis (CSDA).

First, SAAR provides basic GIS and geovisualization functions. Because SAAR is specialized for advanced SDA techniques, it furnishes fundamental GIS functionalities for spatial data exploration and tabular manipulation. The fundamental functions for spatial data display and navigation primarily are accomplished using controls in ArcGIS Engine. For tabular data manipulation, SAAR offers functions for adding and deleting fields, as well as a field calculator. In addition, SAAR supports various geovisualization functions, ranging from general thematic mapping techniques to advanced bivariate mapping for uncertainty visualization. General thematic mapping techniques, which include choropleth and proportional symbols, are implemented mainly with ArcObjects. They support an interactive GUI to provide a flexible manipulation of maps for various color schemes and map classification methods. The available map classification methods include not only common map classification methods such as Jenks's natural breaks (Jenks 1977), quantile, and equal interval method, but also recent map classification methods incorporating uncertainty information (Sun et al. 2014; Koo et al. 2017). In addition, SAAR furnishes uncertainty visualization tools as both static and dynamic methods. Static methods include bivariate mapping techniques for uncertainty visualization are implemented, which support a simultaneous display of attributes with their corresponding uncertainties (Koo et al. 2018). Dynamic methods include the form of animation, in which a longer duration represents a smaller uncertainty (e.g., Fisher 1993), and interactive control, which changes the appearance of a thematic map based on a level of uncertainty (e.g., Rheingans 1992).

Second, SARR furnishes both exploratory data analysis (EDA) and ESDA tools. A number of general statistical graphs (e.g., histogram and scatter plot) are supported, and dynamic linking and brushing, which are central techniques in EDA and ESDA (Cleveland and McGill 1988, Symanzik et al. 2000), are implemented. The usefulness of dynamic linking and brushing in ESDA has been demonstrated within various ESDA applications (e.g., Brunsdon et al. 1998; Dykes 1998; Haining et al. 1998). SARR also supports advanced dynamic linking and brushing to explore associations for spatial neighbors in various contexts, such as one point in a data space corresponding to pairwise locations (e.g., a variogram-cloud), one location with its neighbors on a map (e.g., spatial correlogram), and a general one-to-one matching between a location in a data space and a map. Technically, in

SARR, dynamic linking and brushing are achieved with the unique identification numbers of both an input and outputs. Thus, one graphics window or a single layer in map view serves as an input to define a subset of values or locations among observations, but the number of linked maps and graphs (i.e., outputs) is unlimited. Also, different color symbols are used for an input (in cyan) and target sources (in red) so that users can distinguish them easily from each other (see Figure 5).

Third, various CSDA tools are implemented in SAAR. Specifically, SAAR mainly uses functions in the *spdep* package (Bivand 2002) to construct a spatial weights matrix, test for spatial autocorrelation, and estimate spatial regression model parameters. SAAR supports construction of a spatial weights matrix based on polygon contiguities (i.e., rook and queen definitions), point pattern by distance (i.e., a fixed distance and *k*-nearest neighbors), and Delaunay triangulations. In addition, SAAR furnishes graphical tools to examine the spatial configuration affiliated with a spatial weights matrix, including a connectivity histogram and a map with dynamic linking and brushing. SAAR contains functions to measure spatial autocorrelation for both global contexts—Moran Coefficient (MC) and Geary Ratio (GR)—and local contexts—local MC and G_i^* (Getis and Ord 1992). The significance tests for these measures can be conducted under normality and/or randomization assumptions, with a multiple-testing adjustment option also being available. Furthermore, bivariate spatial autocorrelation measures for both global and local tests are implemented, which simultaneously consider the correlation between two variables as well as the spatial autocorrelation of these variables (Lee 2001, 2004, 2009). SAAR supports various types of spatial regression models, including the simultaneous autoregressive (SAR) (i.e., spatial error and spatial lag), conditional autoregressive (CAR), spatial moving average (SMA), and spatial Durbin specifications (Anselin 1988). In addition, the various Jacobian computation methods in the *spdep* package [e.g., Cholesky, Chebyshev, and Monte Carlo approximate log-determinant methods as well as the exact eigenvalue approach (Bivand et al. 2013a)] are available for different types of spatial weights matrices (i.e., sparse and dense), which makes the estimation of a spatial regression model possible for large size datasets. Furthermore, SAAR furnishes Moran eigenvector spatial filtering (MESF, Griffith 2003) tool for both linear and generalized linear models (GLM) (e.g., Poisson and binomial). Briefly, MESF extracts eigenvectors from a spatial weights matrix, and uses them as independent variables in a stepwise regression specification (Griffith 2003); more computationally efficient calculations for this technique are available (e.g., Chun et al. 2016), and also are implemented in SAAR.

5. An application

This section demonstrates the functionalities of SAAR using an empirical dataset. It follows a general SDA process, ranging from simple mapping, ESDA, and CSDA, to an uncertainty exploration in the analyses and data. The empirical data are obtained from the 2010–2014 five-year American Community Survey (ACS) data for Texas counties. Specifically, this section mainly presents a SDA using the estimates of median house value (*hvalue*) for the 254 counties in Texas. Two additional variables, the unemployment rate (*unemploy*) and median year in which a structure was built (*year*), are used as covariates in the context of regression. ACS reports margins of error (MOEs) with estimates. The MOEs are used for

uncertainty visualization and classification as a form of the coefficient of variation (CV) and/or standard error. Table 1 summarizes descriptive statistics for these variables.

5.1 A simple mapping and geovisualization

Figure 4 presents the simple choropleth map of *hvalue* in the main GUI of SAAR. The GUIs of SAAR are user-friendly and compatible with ones in the contemporary computing environment (Figure 4a). Basically, all functions are arranged as menu items and toolbar buttons in the GUIs, and corresponding context menus are structured based upon user interaction. In Figure 4, the choropleth maps are prepared with different map classification methods and color schemes. Specifically, the choropleth maps of *hvalue* and *unemploy* are constructed based on a quantile method to properly reflect the correlation between two variables (Slocum et al. 2009), and the map of *year* is drawn with an equal interval method. With the three choropleth maps, positive correlation is visually observed between *hvalue* and *year*, and negative correlation is observed between *hvalue* and *unemploy*, especially in the three major Metropolitan Statistical Areas (MSA): the Dallas-Fort Worth-Arlington, the Houston-The Woodlands-Sugar Land, and the San Antonio-New Braunfels MSAs. Furthermore, SAAR furnishes tools to construct a page layout for map printing, where map elements (e.g., legend, north arrow, and scale bar) can be arranged. The map layout including map elements (i.e., Figure 4b and 4c) also can be printed or exported as an image file employing several image file formats, including JPEG, Tagged Image File Format (TIFF), and bitmap image (BMP).

5.2 Exploratory data analysis

SAAR contains EDA tools with the support of dynamic linking and brushing techniques among statistical graphs and a layer map in a map view. EDA tools support not only generic graphics such as histogram, boxplot, and scatter plot, but also violin (Hintze and Nelson 1998) and quantile-comparison plots. Figure 5 demonstrates these EDA functions in SAAR, with the distribution of *hvalue*, and a relationship between *hvalue* and the selected covariates; *unemploy* and *year*. The histogram in Figure 5c shows the numerical distribution of *hvalue*, with some extremely high values highlighted. It indicates that the distribution of *hvalue* is positively skewed. This skewed distribution implies the need for an appropriate data transformation to make the variable comply with the normality assumption of many statistical techniques (e.g., linear regression). As observed by visual inspection of the choropleth maps in the previous section, the scatter plot in Figure 5a depicts a positive relationship between *hvalue* and *year*, whereas the scatter plot in Figure 5b portrays a negative relationship between *hvalue* and *unemploy*.

The relationship among the three variables can be explored further with dynamic linking and brushing. In Figure 5a, the observations with 1995 or later for *year* are selected and brushed as an input, and are highlighted in cyan. Dynamic linking and brushing highlight the same counties in the histogram with the cross-hatched red symbol (Figure 5c), in the scatter plot in red (Figure 5b), and in the map in cyan (Figure 5d). Specifically, the histogram in Figure 5c shows that the selected counties with 1995 or later for *year* comprise a large proportion of counties with high median house values. In addition, the highlighted observations in the scatter plot (Figure 5b) roughly show there is no significant correlation between *year* and

unemploy, which presumes that inclusion of both variables as covariates in a regression model may not create a multicollinearity problem. Finally, the map shows that the newly built houses are mainly located in suburban areas of the three major MSAs in Texas.

5.3 Exploratory spatial data analysis and spatial autocorrelation

SAAR provides a set of global and local spatial autocorrelation measures. Figure 6 illustrates the test results for both global and local spatial autocorrelation for *hvalue*. Figure 6a shows the global MC result, which indicates significantly strong and positive spatial autocorrelation (MC = 0.502 and p-value < 0.001 under a normality assumption). Figure 6c portrays a local MC map that exhibits the counties with significant local MC values (p-value < 0.05). These counties are differentially symbolized based on the different types of local spatial associations (i.e., high-high, high-low, low-high, and low-low). In the local MC map, the significant high-high clusters, which mean a high value is surrounded by neighboring high values, are founded in the four major MSAs in Texas, whereas the significant low-low clusters are displayed in north-central Texas. In addition, Kenedy County in southern Texas is the significant high-low spatial outlier, which means a high value is surrounded by low neighboring values. This outlier might occur due to relatively higher median house values in Padre Island compared with its surrounding regions.

SAAR also supports ESDA tools, including Conditioned Choropleth maps (Car et al. 2000), and a connectivity histogram and map (Anselin et al. 2006) for exploring spatial weights. In addition, a Moran scatter plot is implemented as a visual tool to explore the global MC (Anselin 1996) (Figure 6b), where the slope of a regression line represents a global MC value when using standardized variables (here, the slope = 0.502, and MC = 0.502 for *hvalue*). Also, SAAR furnishes a spatial correlogram tool to explore spatial autocorrelation at a particular spatial lag and its trend across spatial lags (Bailey and Gatrell 1995). Figure 7a shows a spatial correlogram based on MC for *hvalue*, where blue dots display local MC values at corresponding spatial lags. Bold lines at the center of box plots represent the averages of local MC values at each spatial lag, which corresponds to the global MC values (Anselin 1995), and a red horizontal dotted line shows the expected value of MC at corresponding spatial lags. In Figure 7a, the spatial correlogram illustrates positive spatial autocorrelation at the first order spatial lag, with spatial autocorrelation decreasing as the order of spatial lags increases.

SARR provides advanced dynamic linking and brushing, and a Moran scatter plot and a spatial correlogram can be used to illustrate this technique. That is, one point in both graphs is brushed on a map with not only a corresponding location, but also its neighbors. The advanced dynamic linking and brushing help to explore a spatial configuration especially at higher order spatial lags and for distance-based spatial weights (e.g., k-nearest neighbors and Delaunay triangulations). For example, Figures 6c and 7b clearly show the selected counties and their first order and fourth order queen's case neighbors, respectively. Furthermore, in a spatial correlogram (e.g., Figure 7a), the selected observations at a particular spatial lag also are connected through lines to the same observations at other spatial lags, visually highlighting spatial autocorrelation trends for the observations.

5.4 Data transformations and regression analysis

SAAR supports various regression analysis tools, including simple linear regression, GLM, various spatial regressions, and MESF for linear and GLM specifications. In addition, SAAR supports a Box-Cox transformation tool to convert an input variable to one that better mimics a normal distribution (i.e., bell-shaped curve). Figure 8 demonstrates the Box-Cox transformation tool and its result for *hvalue*. The maximum likelihood estimate of the exponent parameter for this Box-Cox transformation is –0.24. The transformed variable conforms reasonably well to a normal distribution based on the normal quantile-quantile (QQ) plot, histogram, and Shapiro-Wilk test (0.992) compared to the positively skewed distribution of the variable before its being subjected to this transformation (see Figure 5c).

Figure 9a reports the result of linear regression with the Box-Cox transformed *hvalue* as the dependent variable, and *year* and *unemploy* as independent variables. Results of the linear regression analysis reveal a positive relationship between *hvalue* and *year*, and a negative relationships between *hvalue* and *unemploy*. That is, the coefficient estimates of *year* and *unemploy* are 0.002 and –0.003, respectively. Also, these two covariates explain a sizeable proportion of the variance in the dependent variable, with an adjusted-R^2 of 0.576. However, the linear regression residuals have significant positive spatial autocorrelation (MC = 0.261 and *p*-value < 0.001) (Figure 9a), which indicates that the linear regression specification should be replaced with a spatial regression specification. The map of the linear regression residuals (Figure 9b) also clearly displays spatial structure in the regression residuals. This result implies that a spatial regression model is necessary to properly account for unexplained spatial autocorrelation in the residuals of the linear regression.

Figure 10 illustrates spatial regression results for a SAR model (i.e., spatial error model), with the same variables. The linear and SAR specifications yield similar estimated coefficient magnitudes with the same signs. However, the SAR (Figure 10a) clearly shows an improvement compared to the linear regression (Figure 9a). The estimated spatial autocorrelation parameter of the SAR specification is significant ($\lambda = 0.471$ and *p*-value < 0.001), which coincides with the positive spatial autocorrelation in the linear regression residuals. The SAR specification properly accounts for spatial autocorrelation, and spatial structure is not observed in the SAR residual map (Figure 10b).

SAAR also furnishes an MESF tool for linear and GLM models. Figure 11 presents the results of MESF for linear regression with the same variables. In Figure 11a, the MESF selects 27 of 59 candidate eigenvectors[2], and, with the selected eigenvectors, shows a considerable improvement compared to the conventional linear regression results (see Figure 9a). Specifically, spatial autocorrelation in the linear regression residuals is successfully accounted for by the MESF model (MC = −0.133 and *p*-value = 0.892), which also is suggested by no prominent spatial structure in the residuals map from the MESF analysis (Figure 11b). In addition, the MESF model has a better fit than the linear regression model; the MESF model has a much higher adjusted-R^2 value (0.762 > 0.579).

[2]The candidate eigenvector set in this application was determined by the equation proposed in Chun et al. (2016).

5.5 Geovisualization and map classification incorporating uncertainty information

SAAR provides toolsets to explore uncertainty in spatial data and SDA output in two different contexts: geovisualization, and map classification. Extending bivariate mapping techniques, geovisualization tools in SAAR are implemented in three different ways: the coloring properties to proportional symbols (CPPS), overlaid symbols on a choropleth map (OSCM), and composite symbol (CS) methods (Koo et al. 2018). Figures 12a and 12b visualize the estimates of *hvalue* and their corresponding uncertainty levels (i.e., CV) using OSCM and CPPS, respectively. In Figure 12a, a choropleth map represents the estimates in color, and the overlaying textures portray their uncertainty levels: a smaller spacing (more dense symbols) denotes a higher level of uncertainty. In Figure 12b, the symbol sizes of the circles represent the estimates, and lightness denotes their uncertainties: a lighter color represents a higher uncertainty.

Although uncertainty visualization methods offer map users reliability information about estimates and its spatial pattern in a choropleth map (see Figures 12a and 12b), this spatial pattern still might be unreliable because the map classification is constructed in the presence of sampling error (Sun et al. 2017). Hence, SAAR furnishes map classification methods incorporating uncertainty information based on the separability criterion devised by Sun et al. (2014), and also on optimal classification methods (Koo et al. 2017). Figure 13 exhibits results of the two map classification methods for the estimates of *hvalue* with their uncertainty information. Figure 13a shows the result of the separability method. This map classification method is useful for highlighting statistical outliers incorporating uncertainty information because it heuristically maximizes the statistical difference between classes (Sun et al. 2014). Meanwhile, Figure 13b, the outcome of an optimal classification method, is based upon minimizing a total sum of pairwise Bhattacharyya distance in a class, and hence achieves a homogeneity among the classes, simultaneously accounting for estimates and uncertainty information. It also has a more balanced number of observations (i.e., counties) for each class compared with its separability based counterpart. Detailed discussions of these methods can be found in Sun et al. (2014), and Koo et al. (2017).

6. Software evaluation

A focus group was convened to evaluate the usefulness and requirements of SAAR. The group consisted of thirteen graduate students, enrolled in both master's and Ph.D. programs of study, with different backgrounds and experiences working with spatial data. Prior to the survey, the functionalities of SAAR were introduced, and the participants replicated the SDA described in the previous application section with the same empirical dataset: the three variables of *hvalue*, *unemploy*, and *year* for the 254 counties in Texas. The participants explored SAAR further, conducting a SDA with their own datasets, and then they each completed an evaluation questionnaire. Specifically, this questionnaire consists of six open questions. The first three questions ask for background information about the participants: their majors, educational backgrounds, and SDA and GIS usage experiences. The fourth question asks about the usefulness of SAAR for their research. And, the fifth and sixth questions seek user input pertaining to additional functionality and the fulfillment of design goals, respectively.

The participants have various disciplinary backgrounds: economics, public policy, statistics, and marketing, as well as GIS. Nine participants were Ph.D. students, and the remaining four were master's students. Six participants, the non-GIS students, had less than one year of working experiences with spatial data. With regard to the usefulness of SAAR in their fields of study, overall the participants indicate that they found SAAR to be useful and easy-to-use for SDA. Specifically, the practical usefulness generally was discussed in terms of two aspects. First, the participants from economics and statistics pointed out the usefulness of SAAR because of its easier-to-use toolset for spatial statistical analyses, especially spatial regression models and MESF, vis-à-vis R. Second, the GIS participants placed more emphasis on the effective data visualization tools in SAAR (e.g., dynamic linking and brushing) supporting examinations of spatial relationships in data. Also, they mentioned that R and other statistics software require considerable effort to investigate such relationship.

The participants' responses for additional desirable functionality in SAAR suggest three broad themes: an extension of existing functions, inclusion of additional functions, and the supporting of various data types. First, the participants suggest extending the functionalities of existing tools. For example, they stressed that distance-based spatial weights (e.g., Euclidean, inverse, and great circle distances) can be useful to explore spatial associations at a given spatial scale. In addition, they recommended an easy user interface to generate an eigenvector map, which can be a useful tool for MESF, because it can represent a wide spectrum of spatial autocorrelation scenarios (e.g., Griffith 2003). Also, the limited mapping functions, which support a fixed style of each map component (e.g., scale bar, north arrow, and legend), was singled out for improvement so that users can prepare a high quality map with more styles and options. Second, additional functions for SAAR that are particularly relevant to the participants' disciplines were suggested. Specifically, the GIS participants suggested supporting interpolation methods (e.g., kriging) and a semi-variogram to furtherly explore the degree of spatial autocorrelation along with distances between spatial units. The participants from marketing recommended inclusion of spatial analysis tools for point data, with specific mention of cluster detection methods. The participants from statistics underlined the importance of spatial panel analysis tools (e.g., Elhorst 2010). Finally, SAAR currently supports the shapefile format for SDA input, which should be augmented by support for more GIS data formats. With regard to the question about the fulfillment of GUI design goals, there is a clear consensus among the participants that the simplicity of SAAR's GUI helped them to find appropriate tools easily, and, furthermore, they suggested preserving this GUI simplicity as much as possible in future extensions of SAAR.

7. Concluding remarks

This paper presents SAAR, which provides an integrated environment for GIS and SDA. SAAR can benefit both general and advanced GIS users. For the former, SAAR provides powerful ESDA and CSDA techniques as well as fundamental GIS functions and geovisualization methods in a popular GIS environment with a user-friendly GUI. For the latter, SAAR can provide a flexible framework to help users customize and extend SDA functions in a popular GIS environment with relatively little development effort. This possibility fundamentally is enabled with the framework of SAAR, the integration of the high extensibility of R (Bivand et al. 2013b), and a simple integration process between R

and .NET. Furthermore, because of its open source environment, SAAR can function as a platform with which an advanced GIS user easily can add more SDA tools on his/her own choosing.[3]

SAAR still is under active development and will be further extended in the future. First, making the source code of SAAR available as cross-platform software will be a definite way to increase accessibility to this application for end-users. Currently, the visualization and the analytic engines for GIS functionalities in SAAR are developed using proprietary libraries in ESRI's ArcGIS Engine, which is available only on the MS Windows platform. This change to a cross-platform format requires considerable effort to modify the GUIs and internal functions for visualization and GIS functions. Second, like ArcGIS for Desktop, ArcGIS Engine is a native 32-bit application, which makes SAAR run as a 32-bit application even on 64-bit MS Windows. Thus, SAAR is limited in handling large datasets because of the accompanying computer memory size limitation of the 32-bit Windows systems. Third, the SDA functionalities need to be further extended. Specifically, implementing nonlinear regression (e.g., negative binomial) together with its MESF specification would allow the modelling of various types of variables. In addition, an extension of the Jacobian approximation methods (e.g., Griffith and Sone 1995) would bolster the spatial regression tool's capabilities with large datasets. Fourth, more SDA functionalities for spatio-temporal data analysis are desirable, ones similar to STARS (Rey and Janikas 2006) and SpaceStat (Jacquez et al. 2014). Because of the high extensibility of the analytic engine in SAAR, implementation of spatio-temporal analysis functions can be achieved relative easily. However, due to a limited capability to handle temporal data in a GIS, an additional visualization engine is necessary to efficiently display spatial-temporal data (e.g., a 3-dimensional display), although SAAR already supports interactivity of the various charts and a map view to examine data across various dimensions. Finally, the integration based on R.NET may have a potential issue arising from the support of R.NET in the future. However, the R.NET source code is available, which allows researchers to address potential issues such as supporting a newer version of software. Furthermore, this limitation merits further future investigation with regard to other interoperability bridge programs (e.g., statconn.NET[4]).

References

Andrienko G, Andrienko N, and Voss H 2003 GIS for everyone: the CommonGIS project and beyond. In: Peterson MP (ed). Maps and the Internet. Elsevier Science: Amsterdam, 131–46.

Andrienko N, Andrienko G, Voss H, Bernardo F, Hipolito J, and Kretchmer U 2002 Testing the usability of interactive maps in CommonGIS. Cartography and Geographic Information Science 29: 325–42.

Anselin L 1988 Spatial Econometrics: Methods and Models. Kluwer Academic Publishers: Boston.

Anselin L 1995 Local indicators of spatial association — LISA. Geographical Analysis 27: 93–115.

Anselin L 1996 The Moran scatterplot as an ESDA tool to assess local instability in spatial association In: Fischer M, Scholten H, Unwin D (eds). Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences. Taylor & Francis: London, 111–25.

[3]The source code of SAAR will be made available in github upon publication of this manuscript. The installation file is available in https://utdallas.app.box.com/v/saar-b-102.

[4]http://www.autstat.com

Anselin L, and McCann M 2009 OpenGeoDa, open source software for the exploration and visualization of geospatial data. GIS '09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems: 550–1.

Anselin L, Syabri I, and Kho Y 2006 GeoDa: An introduction to spatial data analysis. Geographical Analysis 38: 5–22.

Bailey TC, and Gatrell AC 1995 Interactive Spatial Data Analysis. Longman Scientific & Technical: Essex, England.

Bivand R, and Gebhardt A 2000 Implementing functions for spatial statistical analysis using the R language. Journal of Geographical Systems 2: 307–17.

Bivand R, Hauke J, and Kossowski T 2013a Computing the jacobian in gaussian spatial autoregressive models: An illustrated comparison of available methods. Geographical Analysis 45: 150–79.

Bivand RS 2000 Using the R statistical data analysis language on GRASS 5.0 GIS database files. Computers and Geosciences 26: 1043–52.

Bivand RS 2002 Spatial econometrics functions in R: classes and methods. Journal of Geographical System 4: 405–21.

Bivand RS, Pebesma EJ, and Gómez-Rubio V 2013b Applied Spatial Data Analysis with R. Springer: New York.

Brunsdon C, Fotheringham S, and Charlton M 1998 Geographically Weighted Regression-Modelling Spatial Non-Stationarity. The Statistician 47: 431–43.

Chun Y, Griffith DA, Lee M, and Sinha P 2016 Eigenvector selection with stepwise regression techniques to construct eigenvector spatial filters. Journal of Geographical Systems 18: 67–85.

Delmelle E, Delmelle EC, Casas I, and Barto T 2011 H.E.L.P: A GIS-based Health Exploratory AnaLysis Tool for Practitioners. Applied Spatial Analysis and Policy 4: 113–37.

Dykes J 1998 Cartographic visualization: exploratory spatial data analysis with local indicators of spatial association using Tcl/Tk and cdv. The Statistician 47: 485–97.

Elhorst JP 2010 Spatial panel data models In: Fischer MM, Getis A (eds). Handbook of Applied Spatial Analysis. Springer Berlin Heidelberg: Berlin, Heidelberg, 377–407.

ESRI (Environmental Systems Research Institute) 2016 ArcGIS Desktop: release 10.5. Environmental Systems Research Institute: Redlands, California, USA.

Fisher P 1993 Visualizing uncertainty in soil maps by animation. Cartographica 30: 20–7.

Getis A, and Ord JK 1992 The analysis of spatial association by use of distance statistics. Geographical Analysis 24: 189–206.

Goodchild M, Haining R, and Wise S 1992 Integrating GIS and spatial data analysis: problems and possibilities. International journal of geographical information systems 6: 407–23.

Griffith DA 2003 Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer: Berlin.

Griffith DA, and Sone A 1995 Trade-offs associated with normalizing constant computational simplifications for estimating spatial statistical models. Journal of Statistical Computation and Simulation 51: 165–83.

Haining R, Wise S, and Ma J 1998 Exploratory spatial data analysis in a geographic information system environment. The Statistician 47: 457–69.

Hintze JL, and Nelson RD 1998 Violin plots: A box plot-density trace synergism. American Statistician 52: 181–4.

Jacquez GM, Goovaerts P, Kaufmann A, and Rommel R 2014 SpaceStat 4.0 user manual: Software for the space-time analysis of dynamic complex systems.

Jenks GF 1977 Optimal Data Classification for Choropleth Maps. Occasional Paper No. 2, Department of Geography, University of Kansas.

Koo H, Chun Y, and Griffith DA 2017 Optimal map classification incorporating uncertainty information. Annals of the American Association of Geographers 107: 575–90.

Koo H, Chun Y, and Griffith DA 2018 Geovisualizing attribute uncertainty of interval and ratio variables: A framework and an implementation for vector data. Journal of Visual Languages & Computing 44: 89–96. [PubMed: 29503517]

Lee S, II 2001 Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. Journal of Geographical Systems 3: 369–85.

Lee S, II 2004 A generalized significance testing method for global measures of spatial association: An extension of the Mantel test. Environment and Planning A 36: 1687–703.

Lee S, II 2009 A generalized randomization approach to local measures of spatial association. Geographical Analysis 41: 221–48.

LeSage JP 1999 Applied econometrics using MATLAB. University of Toronto.

Levine N 2010 CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v 3.3). Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC 7.

NET R. R.NET for users 1.6. 2015. Retrieved from http://jmp75.github.io/rdotnet/.

Rey SJ, and Janikas MV. 2006 STARS: Space-Time Analysis of Regional Systems. Geographical Analysis 38: 67–86.

Rheingans P 1992 Color, change, and control of quantitative data display In: Visualization '92. IEEE Computer Society Technical Committee on Computer Graphics: Boston, Massachusetts, 252–9.

Rura MJ, and Griffith DA 2010 Spatial statistics in SAS In: Getis A, Fischer MM (eds). Handbook of Applied Spatial Analysis. Springer-Verlag: Berlin, 43–52.

Slocum TA, McMaster RB, Kessler FC, and Howard HH 2009 Thematic Cartography and Geovisualization. Prentice Hall: Upper Saddle River, New Jersey.

Sun M, Wong DW, and Kronenfeld B 2014 A classification method for choropleth maps incorporating data reliability information. The Professional Geographer 67: 72–83.

Sun M, Wong DW, and Kronenfeld B 2017 A heuristic multi-criteria classification approach incorporating data quality information for choropleth mapping. Cartography and Geographic Information Science 44: 246–58. [PubMed: 28286426]

Symanzik J 2014 Exploratory spatial data analysis. In: Fischer MM, Nijkamp P (eds). Handbook of Regional Science. Springer: Heidelberg, 1295–310.

Symanzik J, Cook D, Lewin-Koh N, Majure JJ, and Megretskaia I 2000 Linking ArcView and XGobi: Insight behind the Front End. Journal of Computational and Graphical Statistics 9: 470–90.

Unwin A, and Hofmann H 1997 New interactive graphics tools for exploratory analysis of spatial data In: Carver S (ed). Innovations in GIS. Taylor & Francis: London, UK, 46–55.

Vandergast AG, Perry WM, Lugo RV., and Hathaway SA 2011 Genetic landscapes GIS toolbox: Tools to map patterns of genetic divergence and diversity. Molecular Ecology Resources 11: 158–61. [PubMed: 21429115]

Yamada I, Rogerson PA, and Lee G 2009 GeoSurveillance: A GIS-based system for the detection and monitoring of spatial clusters. Journal of Geographical Systems 11: 155–73.

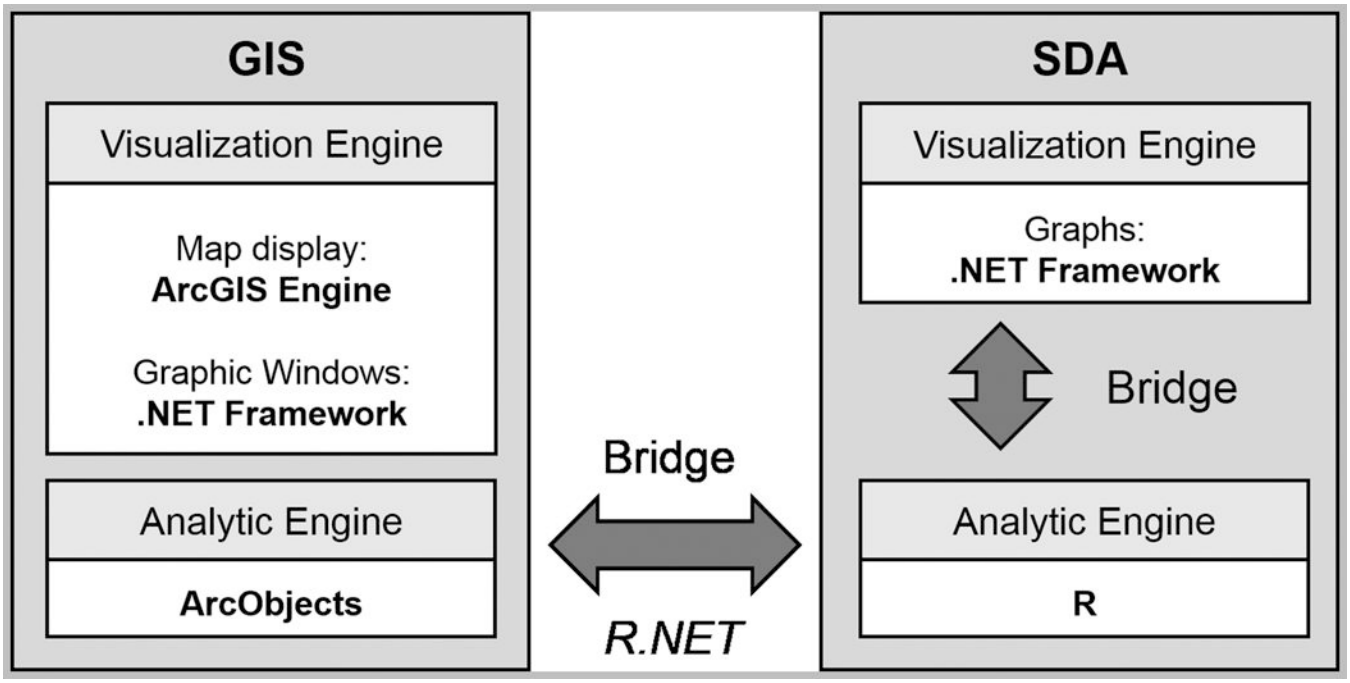

Figure 1.
The basic architecture of SAAR

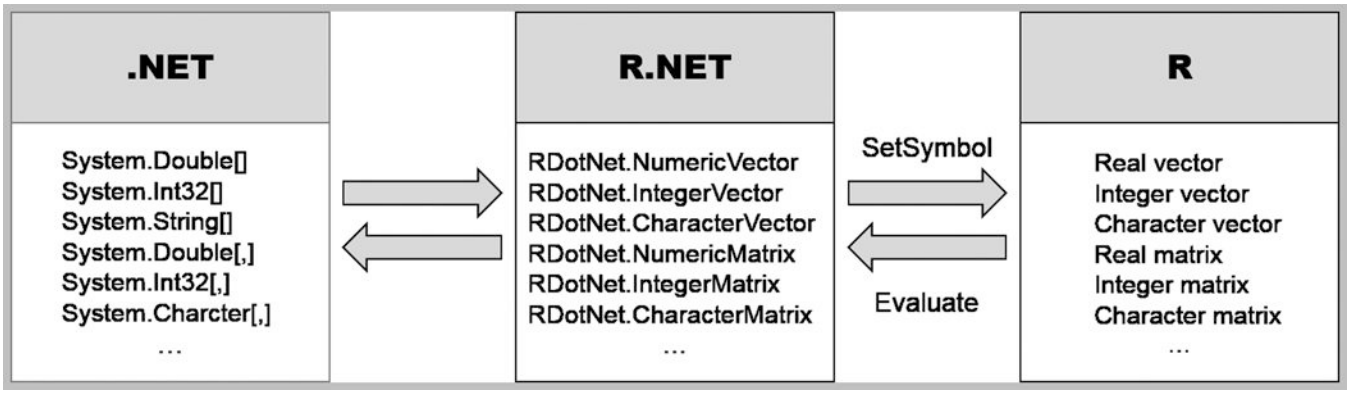

Figure 2.
The integration process between R and .NET Framework using R.NET

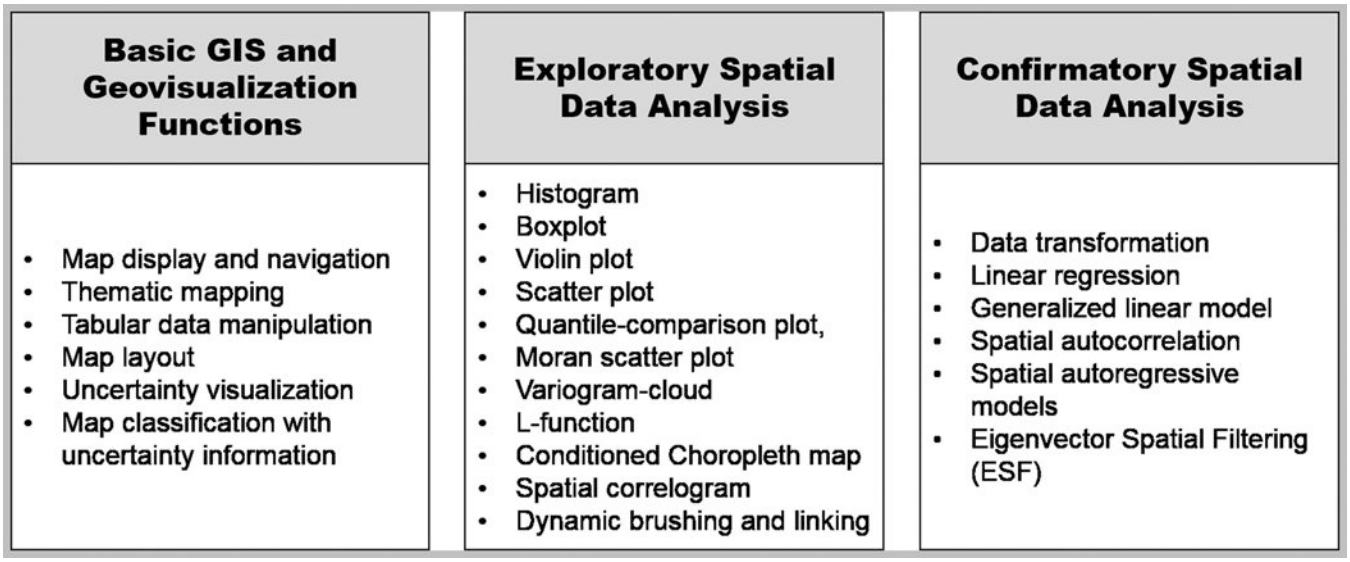

Figure 3.
An overview of the existing SAAR functions

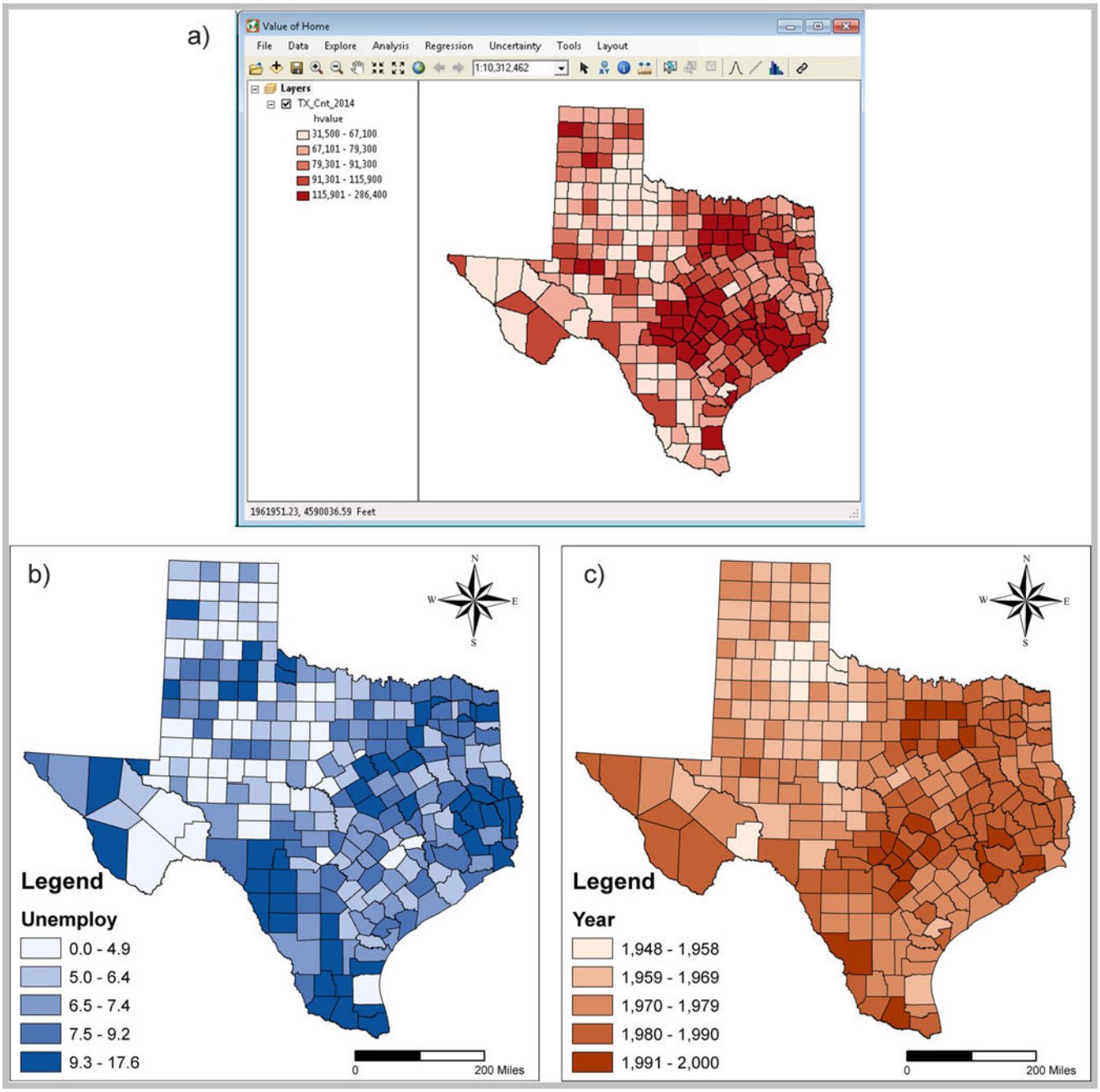

Figure 4.

The graphic user interface of SAAR, and choropleth maps of selected variables: a) median house values, b) unemployment rates, and c) median years in which a structure was built

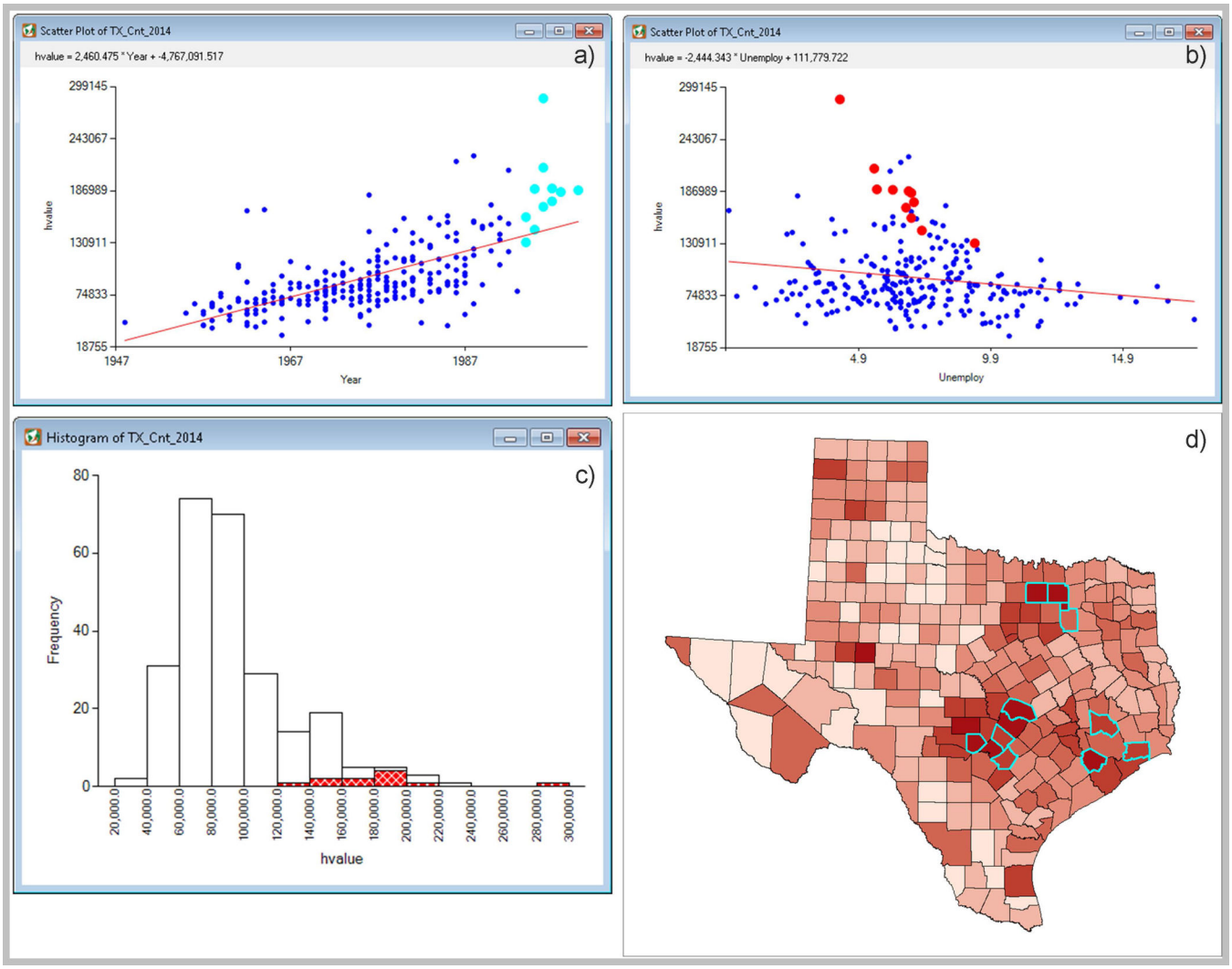

Figure 5.

Exploratory data tools with dynamic brushing and linking: a) the scatter plot for median house value and median year in which a structure was built, b) the scatter plot for median house value and unemployment rate, c) the histogram of median house value, d) the choropleth map of median house value

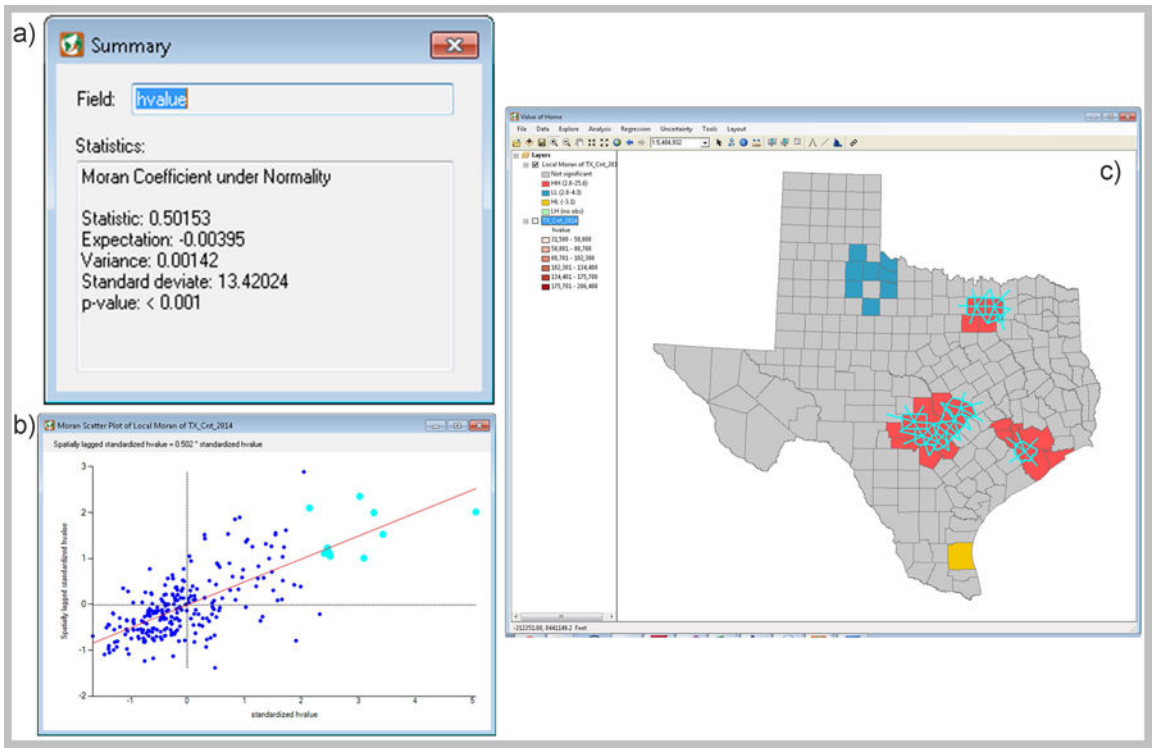

Figure 6.

Spatial autocorrelation in median house values: a) global Moran Coefficient, b) Moran scatter plot, c) the cluster map of local Moran Coefficients (p-value < 0.05)

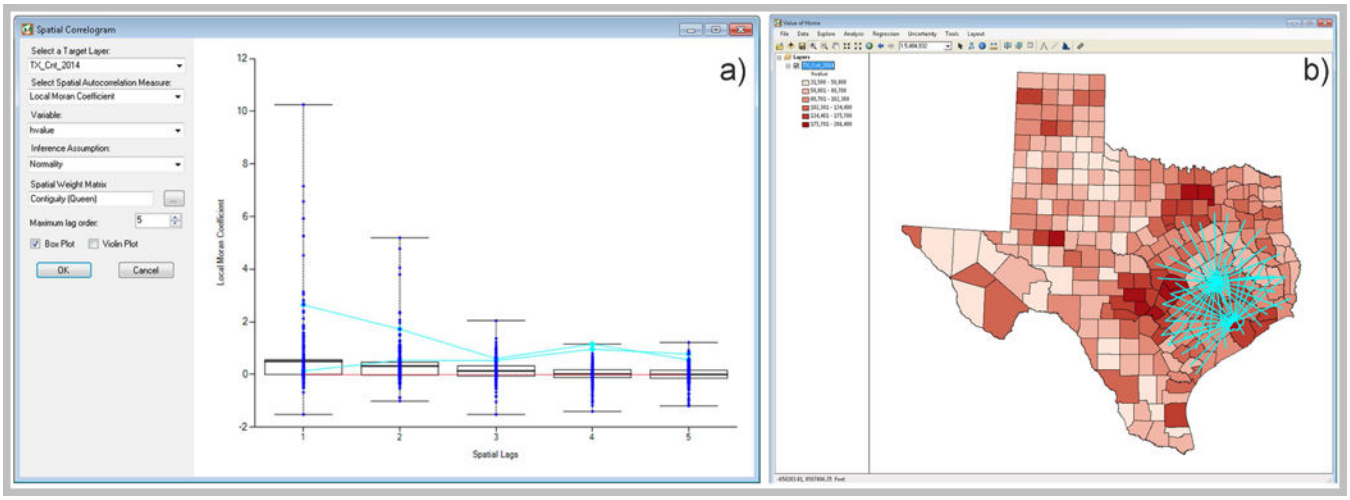

Figure 7.

A spatial correlogram with advanced dynamic linking and brushing for median house value: a) spatial correlogram, b) brushing on the linked map

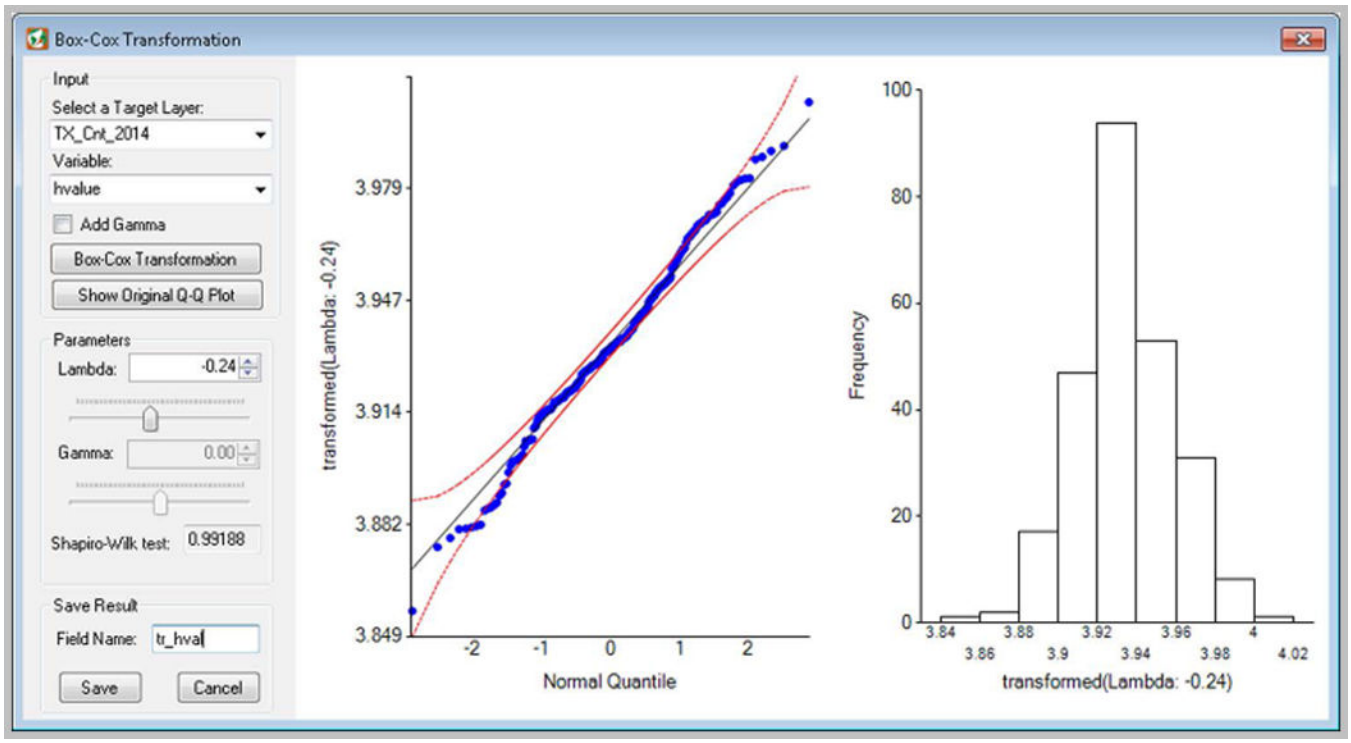

Figure 8.
Box-Cox transformed median house values

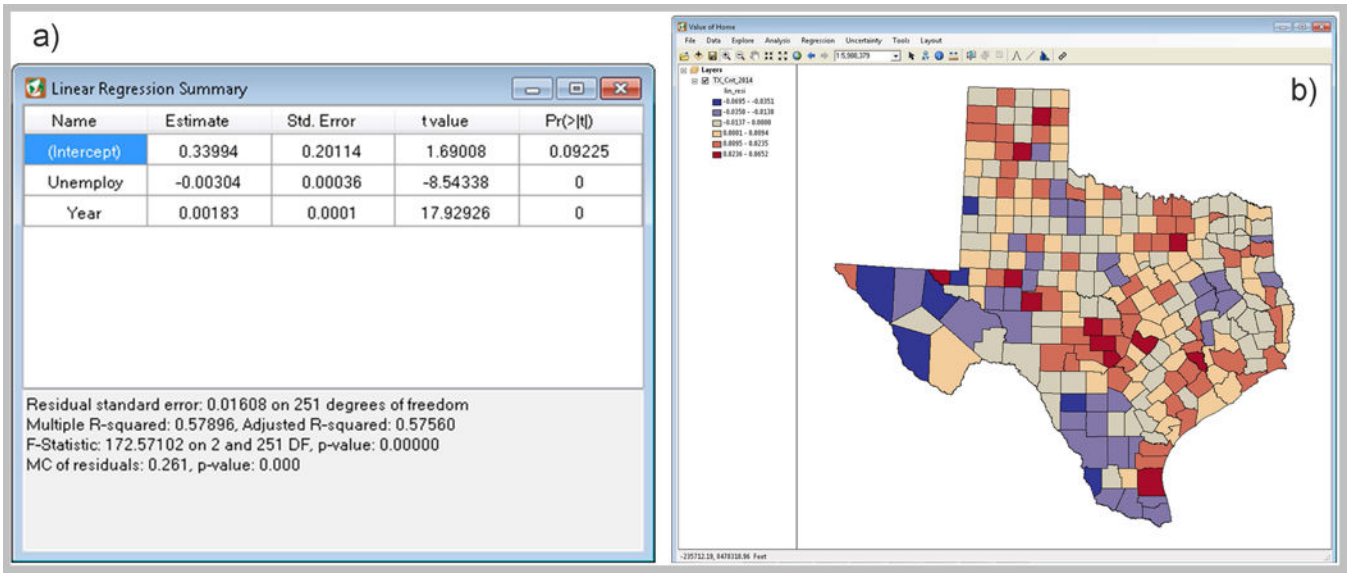

Figure 9.

A linear regression analysis result: a) a summary of the linear regression results, b) a map of the linear regression residuals

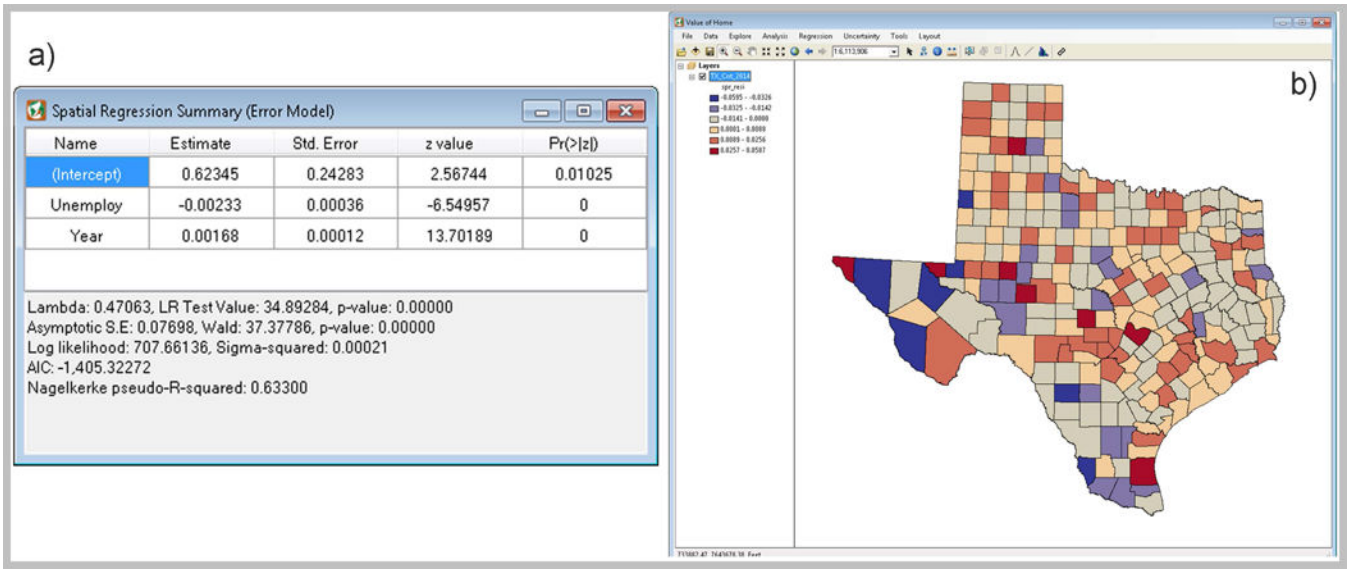

Figure 10.
A SAR analysis result: a) a summary of the SAR results, b) a map of the SAR residuals

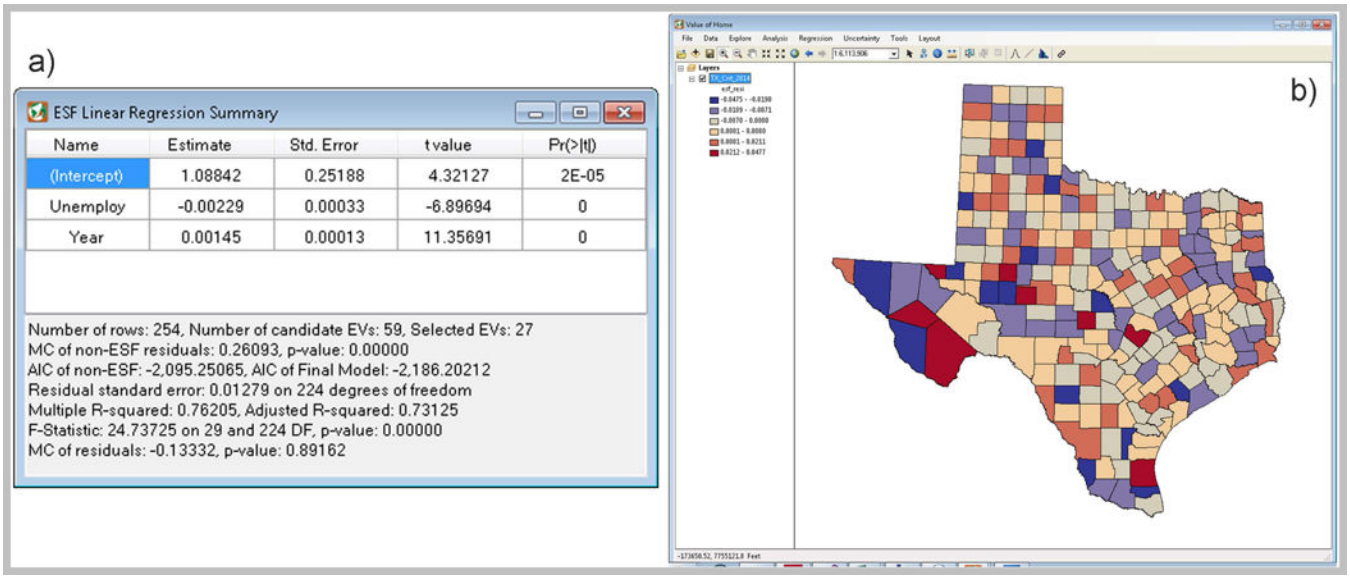

Figure 11.
A MESF analysis result: a) a summary of the MESF results, d) a map of the MESF residuals

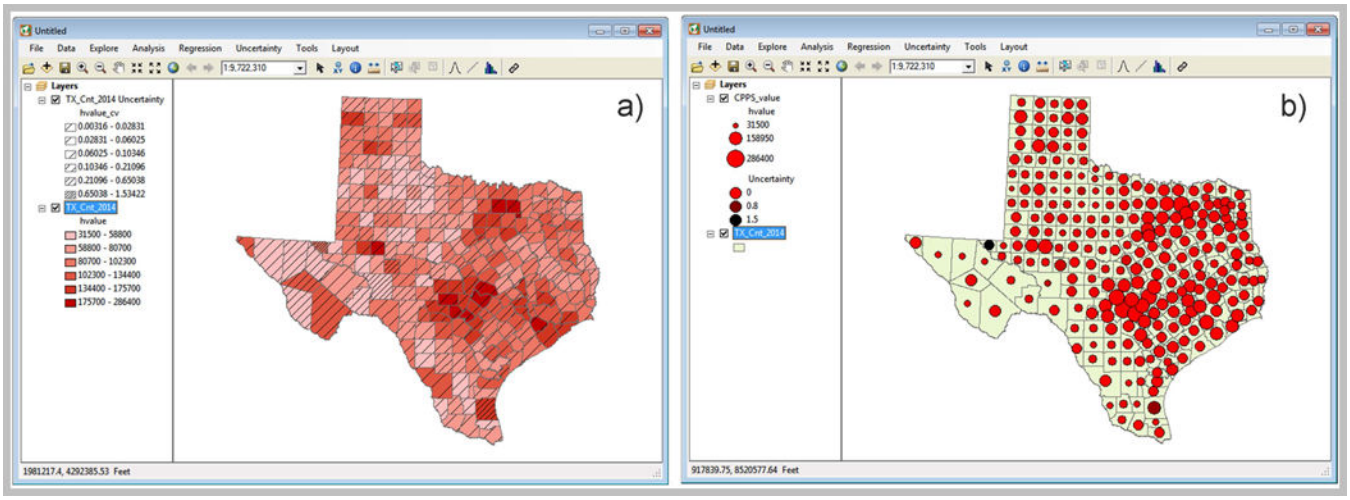

Figure 12.

Geovisualization methods for uncertainty exploration: a) overlaid symbols on a choropleth map, b) coloring properties combined with proportional symbols

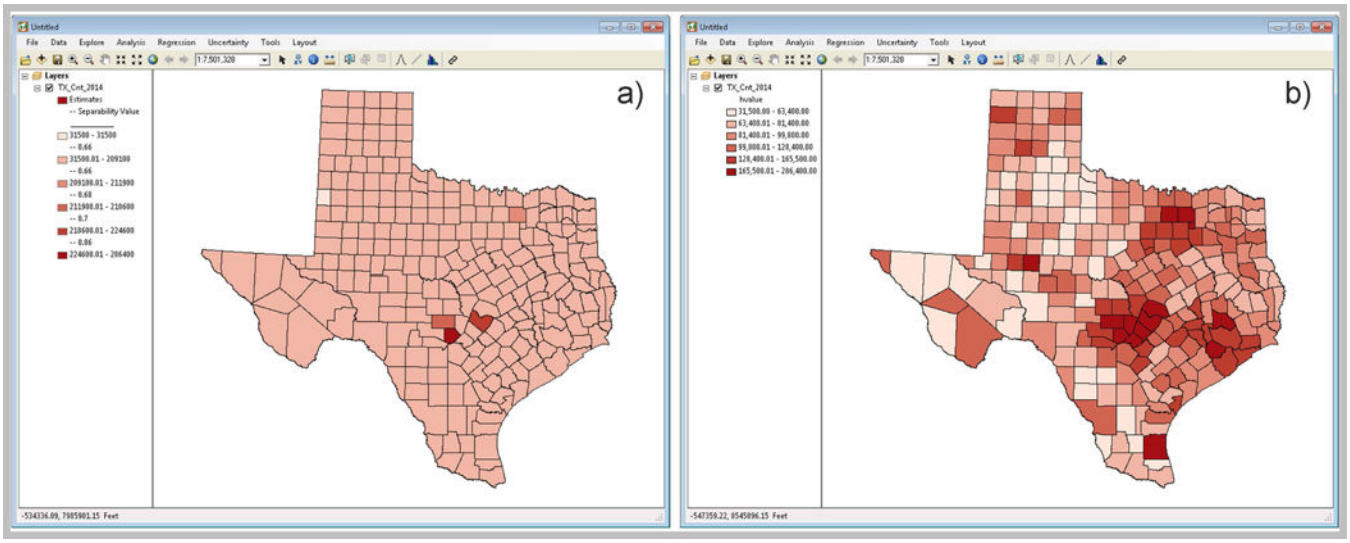

Figure 13.

Map classification methods incorporating uncertainty information: a) the separability method (Sun et al. 2014), b) the optimal classification method with Bhattacharyya distance (Koo et al. 2017)

Table 1.

Descriptive statistics of the selected variables for the 254 counties in Texas

Variables	Min	Max	Mean	STD	Median	IQR
Median house value	31,500	286,400	94,342	37,872	85,200	40,300
Median year in which a structure was built	1948	2000	1976	10.1	1976	14
Unemployment rate	0	17.6	7.1	2.9	6.9	3.3