GEODA: AN INTRODUCTION TO SPATIAL DATA ANALYSIS

EE464: Urban Economics
EE562: Selected Topics in Development Economics 2
Semester 1 / 2020
Faculty of Economics, Thammasat University

Main contents

(1) Design and Functionality
(2) GeoDa Functionality Overview
(3) Mapping and Geovisualization
(4) Multivariate EDA
(5) Spatial Autocorrelation Analysis
(6) Spatial Regression

(1) Design and Functionality

In broad terms, the functionality can be classified into six categories:

- **Spatial data manipulation and utilities:**
  - data input, output, and conversion
- **Data transformation:**
  - variable transformations and creation of new variables
- **Mapping:**
  - choropleth maps, cartogram and map animation
- **EDA: statistical graphics**
- **Spatial autocorrelation:**
  - global and local spatial autocorrelation statistics, with inference and visualization
- **Spatial regression:**
  - diagnostics and maximum likelihood estimation of linear spatial regression

Source: Directly quoted from Anselin, L., Syabri, I, and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(2) GeoDa Functionality Overview

Spatial Data

- Data input from shape file (point, polygon)
- Data input from text (to point or polygon shape)
- Data output to text (data or shape file)
- Create grid polygon shape file from text input
- Centroid computation
- Thiessen polygons

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(2) GeoDa Functionality Overview

Data Transformation

- Variable transformation (log, exp, etc.)
- Queries, dummy variables (regime variables)
- Variable algebra (addition, multiplication, etc.)
- Spatial lag variable construction Spatial lag variable construction
- Rate calculation and rate smoothing
- Data table join

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(2) GeoDa Functionality Overview

Mapping

- Generic quantile choropleth map
- Generic quantile choropleth map
- Standard deviational map
- Percentile map
- Outlier map (box map)
- Circular cartogram
- Map movie
- Conditional maps
- Smoothed rate map (EB, spatial smoother)
- Excess rate map (standardized mortality rate, SMR)

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(2) GeoDa Functionality Overview

EDA histogram

- Box plot
- Scatter plot
- Parallel coordinate plot
- Three-dimensional scatter
- Conditional plot (histogram, box plot, scatter plot)

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(2) GeoDa Functionality Overview

Spatial Autocorrelation spatial weights creation (rook, queen, distance, k-nearest)

- Higher order spatial weights
- Spatial weights characteristics (connectedness histogram)
- Moran scatterplot with inference
- Bivariate Moran scatterplot with inference
- Moran scatterplot for rates (EB standardization)
- Local Moran significance map
- Local Moran cluster map
- Bivariate Local Moran
- Local Moran for rates (EB standardization)

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(2) GeoDa Functionality Overview

Spatial Regression

- OLS with diagnostics (e.g., LM test, Moran's I)
- Maximum Likelihood spatial lag model
- Maximum Likelihood spatial error model
- Predicted value
- Residual map

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(3) Mapping and Geovisualization

- The bulk of the mapping and geovisualization functionality consists of a collection of specialized **choropleth maps**, focused on highlighting **outliers** in the data, **so-called box maps** (Anselin 1999).
- For example, using data on prostate cancer mortality in 156 counties contained in the Appalachian Cancer Network (ACN) for the period 1993-97, a **box map** was constructed by specifying the number of deaths as the numerator and the population as the denominator.
- The resulting map for the **crude rates** (i.e., without any adjustments for differing age distributions or other relevant factors) is shown as the **upper-left panel** of Figure 1.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(3) Mapping and Geovisualization

- **Three counties** are identified as **outliers** and shown in **dark red**.
- These match the outliers selected in the box plot in the **lower-left panel** of the figure.
- The linking of all maps and graphs results in those counties also being **cross-hatched** on the maps.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

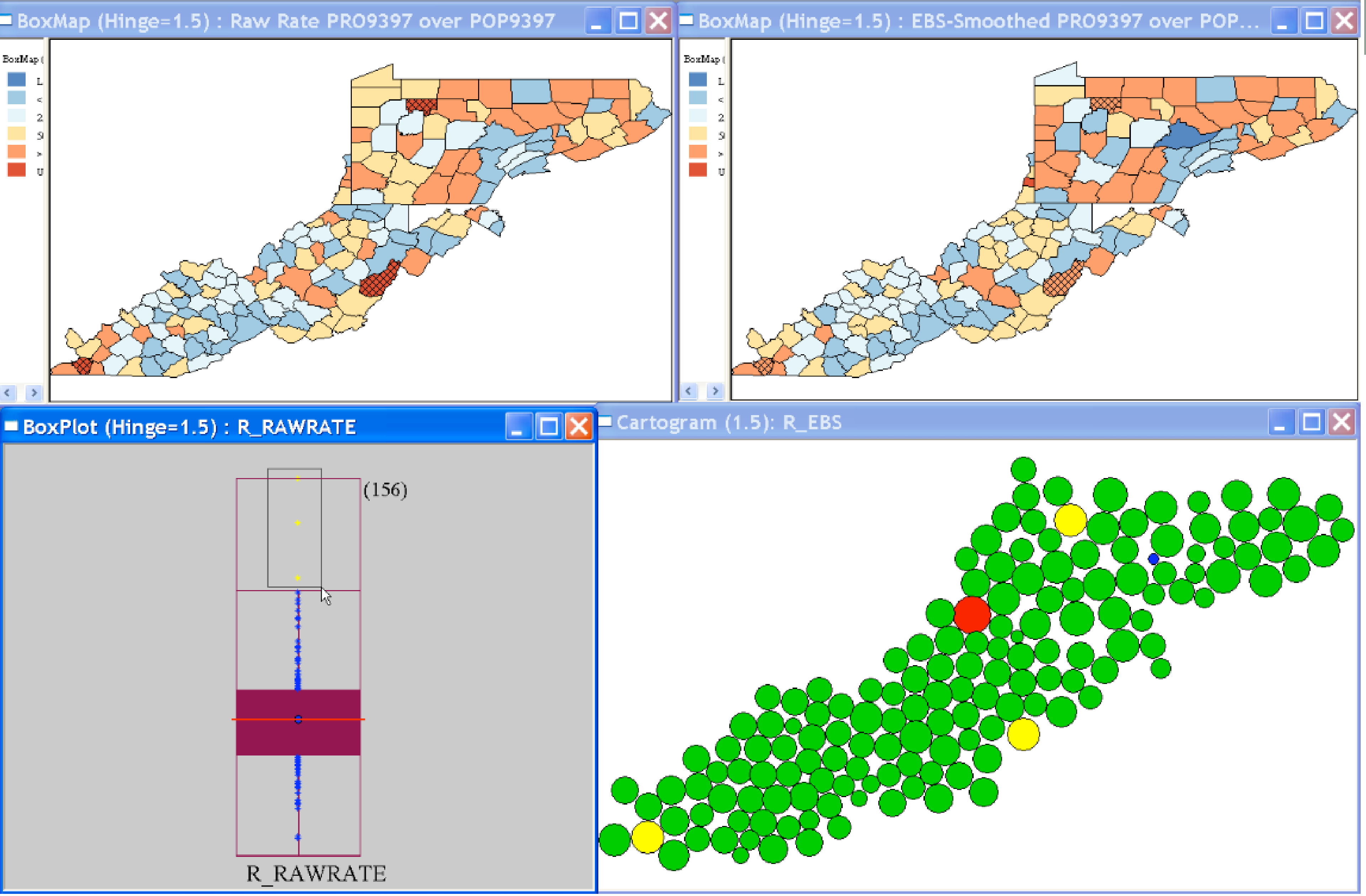

Figure 1: Linked box maps, box plot and cartogram, raw and smoothed prostate cancer mortality rates

(3) Mapping and Geovisualization

- In order to **remove the potentially misleading effect** of area on the perception of interesting patterns, a **circular cartogram** is shown in the **lower-right panel**, where the **area of the circles** is proportional to the value of the **EB smoothed rate**.
- The **upper outlier** is shown as a **red** circle, the **lower outlier** as a **blue** circle.
- The **yellow circles** are the counties that were **outliers** in the **crude rate map**, highlighted here as a result of **linking** with the other **maps and graphs**.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(4) Multivariate EDA

- **Multivariate exploratory** data analysis is implemented in GeoDa through **linking and brushing** between a collection of statistical graphs.
- These include the usual **histogram**, **box plot** and **scatter plot**, but also a parallel coordinate plot (PCP) and **three-dimensional scatter plot**, as well as conditional plots (conditional histogram, box plot and scatter plot).

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

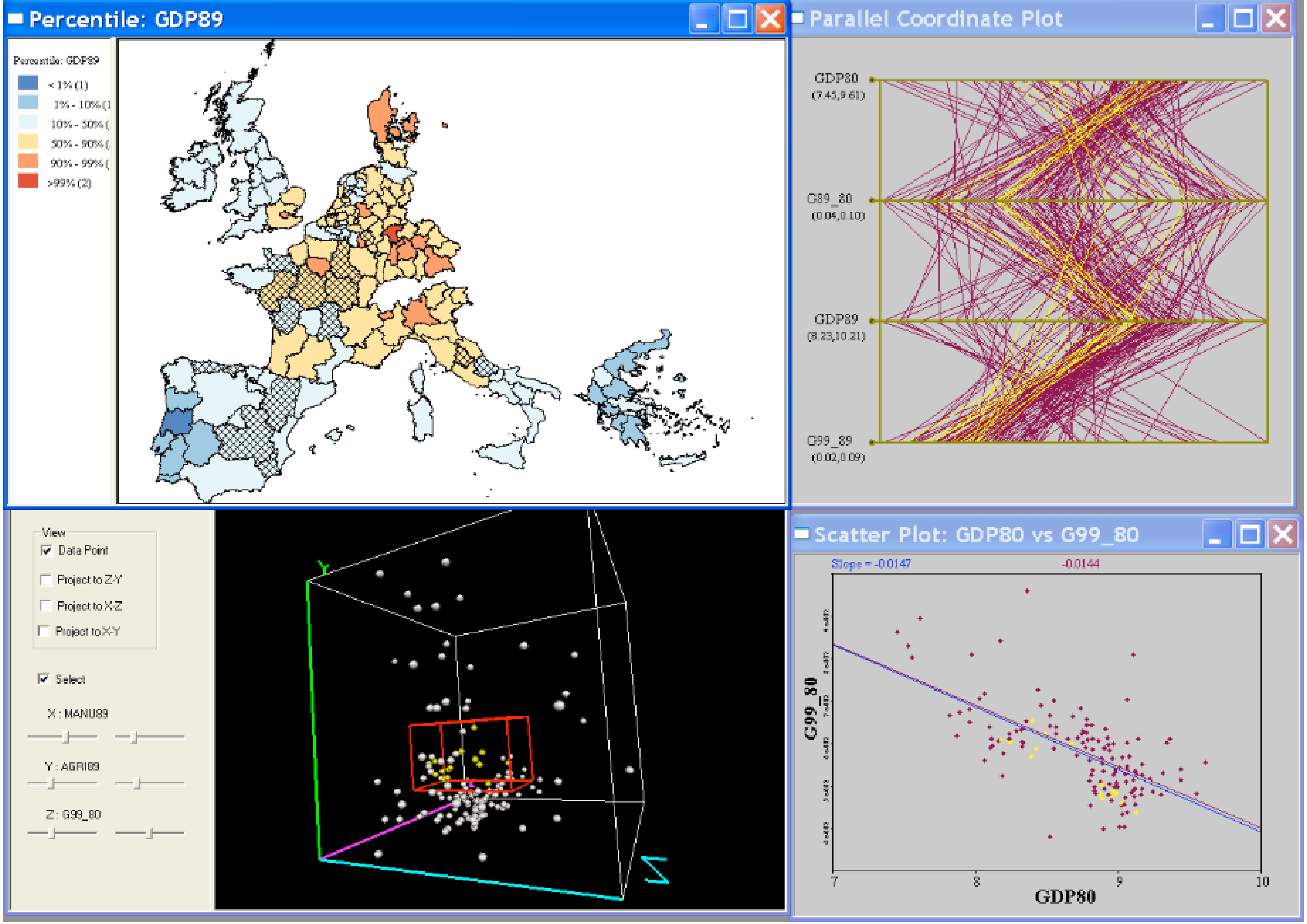

Figure 3: Multivariate exploratory data analysis with linking and brushing

Note on Data Set

- The data are from the most recent version of the **NewCronos Regio database** by Eurostat.
- NUTS stands for "Nomenclature of Territorial Units for Statistics" and contains the definition of **administrative regions** in the **EU member states**.
- NUTS II level regions are **roughly comparable to counties** in the U.S. context and are available for all but two countries.
- **Luxembourg** constitutes only a **single region**.
- For the **United Kingdom**, data is **not available** at the NUTS II level, since these regions do not correspond to local governmental units.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(4) Multivariate EDA

- We illustrate some of this functionality with an exploration of the **relationships** between **economic growth** and **initial development**, typical of the recent **“spatial” regional convergence literature** (for an overview, see Rey 2004).
- We use economic data over the period **1980-1999** for **145 European regions**, most of them at the NUTS II level of spatial aggregation, except for a few at the NUTS II level (for Luxembourg and the United Kingdom).

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(4) Multivariate EDA

- Figure 2 illustrates the various **linked plots and map**. The left-hand panel contains a **simple percentile map** (GDP per capital in 1989), and a **three dimensional scatter plot** (for the percent agricultural and manufacturing employment in 1989 as well as the GDP growth rate over the period 1980-99).
- In the top right-hand panel is a PCP for the **growth rates in the two periods of interest (1980-89 and 1989-99)** and the **GDP per capita** in the base year, the typical components of a **convergence regression**.
- In the bottom of the right-hand panel is a **simple scatter plot of the growth rate** in the full period (1980-99) on **the base year GDP**.
- Both plots on the right hand side illustrate the **typical empirical phenomenon that higher GDP at the start** of the period is associated with a lower growth rate.
- However, as demonstrated in the PCP (some of the lines suggest a positive relation between GDP and growth rate), the pattern is not uniform and there is a suggestion of heterogeneity.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(4) Multivariate EDA

- A further **exploration of this heterogeneity** can be carried out by brushing any one of these graphs.
- For example, in Figure 3, a selection box in the three-dimensional scatter plot is **moved around (brushing)** which highlights the selected observations in the map (cross-hatched) and in the PCP, clearly showing opposite patterns in subsets of the selection.
- Furthermore, in the scatter plot, the **slope of the regression** line can be **recalculated for a subset of the data** without the selected locations, to assess the sensitivity of the slope to those observations.
- In the example shown here, the effect on **convergence over the whole period is minimal (-0.147 vs. -0.144)**, but other selections show a more pronounced effect.
- Further exploration of these patterns does suggest a degree of spatial heterogeneity in the convergence results (for a detailed investigation, see Le Gallo and Dall'erba 2003).

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(5) Spatial Autocorrelation Analysis

- Spatial autocorrelation analysis includes tests and visualization of both **global (test for clustering)** and **local (test for clusters) Moran's I statistic**.
- The **global test** is visualized by means of a Moran scatterplot (Anselin 1996), in which the **slope of the regression line corresponds to Moran's I**.
- Significance is based on a permutation test. The traditional **univariate Moran scatterplot** has been extended to depict **bivariate spatial autocorrelation as well**, i.e., **the correlation between one variable at a location, and a different variable at the neighboring locations** (Anselin et al. 2002a).
- In addition, there also is an option to standardize rates for the potentially biasing effect of variance instability (see Assuncao and Reis 1999).

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(5) Spatial Autocorrelation Analysis

- Local analysis is based on the Local Moran statistic (Anselin 1995), visualized in the form of **significance** and **cluster maps**.
- It also includes several options for **sensitivity analysis**, such as changing the number of permutations (to as many as 9999), re-running the permutations several times, and changing the **significance cut off value**.
- This provides an ad hoc approach to assess the sensitivity of the results to problems due to multiple comparisons (i.e., how stable is the indication of clusters or outliers when the significance barrier is lowered).

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(5) Spatial Autocorrelation Analysis

- The maps depict the locations with **significant Local Moran statistics (LISA significance maps)** and classify those locations **by type of association (LISA cluster maps).**
- Both types of maps are available for brushing and linking.
- In addition to these two maps, the standard **output of a LISA analysis** includes a **Moran scatter plot** and a box plot depicting **the distribution of the local statistic**.
- Similar to the Moran scatter plot, the LISA concept has also been extended to a **bivariate setup** and includes an option to standardize for variance instability of rates.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(5) Spatial Autocorrelation Analysis

- The functionality for spatial autocorrelation analysis is **rounded out by a range of operations to construct spatial weights**, using either **boundary files (contiguity based)** or **point locations (distance based).**
- A **connectivity histogram** helps in **identifying potential problems** with the neighbor structure, such as **“islands”** (locations without neighbors).

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(5) Spatial Autocorrelation Analysis

- We illustrate spatial autocorrelation analysis with a study of the spatial distribution of **692 house sales prices for 1997 in Seattle**, WA.
- This is part of a broader investigation into the **effect of subsidized housing on the real estate** market.
- For the purposes of this example, we only focus on the **univariate spatial distribution**, and the location of any **significant clusters** or **spatial outliers** in the data.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

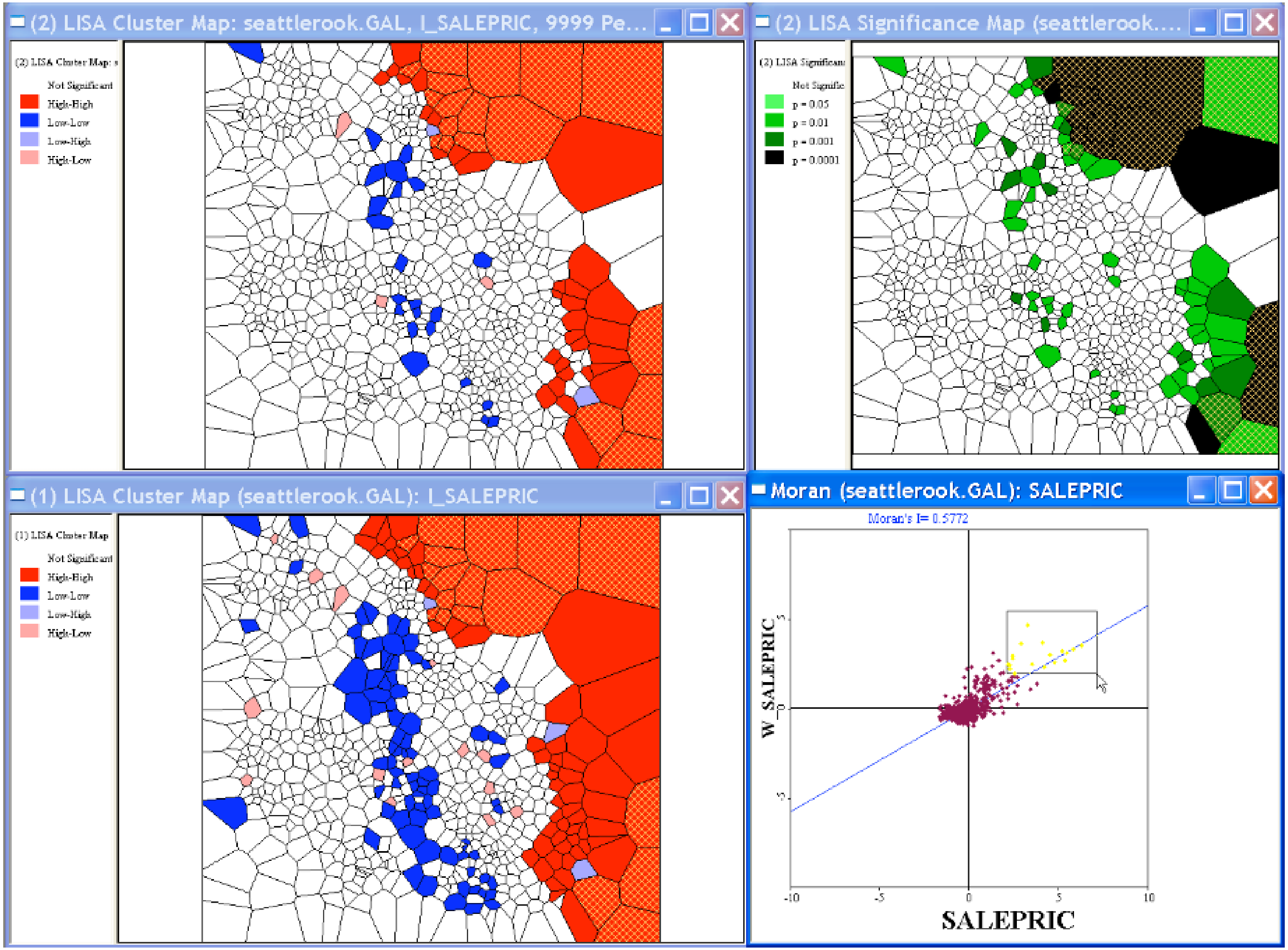

Figure 3: LISA cluster maps and significance maps

(5) Spatial Autocorrelation Analysis

- The original house sales data are for point locations, which, for the purposes of this analysis are converted to Thiessen polygons.
- This allows a **definition of “neighbor”** based on common boundaries between the **Thiessen polygons**.
- On the left hand panel of Figure 3, two LISA cluster maps are shown, depicting the locations of significant Local Moran’s I statistics, classified by type of spatial association.
- The **dark red** and **dark blue** locations are **indications of spatial clusters** (respectively, high surrounded by high, and low surrounded by low).
- In contrast, the **light red** and **light blue** are indications of **spatial outliers** (respectively, **high surrounded by low**, and **low surrounded by high**).

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(5) Spatial Autocorrelation Analysis

- The bottom map uses the **default significance of p = 0.05**, whereas the **top map is based on p = 0.01** (after carrying out 9999 permutations).
- The matching significance map is in the top right hand panel of Figure 3. **Significance** is indicated by **darker shades of green**, with the darkest corresponding to p = 0.0001.
- Note how the tighter significance criterion eliminates some (but not that many) locations from the map.
- In the bottom right hand panel of the Figure 3, the corresponding Moran scatterplot is shown, with the **most extreme "high-high"** locations selected.
- These are shown as **cross-hatched polygons in the maps**, and almost all obtain **highly significant (at p = 0.0001)** local Moran's I statistics.

Source: Directly quoted from Anselin, L., Syabri, I, and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(5) Spatial Autocorrelation Analysis

- The overall pattern depicts a **cluster of high priced houses on the East side**, with a **cluster of low priced** houses following an **axis through the center**.
- Put in context, this is not surprising, since the **East side** represents houses with **a lake view**, while the **center cluster follows a highway axis** and generally corresponds with a **lower income neighborhood**.
- Interestingly, the pattern is **not uniform**, and several **spatial outliers can be distinguished**.
- Further investigation of these patterns would **require a full hedonic regression analysis.**

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(6) Spatial Regression

- We illustrate the spatial regression capabilities with a partial replication and extension of the **homicide model** used in Baller et al. (2001) and Messner and Anselin (2004).
- These studies assessed the extent to which a classic regression specification, well-known in the **criminology literature**, is robust to the **explicit consideration of spatial effects**.
- The model **relates county homicide rates** to a **number of socio-economic explanatory variables**. In the original study, a full ML analysis of all US continental counties was precluded by the constraints on the eigenvalue-based SpaceStat routines.
- Instead, attention focused on two subsets of the data containing 1412 counties in the US South and 1673 counties in the non-South.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set              :
Spatial Weight        : natrook.GAL
Dependent Variable    :      HR80  Number of Observations: 3085
Mean dependent var    :   6.927616  Number of Variables    :    7
S.D. dependent var    :   6.825088  Degree of Freedom      : 3078
Lag coeff. (Lambda)   :   0.293641

R-squared             :   0.462988  R-squared (BUSE)       : -
Sq. Correlation       : -           Log likelihood         :-9369.500716
Sigma-square          :  25.015000  Akaike info criterion :       18753
S.E of regression     :     5.0015  Schwarz criterion      :18795.241581

-----------------------------------------------------------------------
    Variable    Coefficient     Std.Error     z-value      Probability
-----------------------------------------------------------------------
    CONSTANT     8.463455      0.9765372      8.666803    0.0000000
       SOUTH     1.951838      0.2916178      6.693136    0.0000000
        RD80     3.461736      0.1350625      25.63063    0.0000000
        PS80    0.6745796      0.1185432      5.690582    0.0000000
        UE80  -0.04367847     0.03578631     -1.220535     0.2222621
        DV80     1.151325     0.07579833      15.18932    0.0000000
        MA80   -0.2395876     0.02761771     -8.675144    0.0000000
      LAMBDA    0.2936407     0.02562901      11.45736    0.0000000
-----------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                      DF      VALUE          PROB
Breusch-Pagan test                        6        1187.417      0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : natrook.GAL
TEST                                      DF      VALUE          PROB
Likelihood Ratio Test                     1         120.599      0.0000000
========================= END OF REPORT ==============================
```

Figure 4: Maximum Likelihood estimation of the spatial error model

(6) Spatial Regression

- In Figure 5, we show the result of the ML estimation of a spatial error model of county **homicide rates** for the complete set of **3085 continental US counties in 1980**.
- The explanatory variables are the same as before: a Southern dummy variable, a resource deprivation index, a population structure indicator, unemployment rate, divorce rate and median age.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.

(6) Spatial Regression

- The results confirm a **strong positive and significant spatial autoregressive coefficient** ($\lambda = 0.29$).
- Relative to the OLS results (e.g., Messner and Anselin 2004, Table 7.1., p. 137), the coefficient for unemployment has become insignificant, illustrating the **misleading effect spatial error autocorrelation** may have on inference using OLS estimates.
- The model diagnostics also suggest a continued presence of problems with heteroskedasticity.
- However, GeoDa currently **does not include functionality** to deal with this.

Source: Directly quoted from Anselin, L., Syabri, I., and Kho, Y. (2004). GeoDa: An Introduction to Spatial Data Analysis, Department of Agricultural and Consumer Economics University of Illinois, Urbana-Champaign.