

Spatial Data Analysis with SpaceStat and ArcView

Workbook

(3rd Edition)

Luc Anselin

Department of Agricultural and Consumer Economics
University of Illinois
Urbana, IL 61801

luc@spacestat.com
<http://www.spacestat.com/>

[draft — comments and suggestions are welcome]

July 25, 1999

© 1999, Luc Anselin, All Rights Reserved

May not be reproduced without express written permission

List of Exercises

GIS Operations

- 1 ArcView Basics
- 2 Queries in ArcView
- 3 Spatial Data Types in ArcView
- 4 Spatial Buffering
- 5 Spatial Aggregation
- 6 Address Matching
- 7 Accessibility Analysis in ArcView Network Analyst
- 8 ArcView Spatial Analyst

Exploratory Spatial Data Analysis

- 9 SpaceStat Basics
- 10 Linking SpaceStat and ArcView
- 11 Outlier Analysis
- 12 Smoothing Rates
- 13 DynESDA Extension Basics
- 14 Dynamic ESDA

Spatial Autocorrelation

- 15 Constructing Spatial Weights from ArcView Shape Files
- 16 Spatial Weights Based on Distance Metrics
- 17 Spatial Weights and Spatial Lags
- 18 Join Count Statistics
- 19 Spatial Correlogram for Moran's I and Geary's c
- 20 Visualizing Spatial Dependence, Moran Scatterplot
- 21 LISA Maps

Spatial Econometrics

- 22 SpaceStat Regression Basics
- 23 Heteroskedasticity
- 24 Discrete Spatial Heterogeneity: SANOVA and Regimes
- 25 Continuous Spatial Heterogeneity: Trend Surface and Spatial Expansion
- 26 Diagnostics for Spatial Effects
- 27 Systems Models
- 28 ML Estimation of the Spatial Lag Model
- 29 IV Estimation of the Spatial Lag Model
- 30 ML Estimation of the Spatial Error Model
- 31 GM Estimation of the Spatial Error Model

Exercise 1 — ArcView Basics

Topics

- the ArcView interface.
- drawing simple choropleth maps in ArcView.
- attribute tables

Tutorial

Getting Started in ArcView

- double click on the ArcView shortcut icon; the main ArcView window opens, make it larger for better viewing
note the “Untitled” Project window: you can give a Project a name by using the Properties item in the Project menu; save the project (Save As and give it a name)
- projects are the way in which ArcView keeps track of the various maps and tables that you manipulate; projects contain the file names of shape files and data files with the full path of where they were stored when first used; if you move the files after exiting ArcView, the project will NOT be able to find them any more!
- the ArcView interface consists of a collection of Views (maps), Tables (data tables), Charts (graphs), Layouts (for printing) and Scripts (for programming macros); records (observations) in the View-Table and Chart are linked in the sense that highlighting (selecting) one in either one will also select it in the others.

Create a View

- press New button in Project window with Views highlighted
- select View, Add Theme (or click on the + button)
- browse the file system in the Add Theme dialog box and select the columbus.shp file (from the Columbus directory in the sample data collection); double click on the name or select (highlight) and then click OK
- click on theme to make it visible (check box), map outline appears (Note: in ArcView there is a difference between an active theme — a theme you can refer to and manipulate — and a visible theme, a theme for which a map is drawn; see the help file for further details).

Attributes of a Feature in a Theme

= values associated with variables for a spatial unit in a map

- make sure the View is active
- click on the leftmost button on second row (i), move the pointer over an areal unit (Columbus neighborhood) and click: the Identify Results Box displays all the information (field = variables, and associated values) for the selected location

Attribute Table

- in View, make sure a theme is active (click on it), select Theme, Table: the Attributes of theme table appears, rows=records (observations), columns=fields (variables)
- make table active (click on it), new menu headers appear, check properties, Table Properties (checked fields are the ones that appear; uncheck some and see what happens)

Draw a Choropleth Map

- make sure the View is active and the box for the theme is checked (you see the map)
- double click on the rectangle in the theme legend; the Legend Editor opens
- under Legend Type, select “Graduated Color”, you will now also see the “Classification Field”
- select a Field (= variable) from the Classification Field Drop Down List (such as Crime), click on field name; field name will be listed and an initial classification of the values in the Field will be given
- use the Classify button to change the classification, the default is “Natural Breaks”, but you could also try a quantile map, an equal interval map or a standard deviational map
- click Apply: the choropleth map appears in the View window

Print the Contents of a View

- use Print File; for more complex printouts, use Layout and drop the current theme into the layout by clicking on the top-most layout button (the one with the globe on it) (experiment with adding text, a legend, scale, etc., use ArcView Help for instructions)

Assignment

Make choropleth maps for the variables Crime and Inc for the Columbus neighborhoods. Use at least two different classification schemes (e.g., equal interval vs. quantile) and print the results. Be prepared to comment on the different “impressions” of the data given by the two classification schemes, in particular try to assess and compare the degree of “clustering” for the two variables in the data set. If you wish, experiment with other variables and classification types.

Exercise 2 — Queries in ArcView

Topics

- simple queries and spatial queries
- creating new variables in an attribute table
- constructing an indicator variable for selected records
- creating a new shape file with only selected records

Tutorial

- before you start the tutorial, make sure the *SpaceStat Extension* is loaded (click on the Project window, check File>Extension and make sure the check mark is in the box next to SpaceStat)
- create a new View with the Columbus neighborhood shape file (or load a previously saved project with this theme)

Selecting Observations, Simple Queries

- Make the View active, click on hammer button (Edit>Query), the Query Builder dialog box will open. Select a variable from the list, an operator and a value from the Values list to construct a logical expression using Query Builder, e.g., [Inc] < 9.963; select New Set; all observations that satisfy the condition are highlighted in yellow (in both map and attribute table)
- to remove the existing selection, click on Theme>Clear Selected Features or use the corresponding button (on top row, to the right, looks like a ring binder)

Spatial Queries

- ArcView has a wide range of features to carry out spatial selection of locations directly on the map; these are implemented by means of “selection tools”, the buttons on the bottom bar (fourth, fifth and sixth from left)
- in View, use the spatial selection button (fourth from left on bottom row) to select a subset of the locations (draw a rectangle over them and let go); alternatively, click on selected locations while holding shift key down

- make the Attribute table active and look up the selected records; use the switch button (round arrows up and down) or Edit>Switch Selection and see what happens to the View
- use the circle selection tool; place pointer on the middle of the map (approximately) and move outward (hold left mouse button down), when you let go of mouse button, all neighborhoods within the circle are selected (use Switch Selection in Table to select all neighborhoods outside of the circle)

Finding Neighbors

- use the select button to choose a neighborhood near the CBD of Columbus (it will be highlighted in yellow)
- select Theme>Select by Theme and Select features of active themes that “Are within distance of” from the drop down box; a new choice box opens with the “Selection distance”, leave it at 0; click New Set; the neighborhoods that adjoin (are contiguous to) the selected neighborhood are shown highlighted. For a realistic application, you would convert these steps to an Avenue “script”, for example to construct a spatial weights file for the neighborhoods.

Editing the Attribute Table

- to edit contents of table (individual cells, add/delete fields/records), Table>Start Editing
- to add a Field (a new variable), select Edit>Add Field; the Field Definition dialog box appears, enter name and choose type of data (numeric, character), e.g., enter LowIncome, Boolean, click OK and a new column will be added to the table; use Field, Calculate to assign a new value to each observation; the Field Calculator dialog box will open with [LowIncome] =; enter a boolean operation to construct the values for LowIncome, such as [Inc] < 9.963 (double click on Inc, double click on <, enter 9.963 and click OK); values for LowIncome will be entered as True and False; note that if you have selected records, the new variable will only be computed for those selected records!
- when done with changes, select Table>Stop Editing, you will be asked to Save Edits; if you respond YES, the new Field will be permanently added to the data set.

Selection Variable

- the SpaceStat Extension contains a feature that allows you to easily add a dummy variable with values of 1 for the selected records and 0 for the others
- select all records within a distance from the CBD in View (use the circle tool)
- select Data>Add Selected Features Dummy; enter a variable name for the new dummy variable (the default is Select) and click OK
- check the Attribute table, you will see that a new variable is added, Select (or the name you specified), with values of 1 for the highlighted records; this is an easy way to create indicator variables to carry out spatial analysis of variance (if you already have a Select variable in the attribute table, you will be asked to overwrite it).
- you can easily create a dummy variable for the complement by using the “switch selection” tool button to select the unselected records.

Creating a New Shape File with Selected Records

- select neighborhoods in the Columbus CBD (they will be highlighted on the View)
- select Theme>Convert to Shapefile: the selected records will be stored in a new shapefile (make sure to give it a different name or the original file will be corrupted);
- this feature is very useful to select subsets from a larger data set; for example, to create a shape file with only Michigan counties from the US Counties sample data set that comes with ArcView, use the Query Builder tool to select counties with state=Michigan and then Theme>Convert to Shapefile.

Assignment

Create a new project with the West Virginia Housing shape file. Use the spatial selection tools to create a dummy variable for a subset of the data containing only counties in the interior of the state, i.e., counties that do not border another state. Compare the average housing value in 90 (actually, the average of the median housing values) between the internal and boundary counties of the state (use the Field>Statistics command from the Table menu; make sure the column for the Val90 field has been selected first). Carry out similar comparisons for other socio-economic indicators between the “interior” and “exterior” counties.

Exercise 3 — Spatial Data Types in ArcView

Topics

- exporting and importing tables
- adding point data to a view
- ArcView customization and Avenue scripts
- converting polygons to polylines

Tutorial

- load the SpaceStat extension (with the Project window active, select File>extensions and check the SpaceStat extension)
- create a new View with West Virginia county housing data (Wvhouse.shp)
- create a theme for the state West Virginia: use the states.shp file from the ESRI sample data set, select West Virginia and use Theme>Convert to Shapefile; call the new shapefile Wvstate.shp and save in the same directory as the other WV files (i.e., NOT in the Temp directory given as the default).

Computing and Exporting Centroids

- make sure the Wvhouse theme is active.
- from the menu, select Data>Add Centroid Coordinates; this will compute the centroid for each county and add its coordinates to the attribute table as X_Coord and Y_Coord.
- make the attribute table active, and, from the menu, select Table>Properties; uncheck every variable except Fipsno, X_Coord and Y_Coord (the resulting attribute table will only show three fields)
- export the new table as centroid.dbf in the same directory as the other WV files: use File>Export, select dBase format and enter centroid.dbf as the file name. You can check out this file in any data base or spreadsheet program (e.g., Excel)
- go back to the original attribute table and re-check all variables of interest.

Adding Point Data to a View

- create a new (empty) View
- using Tables, add the centroid.dbf table to the project
- with the View window active, select from the menu View>Add Event Theme; a dialog will open, asking you for the name of the table (centroid.dbf), the X field (X_Coord) and the Y field (Y_Coord); click OK, the centroid.dbf will be added as a theme to the View. Check the theme to make it visible. If you use the identify button to click on one of the points, you will see its “shape” as a point, and the three data items (Fipsno, X_Coord and Y_Coord).
- Note that it’s good practice to convert the point theme to a separate shape file (as it is, there is no shape file for the points, but an explicit link is maintained with the table) using Theme>Convert to shapefile (save it as Wvpoints).
- Any dbf file containing coordinates (e.g., locations of accidents or points collected with a GPS) can be added to a View in this manner.

ArcView Customization

- ArcView allows you to add your own menu items, buttons and tools to the interface; this is accomplished by making the Project menu active and using Project>Customize; a customization dialog opens up that lets you add new features to the interface and assign their properties
- make sure the type is View, and the Category is set to Menus, highlight the last item (&Help); click on the New Menu button, an extra item is added, called Menu. Go down the list of properties and click on the field next to “Label”, then change the label to &MyScripts (the &M means you can invoke the menu by typing the Alt-M key). Close the dialog and check out what happened to the interface when View is active.
- as it stands, the &MyScripts menu does not have any items. Go back to the customize dialog (make the project window active, select Project>Customize), set Type to View, Category to Menu and highlight the &MyScripts menu entry. Select New Item and change its label to Script 1. You now have an extra Menu &MyScripts with one item, Script 1, but so far clicking on this menu item doesn’t accomplish anything.

Editing and Compiling a Script

- In the Project window, select Scripts and click on new; a new empty script window will be opened. Scripts are written in Avenue, an object-oriented language that comes with ArcView. Working with Avenue is beyond the current scope, but you can import scripts from the many system routines that come with ArcView, or download them from the ESRI web page. Unless you really like to program, in many instances someone has already done the work and posted it on the ArcView support pages.
- Start ArcView Help and click on Help Topics; use the index to select “Avenue, sample scripts”; a list of Topics Found will be shown. Scroll down the list till you find “View.ConvertPolygonToPolyline”, double click on this item. In the help display, click on “Source Code” and select Options, Copy. Move back to the open script window and select Edit>Paste from the menu. The source code for the system ConvertPolygonToPolyline script will be copied into your script window. You can scroll through the file to get a sense of how Avenue is structured.
- As such, the script doesn’t work, but it needs to be compiled. This is accomplished with the compile button (a check mark) or by selecting Script>Compile from the menu. Make sure the script window is active when you do this. The script is now compiled and ready to run (the running person button). However, we will add it to our customized interface and invoke it from the Script 1 menu item.
- Make the customize dialog active and select the Script 1 menu item (Type is view, Category is menu, select &MyScripts and the Script 1 item). A list of “properties” (of the object associated with the dialog) is listed on the left, with their definition on the right (in the property field). Double click on the property field next to “Click” and scroll down the list of scripts until you find Script1; select this script. Also, next to the Help item, enter “Polygon to Polyline conversion” (this will show up on the bottom left hand side of your ArcView window anytime you have the cursor on this menu item). Close the customize dialog.

Converting Polygons to Polylines

- You have now all the pieces in place to convert any ArcView polygon to a polyline type (the polyline is the same as the polygon, but without the area and perimeter

characteristics; in general, a polyline is any collection of line segments). Open a new View and add the Wvstate.shp as the theme. Note this is a polygon, i.e., a solid shape.

- With the View and the Wvstate theme active, select the menu item &MyScripts>Script 1; a dialog will ask you for the filename, enter Wvline.shp and make sure to save it in the same directory as the other WV files. A summary dialog reports on the number of shapes converted. Click OK in response to the request to add the shape to a view and add it to the view with the original Wvhouse theme. Note that when you only have the Wvline theme active, there is simply a line, no solid area.

Assignment

Use the police data set to create a file with the fips code and centroid coordinates for the Mississippi counties. Customize the ArcView interface by adding a second menu item to &MyScripts, this one pertaining to the conversion of a polyline to a polygon (View.PolylineToPolygon in the system scripts). Convert the state outline for Mississippi (from the ESRI data set States.shp file) to a polyline and test your new menu by converting it back [For the adventurous: create two buttons to carry out the conversion polygon to polyline and polyline to polygon (also create a tool tip)]. Create a map with the state outline and the centroid for each county. Challenge: turn this into a symbol map (use graduated symbol) to illustrate the spatial pattern in police expenditures.

Exercise 4 — Spatial Buffering

Topics

- editing themes
- spatial buffering

Tutorial

- load the SpaceStat extension (with the Project window active, select File>Extensions and check the SpaceStat extension)
- in this exercise, you will need the Mississippi police expenditure data (police.shp and milines.shp (see the assignment for exercise 3)

Editing Themes

- the purpose of this exercise is to obtain a line theme for the Mississippi river, or the western border of the state, by extracting it from the state line theme
- create a View with the Mississippi state outline (milines.shp) as a polyline.
- make the theme active and select Theme>Start Editing from the menu (the check box for the theme will become dashed)
- first check the overall properties of the polyline theme; make sure the cursor is the pointer tool (black arrow), double click on the view (near the polyline), the polyline should become enclosed in six or so black squares (graphics handles). Right-click the mouse and select shape properties: you will see a list of points (vertices) that make up the Mississippi state boundary. You can now delete points one at a time, but this will take a while, given that there are over 400 points. Press cancel to return to the view.
- select the line split tool (under the draw tool in the tool buttons, go down to the doubly squiggly lines). You can use this tool to create big chunks of the polyline as separate graphic objects, which can then easily be deleted, saving you the hassle of deleting each individual vertex. Click on the view above the state line and slightly to the right of the northernmost point of the river (left-most line), move the cursor down to below the southern border and double click: there will now be several more graphics handles on the view. Switch the cursor to the pointer shape (black arrow) and select a group of graphic handles on the left and delete the corresponding

segment on the polyline; work your way through the set until you only have two short edges sticking out at the north and southern borders (don't delete the last box, or there will be nothing left)

- select the vertex editing tool (transparent arrow) and click on the polyline: all the vertices will show as small squares. Now move the cursor over each square (it will change into a cross-hair) and delete one by one until only the river is left.
- Save the edit as a different file name (otherwise you will lose the original state boundary), say river.shp, in the same directory as police.shp. The new theme will be added; make it active and stop the editing process by selecting Theme>Stop Editing; click on yes to save changes.

Creating a Buffer Theme

- You can now use the river theme to quickly select the counties in the Mississippi delta of the state.
- Make sure you have a view with both the river theme and the police theme.
- Select the polyline in the river theme (use the select rectangle and make sure the whole of the polyline is included; it will turn yellow). Then make the police theme active and use the menu item Theme>Select by Theme, select features of an active theme that “are within distance of” the selected features of “river.shp”, using a selection distance of 0, and make this into a new set. This will select all the perimeter counties.
- Alternatively, if you want to select all the counties that intersect with a 20 mile buffer along the river, you can use the Theme>Create Buffers command. This opens an interactive “wizard” to compose the buffering properties. Start by selecting “The features of a theme” and “river” and press the Next button. Then enter 20 in the box for “At a specified distance” and make sure the distance units are “miles”, click Next again. Finally make sure the barriers dissolved option is on and select save to a new shape file with rivbuffer as the file name (in the same directory as the other police files). The buffer will be added as a new theme to the current view. Now you can use Theme>Select by theme to select all counties that intersect with the buffer (use the option “Intersect”).

Assignment

Use the buffer just created to compare the average police expenditures in the delta region to the average for the rest of the state (use switch selection and Field>Statistics in the Table menu). For the West Virginia theme, use the spatial buffer technique to compare the average housing value in 90 (Val90) for those counties that have a point within 20 miles from the state perimeter to the average for the interior counties.

Exercise 5 — Spatial Aggregation

Topics

- merging features
- spatial aggregation

Tutorial

- load the SpaceStat extension (with the Project window active, select File>Extensions and check the SpaceStat extension)
- in this exercise, you will need the Miami tract and block group files (miatract.shp and miablgrp.shp)

Merging Features

- create a View with the Miami tract data (miatract.shp), make it visible and active
- before you do anything, save the theme (convert theme to shapefile) under a different name, say Miami2; now, create a new view with the Miami2 shape file.
- you want to merge the two two census tracts on the far west and south into a single unit; there are two aspects to this: one is the merging of the boundaries, the other pertains to the field values (data) contained in these record to make sure that the values for the merged units make sense.
- start by using the identify tool and noting the value for Pop100 in each of the tracts (4283 and 10425)
- from the View menu, select Theme>Start Editing and set the cursor to be a pointer (black arrow); click on the first tract, shift click on the second one and observe the graphics handles
- from the Edit menu, select Edit>Union Features and the two tracts will be merged.
- end the edit with Theme>Stop Editing (make sure to save the edits) and use the identify button to see what happened to the records. The new tract is stored under the tract number for the first one selected (114.98) and the value for Pop100 equals the one for the first selected tract. This is typically not what you want. Also, repeat the same exercise, but use Edit>Combine Features to merge the two tracts. Check with

the identify button what happened; again not what you want (Combine Features is only graphical)

- To obtain the proper aggregation rules, you need to set them explicitly in the Theme>Properties dialog; under the Editing category. Click on “Editing” and change the Attribute Updating “union rule” item for the Pop100 field from “copy” to “add”. Now repeat the procedure using Edit>Union Features and you will have the desired result for Pop100. To obtain this for all the fields, you will need to set the “union rule” explicitly in the Theme Properties.

Spatial Aggregation

- create a new View with the Miami tract data (unmerged)
- make the attribute table active and click on the field Mcd (minor civil division); this is an indicator variable that takes on the same value for all the tracts in the same “region”.
- you create a new shapefile with spatially aggregated tracts by means of the Field>Summarize command (or the big Σ button); start this command to get the Summary Table Definition dialog.
- first create a meaningful file name in the “save as” item; then add the Shape field with the “summarize by” set to “Merge” (click on the add button). You will see a new field added to the list and named Merge_Shape. This will contain the spatially aggregated boundaries. Additional fields are added to the new shape file by selecting them one by one and setting the “summary” rules. Typically, these will be Sum or Average, but they must be set explicitly and each field must be “added” explicitly. Try this for Pop100 (will become Sum_Pop100) and for Hucnt100. You can also add the averages for these fields (e.g., Ave_Pop100). Click on OK to create the summary table. When you include the Shape field, you will be asked to add the new theme to a view, otherwise, you will get a new table with the summary items (e.g., only Pop100 or Hucnt100).
- add the theme to a new view, make it active and visible and check the contents for one of the new spatial units by means of the identify button (note that Mcd = 1 only had one tract in it, so the sum and average for Pop100 should be the same, for the others, check that the average equals the sum divided by the Count value). This is an

effective way to aggregate spatial units into larger regions, as long as the boundaries match, for example, counties into planning regions, electoral wards into tracts (if they match), etc.

Clean Shape File

- the SpaceStat extension contains a simple feature to merge the boundary files into larger (or consistent) units. This is very useful when coverages are used that were generated by Arc/Info, in which each polygon has its own unique identifier. In ArcView, a polygon shape can consist of different subpolygons that each share the same identifier. For example, in Arc/Info, North Carolina counties would have different IDs for each of the islands, while in ArcView they should really be grouped under the same ID (e.g., the county FIPS code).
- the Data>Clean Shape File command queries you for the name of an indicator variable that will yield a unique identifier for each of the newly aggregated units. Use the Miami tract data and specify Mcd as the identifier. Enter the name for the new spatially aggregated file and click OK. The result will be shown in a new view. Note that all the variables (fields) are kept, but their values are only for the first one encountered with the same ID number. This only makes sense when all the subunits (to be aggregated) have the same value, which is typically the case for the Arc/Info conversion problem.

Assignment

Use the Miami block group file (Miablkgrp.shp) to aggregate the block group data for Pop100 up to the tract level. Compare the results to the tract theme. You can experiment with other data sets where spatial aggregation may be meaningful. For example, the Nepal.shp data set contains a variable for “development region” (Devreg). Use this to obtain a spatially aggregated theme for some of the demographic data.

Exercise 6 — Address Matching

Topics

- matchable themes
- address matching
- fixing problems

Tutorial

- in this exercise, you will need the Dallas street network file (str48113.shp in the schools directory), census block group file (dallas_county_blkgrp.shp), and the dBase table with school addresses (dallascitypublic.dbf). Note: these data sources can also be downloaded from the web, from <http://www.esri.com/data/online/tiger/> for the GIS files and <http://askted.tea.state.tx.us/cd-rom/start/quickrpt/school/allpubsc/menu.htm> for the schools data.

Matchable Themes

- start a new project, with a new View containing both the Dallas street file and the Dallas block group file; make both themes visible (these are large files, so it may take a few seconds before the view is created).
- before you can carry out geocoding, you need to make sure that the polyline theme (street file) is “matchable”, that is, that it has the proper fields for a given “address style”. Make the street file active and note that the “locate address” button is solid (between the binoculars and the hammer button on the main toolbar). This means that the street file is matchable, in other words, it contains the proper fields to relate an address to a location on the map. Now, make the census block group theme active and notice what happens: the locate address button is no longer solid. The block group theme is not matchable.
- Make the streets theme active. To get an idea of the logic behind address matching, click on the locate button and enter 1515 Young Street as the address, followed by OK. The location of this address will be shown on the map. Zoom in on the point (with the streets theme active) and use the identify button to see the contents at that point. Note that the address is not shown, but instead the field contents are listed for a

polyline theme. In particular the L_add_from, L_add_to, R_add_from, etc. ensure that the point is located (approximately) on the street network.

- use the locate button again and enter “1515 Yang St”; you will get a “cannot locate address” message. The system is not able to associate “Yang St” with one of the address segments in the polyline theme. However, if you make the system less picky, by changing the “preferences” you can still get a match. Enter 1700 Yang St and click on preferences to change “spelling sensitivity” to 50 and “minimum match score” to 40; now the system will be able to locate your misspelled address.

Address Matching

- now that you have established that the street file is matchable, you need to get a database of street addresses to match. In the project menu, select tables, add and click on the file name “dallascitypublic.dbf”. This loads a table with the addresses of the public schools in Dallas County. Note the field “Site_addr_”; this is the field you will use to match. In order to be useable for geocoding, the address data base must have a field that conforms to one of the address styles recognized by ArcView (check the help file for more details)
- make the streets file active and select View>Geocode addresses from the menu; a dialog will open. Keep the reference theme (the street theme), but make sure the “address table” refers to “dallascitypublic.dbf”; select “Site_addr_” as the address field and enter a file name for the geocoded theme (e.g., schools.dbf). Next, click on “batch match” and check the results: 242 addresses, or 90% of the data base were matched, 1 was partially matched and 27 were unsuccessful. This is fairly typical of address matching. If this is OK, click on Done.
- a point theme with the matched addresses will be added to the view; make the theme visible and use the identify button to locate some schools. Note how the matching procedure has added some fields, such as Av_add, Av_status, Av_score that indicate how the matching was carried out.

Fixing Problems

- To improve the results of the match, you may need to “manually” fine tune the parameters and/or correct the entries in your address data base. Go back and re-run the matching program in batch mode, but rather than clicking “Done”, change the

Geocoding preferences by setting the spelling sensitivity to 50, select “rematch unmatched” and click on “interactive rematch”. Click on “next” for the first address, and focus on the second unmatched address, “4223 Briargrove”: a candidate with the correct address range is suggested, but “Briar Creek Ln” is identified as the street name. If you have reason to believe that this is the correct address, you would click on “match.” However, in this instance, Briar Creek is not correct, since Briargrove is outside Dallas County and thus not included in the street file. This is therefore an unmatchable address.

- keep moving through the file (click on next) until you encounter “325 12th Street Stell”. This is an example of a problem with the entry in your data base. Note how a number of candidates with the correct address range but the wrong street are suggested. Select “Edit Standardize” and delete the “Stell” part from the street name, followed by OK. Now, the correct range for 12 St is suggested as a candidate. Click on match. You can finish now, or continue and tackle a few more obvious typos in the data base to improve your batting rate (e.g., 1201 E Eithth St). If you press done now, you will see that the “partial match” has gone up to 2.
- When finished, click on “Done”. The matched points will show up on the street theme and will include the location of 325 12th Street (use query to check).

Assignment

Use the Dallas street file to geocode the locations of the public libraries. The input file is an ascii text file, libraries.txt. Improve the match until you get as close to a perfect match as possible [hint: Renner Frankford is in Collin County].

Exercise 7 — Accessibility Analysis in ArcView Network Analyst

Topics

- shortest path in network
- accessibility measures
- summary statistics for service areas

Tutorial

- before you start the tutorial, make sure the *Network Analyst Extension* is loaded (click on the Project window, check File>Extension and make sure the check mark is in the box next to Network Analyst)
- create a new View with the three Nepal shape files: nepal.shp, npcity.shp and nproad.shp
- make all three themes visible by checking the box on the left and rearranging their order (draw points and lines on top of polygon, polygon should be bottom theme in table of contents)

Shortest Path in a Network

- Make the point theme in the view active (npcity.shp) and get its attribute table; use Table>Properties to set an “alias” for the variable Name to Label (enter the word Label next to Name in the alias column); close the properties dialog by clicking on OK.
- In the View, click on hammer button (Edit>Query), the Query Builder dialog box will open. Select variable Label = Kathmandu, New Set; this will highlight the location of Kathmandu in yellow (you may want to change the size and color of the points in the theme by means of the legend editor)
- Now make the road network the active theme (nproad.shp) and in the Network menu, select Find Best Route ...; a dialog box named Route 1 will open.
- In the dialog box, click on Load Stops and select npcity.shp; Kathmandu will appear as the single selected “stop”, this will be one of the two points between which you will find the shortest path.

- Make the selector tool active (the button with the flag and down-pointing arrow) and click on any other point in the map; it will be listed as “Graphic Pick 1” (you could also have selected it in the point theme, in which case it would have been listed together with Kathmandu); click on the solution button (the grid with a route on it) and the shortest path between the two points will be highlighted (the length of the path will be given under “cost”). Note that in this example, the cost is the arbitrary line length; if your view is in decimal degrees (Nepal is not), you can also specify cost as miles or any other meaningful “cost” (travel time, etc.).
- You can find the shortest path between any two cities in Nepal in the same fashion.

Accessibility

- You can find the “service area” for any city using the Network > Find Service Area function; start by selecting a city (e.g., Kathmandu) in the points theme (npcity.shp), using a query or spatial select tool (first you have to make the points theme active)
- Now make the line theme active and select Network > Find Service Area; in the dialog box, click on the Load Sites button and select npcity.shp. Again, only Kathmandu should be listed. Double-click on the item in the column next to Kathmandu (value should be around 180,000) and enter 150,000. This will create a service area for Kathmandu ranging 150,000 (arbitrary) distance units over the network; click on the solution button to see the area (both a polygon and the corresponding part of the network will be highlighted)
- The theme for the polygon with the service area will be named Sarea1; this can be converted to a regular shape file using the Theme > Convert to Shape File command.

Summary Statistics

- To compute the number of people that were estimated to live in this service area in 1996, you will use a select by theme operation; make the polygon theme active (nepal.shp)
- select Theme>Select by Theme and Select features of active themes that “Have their center in” from the drop down box; “the selected features of” and select “Sarea1” from the drop list; click on New Set to create the selection; the polygons that meet the specified criterion will be highlighted.

- to compute the total number of people in the selected area, make the attribute table active and select Pop96e as the field; use Field > Statistics. Alternatively, you can create a dummy variable for the selected area and use it as the field for a Field>Summary table.

Assignment

Use the Nepal data to find the shortest route between any two cities. Also, compute the population in a target service area (specified with a given distance) around a given city for 1981 and 1991. Experiment with other variables and with changing distance criteria. Try creating a layout with some meaningful maps and print the result.

Note: you can also save a layout (Export) in a number of graphics file formats (e.g., windows metafile) which can be inserted into most word processors.

Exercise 8 —ArcView Spatial Analyst

Topics

- grid interpolation
- contour lines
- Thiessen polygons

Tutorial

- before you start the tutorial, make sure the *Spatial Analyst Extension* and the *SpaceStat Extension* are loaded (click on the Project window, check File>Extension and make sure the check mark is in the box next to Network Analyst as well as next to SpaceStat)
- create a new View with West Virginia housing value theme

Grid Interpolation

- Based on the WV county centroids, create a point theme that contains the Fipsno, X and Y coordinates and the housing values in 80 and 90 (Val80, Val90). Create a point theme and make a symbol map to represent the values in 90.
- Also create a polyline with the outline of the state.
- The housing values represented by the county centroids can also be visualized as a continuous surface, by means of interpolation to a grid (a set of squares covering the area)
- Make sure the point theme in the view is active, and select Surface > Interpolate Grid from the menu (this menu will only be available with the Spatial Analyst extension); choose “Same as View” for the Output Grid Extent (select from the drop down list) and leave all other options to their default settings; keep IDW (inverse distance weighted) as the “Method” and select Val90 as the Z Value Field, click on OK
- A new theme, called “Surface from points” will be created; check it to view its representation; make sure to move the points and the border lines to the top of the table of contents to keep them visible. The surface illustrates the low value “inner-doughnut” in WV as the darker area and the high values in the Eastern Panhandle as

the lighter colored areas. Note the distorting effect of the surface near the state's edges. Show the points and/or the state outline over the surface.

- Repeat the same procedure, but now use the spline interpolation method and the point theme for the output grid extent. Note how the focus of the high values (light area) has moved around relative to inverse distance weighting.

Contour Lines

- An alternative to grid interpolation is to show the contour lines of a three-dimensional surface representing a variable. Again, make sure the points theme is active and select Surface > Create Contours. Use the View for the Output Grid Extent, IDW for the method and Val90 for the Z Value Field and click on OK. Select a contour interval of 5000; a new line theme will be added to the view, named "contours of points.shp". Check it to view the contour lines.
- Double click on the contour legend to give the lines meaningful colors. Use graduated color for the legend type, or, alternatively, use graduated symbols (thicker lines are higher values), and select "Contour" for the classification field. Click apply to change the contour lines in the View.
- Experiment with alternative contour intervals, such as 1000 (much tighter lines, spikes).

Thiessen Polygons

- In many applications, one needs to be able to move between point features and areal features by means of a tessellation. A common tessellation is a Thiessen polygon, which can be constructed with the Analysis > Assign Proximity command
- Make sure the point theme is active and select Analysis > Assign Proximity, with the Output Grid Extent set to Same as View, and the Proximity Field set to Fipsno; a unique value map will be created with a different color for each Fips code; check the box to make it visible and overlay it on the points and county borders (move them to the top of the table of contents).
- Make the "proximity to points.shp" theme active and select Theme > Convert to Shapefile; set the name of the file to Thiessen.shp and make sure to save it in the same directory as wvhouse.shp. Add the theme to the current view and make it visible (you may want to change its fill pattern to transparent, or convert the polygons

to polylines after you create the shape file); compare its shape to the irregular boundaries for the counties.

Assignment

Load the two shape files `afcon.shp` and `aflin.shp` from the Africa directory. Compute the country centroids and create a point theme with `name`, `totcon` and `totcop` as the variables in addition to the coordinates. Create contour lines for total conflict and total cooperation and compare their shapes (experiment with different intervals). Compare the information provided by the contour lines to that of a continuous grid surface. Create Thiessen polygons for the African countries in the data set (note that several countries are not included). Compare the mosaic for the Thiessen polygons to that of the actual country boundaries.

Exercise 9 —SpaceStat Basics

Topics

- SpaceStat menu structure
- moving around the file system
- options
- data menu commands

Tutorial

- start SpaceStat by double clicking on the SpaceStat shortcut item in your Windows desktop; a welcome screen should appear as a DOS window. You can maximize the window by pressing ALT-Return (and also use this command to convert back to the smaller window size). Press any key to move on to the main SpaceStat window.

SpaceStat Menu Structure

- The main menu of SpaceStat contains four modules (**D**ata, **T**ools, **E**xplore, **R**egress), each of which is invoked by typing the corresponding letter. Also, use **ALT-Q** to quit, **F1** to set the options (see below) and **F2** to move to the file system (see below).
- Each module has a two-level menu structure, which is accessed by typing the number for the command or by using the up and down keys to move in the menu, followed by Return to select a command.
- Explore the structure of the Data module. Type D, followed by any of the nine numbers for a submenu. Go back to a higher level by pressing D. For example, to check out the commands under 9 List, type 9 and you will see the nine listing commands.
- To get a feel for the structure of the other modules, type the corresponding letter, followed by the number for the submenu.

Moving Around the File System

- When SpaceStat starts, its “home directory” is whatever was specified in the shortcut. Often this is not what you want and you need to explicitly “move” SpaceStat to the proper directory.

- Type the F2 key and check the directory where you are. If this is not your working directory, use `cd` (change directory) to move around in the directory structure until you are where you want to be [use `cd ..` to move up one level and `cd \pathname` to move to the desired directory]. Use this approach to move to the directory that contains the Columbus data [if you don't know where this is, someone in the lab will tell you].
- NOTE: SpaceStat uses the old 8 character DOS constraint on file and directory names. If your directory name is too long, it will look something like `xxxxxx~1`, where the first six digits are the same as your directory name. To avoid this guessing game, it's easiest to rename any folder or file to an 8 character name.
- Return to the SpaceStat menu by typing **Exit**. Now check the contents of the Columbus data set. First Type D, then 9 (List), then 1 Summary Data Set. Next follows the typical SpaceStat set of interactive queries for filenames, etc. If you don't know the filename, press return and a list of data sets on the current working directory will be shown. If none are shown, you are probably in the wrong directory.
- In response to the file name query, press return to check the contents of your working directory. The file Columbus should be listed (if it isn't, press return until you are back at the menu, then use F2 to move to the proper directory). Type in "Columbus" followed by return: the contents of the Columbus data set should be listed on your screen.
- If you ever get stuck somewhere, simply press return a couple of times and you will be back in one of the menus: NEVER use ESC to try to accomplish this.

Options

- The output just listed only appears on the screen; after you press return, it disappears. In order to keep the results, you need to explicitly activate an **output file**. This is one of the options.
- Start the options by pressing F1. Select 2 Output to a file and enter a filename at the prompt. After you press return you will see the output file listed. Every result that appears on the screen will also be appended (not overwritten) to this output file. Try this out by setting the output file. Now, go back to the Data menu (press ESC, followed by D) and repeat the listing of the columbus data set. After the results are

cleared from your screen, type F2 (to go to the file system), dir (to locate your file) and use the DOS editor (or any other editor) to check the contents of your output file. Go back to SpaceStat by typing exit.

Data Menu Commands

- The Data module contains commands to create SpaceStat data sets and spatial weights files from ascii sets, to carry out various types of data transformations and to list the contents of SpaceStat data sets and weights files. The best way to create data sets and weights files is to use the ArcView extension, so these commands have become less crucial. An important capability of the Data module is to create constants, sequential observation numbers and different kinds of indicator (dummy) variables.
- Create a constant and an observation number for the Columbus data. Use Data > 3 Var Create > 5 Create Constant; enter columbus as the name for the data set (make sure you are in the right directory), followed by const as the variable name and 1 as the value for the constant. The screen will show the contents of the new data set. If you have kept the output file option on, the contents will be appended to it. Now, chose Data > 3 Var Create > 6 Create Observation Numbers and enter columbus and obs in response to the queries. Again, note that a new variable is added to the data set.
- Check the values of the new variables by using Data > 9 List > 3 List Selected Variables. Enter Columbus as the data set, obs and const as the variables and 0 for the default format: the list of values will appear on the screen (and will be appended to the output file).
- Create a dummy variable for all observations with Crime < 34.0 (the median). Use Data > 3 Create Variables > 4 Create Dummy Variables (Range), enter columbus, and use crime for the existing variable and crdum for the new variable; press return in response to the second “variables” query. Enter 0 for the lower bound and 34.0 for the upper bound. The contents of the columbus data set will be shown. Check the value for the dummy variable by means of the Data > 9 List > 3 List Selected Variables command.

Assignment

Use the West Virginia Housing shape file and the techniques described above to create observation sequence numbers and a constant = 2. Use the Var Algebra commands to create a new variable that is the average of the 1980 and 1990 housing values $(Val80 + Val90)/2.0$ (with 2.0 as the constant) [use add and divide commands]. Create a dummy variable for those counties where the vacancy rate in 80 (Vac80) was greater than in 70 (Vac70) [use the subtract command to compute the difference and use dummy (range) to create the dummy].

Exercise 10 — Linking SpaceStat and ArcView

Topics

- SpaceStat extension for ArcView
- creating a SpaceStat data set
- SpaceStat report files
- joining SpaceStat output with ArcView themes

Tutorial

- start SpaceStat **before** you start ArcView (use the SpaceStat shortcut button and press enter to get the main menu); for now, minimize the Spacestat window
- before you start the tutorial, make sure the *SpaceStat Extension* is loaded in ArcView
- create a new View with the Columbus neighborhood shape file (or load a previously saved project with this theme) and make the theme active and visible

SpaceStat Extension for ArcView

- The SpaceStat Extension adds two menu items to the View menu in ArcView (**Data** and **SpaceStat**) and one menu item to the Tables menu (**Data**).
- The Data menu contains various commands to export and import data and to construct some useful spatial variables (with an active View, select Data to see the list).
- The SpaceStat menu contains commands to visualize spatial distributions and indicators of spatial autocorrelation (select SpaceStat to see the list). These are based on computations carried out in SpaceStat
- The SpaceStat-ArcView link is established by means of so-called loose coupling: ArcView creates certain files that SpaceStat can read and vice versa. This is transparent to the user.

Creating a SpaceStat Data Set

- In ArcView, open the Data menu of the SpaceStat extension and select Table to SpaceStat Data Set (or Alt-T). You will be asked to export all variables? For now, enter yes. Two new files will be created on the current working directory with the names columbus.dat and columbus.dht (NOTE: previous files with the same name

will be overwritten without warning). These files are in the binary format that SpaceStat uses.

- NOTE: if your data set contains missing values, the file created by this command may be corrupt, due to SpaceStat's inability to deal with the missing value code. Therefore, it is good practice to always list some selected variables of the created data set as a check.
- minimize ArcView and switch to SpaceStat; if necessary, use the F2 key to move to the directory where the columbus shape file is located (switch back from the system to SpaceStat by typing exit).
- Use Data > 9 List > 1 Summary data set to check the contents of the data set columbus. The SpaceStat data set will contain all the numeric variables from the original shape file, i.e., 49 observations on 21 variables (character variables are skipped).
- Switch back to ArcView, use Data > Table to SpaceStat Data Set, but now click on "no" for all variables. The following dialog will allow you to select the variables you want to move to SpaceStat (by clicking on the variable name); for example, choose polyid, crime, inc and hoval, followed by OK. Now, switch back to SpaceStat and check the contents of the new columbus file (the first one will have been overwritten, unless you renamed it). You will notice only the selected variables are present. (NOTE: there is a bug in the current version of the extension; when you try to export only the selected records, this is ignored and a data set is created with all the observations).

SpaceStat Report Files

- In addition to the screen dumps, SpaceStat also creates output files in a particular format that can easily be joined with Attribute tables in ArcView. These files are referred to as "Report Files"; Report Files are set with the Options command (F1 key) in SpaceStat
- set the following options in SpaceStat (F1 key): (2) Output to a file to yes, with a file name specified (this will contain the output of your analyses); (4) Indicator Variable to the key variable you will use in both SpaceStat and ArcView to identify the observations (this is the variable on which the join will be based — it should be

numerical in ArcView!), enter **POLYID** (default is observation number, but that is not always very useful); (5) **Report Format to 2, Comma Delimited** (return to main menu by pressing Esc)

- NOTE: a common problem occurs when the “Indicator Variable” set in option (4) changes between data sets. If you do not also change this in the options, you will get an error message suggesting some variables were not in the data set. SpaceStat looks for the “old” ID variable, which may not be present in the current data set. Always make sure to check the setting of the Indicator Variable option before carrying out any analyses.

Joining SpaceStat Output to ArcView

- Make sure you have a SpaceStat data set for columbus that contains at least the ID variable Polyid, as well as some other variables, say Crime.
- Make sure the Report Format is set to [2], Comma Delimited
- Use Data > 3 Var Create > 1 Relabel Variables, select columbus and rename Crime to NewCrime (Crime is “existing variable”, NewCrime is “new variable”); make sure to press return to clear the screen at the end
- Select Data > 9 List > 3 List Selected Variables to output Polyid and NewCrime.
- Switch to ArcView and, with the Columbus View active, select Data > Join SpaceStat Report File and select “columbus.txt” as the report file. Check the “attributes of columbus.shp” table to ensure that the new variable has been added to the table. Any newly created variables in SpaceStat can be added to the shapefile in this way (convert theme to shape file to make the addition permanent).

Assignment

Use the West Virginia Housing shape file and the techniques described above to create a SpaceStat data set and to compute the % change in housing value between 80 and 90 (use the var algebra commands). Add the new variables to the existing shape file.

Exercise 11 — Outlier Analysis

Topics

- simple descriptive statistics in SpaceStat
- outlier and percentile maps

Tutorial

- start SpaceStat **before** you start ArcView (use the SpaceStat shortcut button and press enter to get the main menu); for now, minimize the Spacestat window
- before you start the tutorial, make sure the *SpaceStat Extension* is loaded
- create a new View with the St Louis region shape file (stl.shp in the StLouis directory)

Simple Descriptive Statistics in SpaceStat

- create a quantile map for the variable a177995 (homicide rate over the period 1979–95)
- with the view active, select Data > Table to SpaceStat Data Set to create a data set with the variables Fipsno, A177984, A178488, A178893 and A177995.
- switch to SpaceStat. Make sure you are in the right directory (with the St Louis data) and set the options for an output file, for Fipsno as the indicator variable and for Report Format to comma delimited.
- compute descriptive statistics: E(xplore)>1 Describe> 1 Descriptive statistics>2 Interactive (press return for problem file name); enter **stl** as the file name for the data set, skip the spatial weights part (press return) and enter the variable name A177995 (press return to stop)
- you will see screens with the basic descriptive statistics; note the Fips codes for the two outlier counties.
- NOTE: make sure to press the Return key after every screen until you get back to the SpaceStat menu; this ensures that the correct Report File is created, otherwise, you may not be able to create the correct map in ArcView.

Outlier and Percentile Maps

- return to ArcView and make sure the View and the St Louis region theme are active; select SpaceStat>Box Map and double click on B_a17799; a new view will be created with a quartile map in which the upper outliers are highlighted in dark brown; use the identify button to check that these are the same two counties listed in the SpaceStat output (of course, in the output, you had no idea that they were also next to each other).

Note: if the dialog does not show the B_ variables, but instead a list of *.txt files, you may not have completed all screens in SpaceStat; make sure you are back in a SpaceStat menu.

- Now select SpaceStat>Percentile Map and double click on C_a17799; a new view will be created showing the values in the lowest percentile, 1-10%, 10-50%, 50-90%, 90-99% and the upper percentile. Note that only St Louis city qualifies as an “outlier” using the upper percentile criterion. Also note that since there are less than 100 observations in the data set, there is no category for 0-1%, due to rounding.

Assignment

Use the St Louis region shape file and the techniques described above to compare outlier maps (box map and percentile map) for the crime rates in the subperiods 79–84, 84–88 and 88–93 (you will need to use the SpaceStat data set just created and carry out the descriptive statistics). Also compare the impression of “outliers” in the quartile map to that suggested by a standard deviational map in ArcView (use this in the Classify options of the Legend Editor). Try to compose a Layout with the quartile maps for all three periods on the same page.

Exercise 12 — Smoothing Rates

Topics

- computing rates in SpaceStat
- empirical Bayes smoother
- spatial window smoother

Tutorial

- start SpaceStat **before** you start ArcView (use the SpaceStat shortcut button and press enter to get the main menu); for now, minimize the Spacestat window
- before you start the tutorial, make sure the *SpaceStat Extension* is loaded
- create a new View with the St Louis region crime data (stl.shp in the StLouis directory).

Computing Rates in SpaceStat

- use Data > Add Centroid Coordinates to compute the centroids for the counties in the data set
- create a SpaceStat data set with the following St Louis variables: Fipsno, dc7984, dc8488, dc8893, po7984, po8488, po8893, X_Coord and Y_Coord.
- in SpaceStat, make sure the Indicator Variable option is set to Fipsno, Output file is on and Report Format is Comma Delimited
- create the “raw” homicide rates for each of the three periods using Data > 6 Rate Transform > 1 Create Proportions; enter dcxxxx for the counts, poxxxx for the base, and raxxxx for the new variable, select 100000 as the multiplier
- use the List selected variables command to create a file with the new variables (make sure the Fipsno is the first variable); a Report file will be created with the same name as the SpaceStat data set, but with a .txt extension.
- go back to ArcView and use Data > Join SpaceStat Report file to add the computed rates to the attribute table; compare them to the values for the corresponding a17xxxx variables (they should be identical).

Empirical Bayes Smoother

- in Spacestat, select Data > 6 Rate Transform > 7 EB Smoother and enter the homicide (dcxxxx) and population (poxxxx) variables as before, with ebxxxx as the new variable (again, choose 100000 as the multiplier). The new variables will be added to the data set.
- a new report file will be created, with the file name SPTRAN.TXT; switch back to ArcView and join the report file to the St Louis shape file. Compare the attributes of the new Z_dc7984 variable to that of W_dc7984 (use the identify button on any county, or look in the attribute table itself). The Z_ variable is the same as the original raw rate (also a17xxxx), the W_ variable is the Empirical Bayes smoother.
- find the counties with zero homicides and check what happened to the smoothed rate. Also, check the St Louis county and check what happened (note that high population counties have very little change, whereas smaller counties change a lot; also, all the zeros disappear)
- make a box map of the EB smoothed rates to see if the same outliers persist (over time)

Spatial Window Smoother

- before you can implement the spatial window smoother, you need to construct a “spatial weights” file that can be used to select the neighbors in the moving window; this is where you will use the county centroids computed earlier.
- in SpaceStat, select Tools > 4 Distance Weights > 2 Create Arc Distance Matrix, enter stl as the name of the data set and enter dis as the prefix for the distance matrix. In response to the prompts, enter Y_Coord for the latitude and X_Coord for the longitude; a distance matrix will be computed and stored as dis.fmt in the current directory.
- check the characteristics of the distance matrix using Tools > 4 Distance Weights > 3 Characteristics of the Distance Matrix; note the minimum and maximum distances and compare them to the distances in your View (use the distance tool)
- construct two weights matrices based on the distance data; use Tools > 4 Distance Weights > 4 Distance to Binary Weights and enter dis for the distance matrix and wdis as the prefix for the spatial weights matrix. Choose 35 as the upper boundary

and 0 as the lower for the first weights, and 70 and 0 for the second; press Return and the weights files will be constructed, respond No to the row-standardization question. The files will be saved as wdis_1.fmt and wdis_2.fmt in the current directory.

- go to the Data Module and select > 6 Rate Transform > 6 Spatial Rate Smoother, enter stl for the data set wdis_1 for the weights and then construct the rates as before (dcxxxx for the counts, poxxxx for the base, use spxxxx for the rate and 100000 for the multiplier)
- back in ArcView, select SpaceStat > Spatial Smoother from the menu and a new view will open with a quantile map of the spatially smoothed rates. Note how much more pronounced the overall patterns appear. Use the identify tool to compare the original rate (Z_xx) to the spatially smoothed one (W_xx). Alternatively, use a query to find out which smoothed rates are higher (lower) than the original ones. You can also repeat the procedure using the wider window (wdis_2) for even more smoothing.
- NOTE: there may be a problem when several windows are open in ArcView that all refer to the same SPTRAN.TXT file name, when different files with that name have been created in SpaceStat. As long as you leave the windows open, things should be OK. However, there have been problems with this under Windows 95/98 (not NT). One solution is to convert each theme that contains new variables to a different shapefile. Another solution is not to use the SpaceStat > Spatial Smoother item, but to first explicitly list all the new variables in SpaceStat (using the ID variable as the first) with Data > 9 List > 3 List Selected Variables, and then to join the output file (which has the same name as the coverage, with a txt extension) using the Data > Join SpaceStat Report File command. The only difference is that you will then have to construct the maps explicitly using the standard legend editing tools.

Assignment

Use Cressie's North Carolina Sids data (sids.shp in the sids directory) to construct raw rates, EB smoothed rates and spatially smoothed rates for the sids death rate (sid74 over bir74) and compare the overall patterns [Cressie expresses these rates per 1000]. Also compare the patterns over time, using sid79 and bir79.

Exercise 13 — DynESDA Extension Basics

Topics

- basic DynESDA tools
- dynamic linking of windows
- dynamic scatterplots

Tutorial

- Before you start the tutorial, make sure the *DynESDA Extension* is loaded (click on the Project window, check File>Extension and make sure the check mark is in the box next to DynESDA).
- Create a new View with the Amazon Frontier municipalities (frontier.shp in the Frontier directory)

Basic DynESDA Tools

- Make the View active, and create a standard natural breaks map of the % deforestation in 1991 (actually 1992, but the variable is called defpc91). Note the broad regional patterns and the vast difference in scale of the units. Rearrange your desktop such that the ArcView window takes up the right upper quadrant (approximately). You will need the other space for the windows created by the DynESDA extension.
- Invoke the DynESDA extension by clicking on the “S” button; a floating toolbar appears with File, Explore, Window as the menus and five buttons.
- Select the histogram option, either by using the menu of the toolbar (Explore > Histogram), or by clicking on the histogram button (the left most button). Scroll down the list of variables until you find “defpc91”, double click (or click and then click on OK); a new window will open with a standard histogram of the deforestation %. Adjust the histogram to be a bit more informative by changing the number of intervals from the default 6 to 12, by means of the Tools > Intervals menu item. Type in the desired number (12) or move up the list and then click OK (you can also “test” various intervals by means of the “apply” button; these changes are not permanent,

once you click OK, they are). Resize and/or move the histogram window until you can see it along with the choropleth map.

- Now select the boxplot option in the toolbar (select Explore > Boxplot or click on the boxplot button). Again, scroll down and double click on defpc91. A boxplot window will open, with two points as outliers [the box plot has the familiar form, with the dark part corresponding to the quartiles around the median — the blue dot in the middle — the thick lines below and above the box are the “fences”; observations outside the fences are referred to as outliers]. You can change the definition of the fences from 1.5 interquartile range (the default) to 3 times the IQR with the Tools > Hinges menu item (notice how this affects the impression of outliers). Resize and/or move the window so that you can see all three relevant “views” of the data (the map, the histogram and the box plot).
- Finally, select the scatterplot option in the toolbar (select Explore > Scatterplot or click on the scatterplot button). Scroll down the variable list and double click on defpc80 for the x-axis, followed by defpc91 for the y-axis. A scatterplot window will open, indicating a regression slope of 0.8125 between the deforestation rates at the two time periods. Again, resize and/or move the window so that you can see all four “views” of the data.
- **IMPORTANT:** make sure to always close the DynESDA functions (File > Quit or close the floating toolbar) BEFORE you exit ArcView. Otherwise, strange things can happen ... Also, the toolbar is connected with a particular view. If you want to close that view, you need to close the toolbar first (another strange quirk that will be fixed at some point).

Dynamic Linking of Windows

- So far, all you have is four different ways to summarize the distribution of the data. Now, start by highlighting (clicking) on the uppermost category in the histogram (if you selected 12 categories, it should have 3 elements in it. Once you click on the histogram, the corresponding points are also highlighted (in yellow) or “linked” on the three other views. You see that two of the three highest value locations are also outliers in the box plot, and also corresponded to the upper right most points in the scatterplot. These municipios are in the north eastern part of the frontier region (zoom

in and use the identify tool to find their names: Olho D'Agua Das Cunh – truncated –, Lago Do Junco, Altamira do Maranhao).

- You can link any observation in any view of the data to the other views by clicking on it. For example, use the select tool in Arcview to select the easternmost municipios in the frontier region. The matching items highlighted in the other graphs show the “spatial” distribution for this subregion. While it follows the overall pattern for the region, it tends to be more skewed towards the higher deforestation %. This tool can be used to assess the extent to which the distribution of subregions matches the overall patter, or, as a visual approach to assess spatial heterogeneity.
- In the box plot and scatterplot, you can select individual observations by clicking on them, and add unselected observations (or unselect selected observations) with SHIFT-Click. Unselect the municipios in the map in ArcView. Move to the box plot to select the observations with the three highest values (you may need to make the box plot large enough so that you can identify individual observations).
- Alternatively, you can draw a box around a number of observation points to select them. Move to the scatterplot and select a group of points by clicking at the location of one corner of the selection rectangle, drag the mouse and let go at the other corner.
- Finally, you can switch the selection by double clicking in the graph. Assuming you selected some points in the previous step, double click in the scatterplot and observe what happens (this is the equivalent of select unselected in ArcView). Double click again to restore the selection.
- The easiest way to clear all selected observations is to do it in ArcView (click the selection tool anywhere outside the map); you can also select a single point and then deselect it with SHIFT-click, or select all the points and double click to unselect. So far, there is no command to clear the selection. Try the various combinations to make sure you have the mechanics down.
- Now address the following question: given that primary forest is a stock and does not re-grow (strictly speaking, that is), is there something fishy about the data in this example? Carry out a query in ArcView to find those municipios for which $defpc91 < defpc80$ and locate them in the various plots.

Dynamic Scatterplots

- You can have several scatterplots open at a time, and even construct a scatterplot matrix. Say, you would be interested in finding out if there was some relationship between the strange patterns you observed above and the amount of cloud cover. Open an additional scatterplot for the variables cloud80 and cloud91 and observe where the previously selected points are.
- The scatterplots can adjust to the selected data in the sense that the regression slope can be recomputed for the “unselected” points as the selection changes. In one of the scatterplots, use Tools > Exclude Selected. There are now two slopes reported in each scatterplot: one pertaining to the full set (in blue) and one only pertaining to the unselected points (in brown). To get the slope for the selected points, switch the selection with a double click. Adjust the selection of points and observe how the slope changes. You can use this to assess the “leverage” of particular observations, or to identify visually observations that have high influence on the regression slope.

Assignment

Make sure the SpaceStat Extension is active as well. Join the rates you computed in Assignment 12 for Cressie’s North Carolina Sids data with the matching ArcView sids shape file and convert the theme to a new shape file (e.g., newsids). Use the dynamic box plots and/or histogram tools to assess Cressie’s suggestion (Cressie 1993, p. 395 and p. 400) that Anson county is an “outlier” and should be removed from the analysis. Also assess the effect of smoothing on the identification of outliers. Assess the relationship between the sids rates for the two time periods by means of a scatterplot and how the “outlier” affects the regression slope.

Exercise 14 — Dynamic ESDA

Topics

- brushing box plots and scatterplots
- exploring spatial heterogeneity

Tutorial

- Before you start the tutorial, make sure the *DynESDA Extension* is loaded in ArcView (click on the Project window, check File>Extension and make sure the check mark is in the box next to DynESDA).
- Create a new View with the St Louis region crime data (stl.shp in the StLouis directory))

Brushing Box Plots and Scatterplots

- In separate views, create a simple choropleth map for the homicide rate in 84–88, that is, variable “a178488” (a relatively stable period), and another one for 88-93, variable “a178893” (a period of increasing homicide). Make the theme in one of the views active.
- Using the DynESDA floating toolbar (press the S button), create a box plot for a178488 and one for a178893. Also create two scatterplots, one with rdac85 (“resource deprivation”, a composite measure that is a classic covariate of homicide) on the x-axis and a178488 on the y-axis and one with rdac91 on the x-axis and a178893 on the y-axis. Arrange and resize as necessary so that all four graphs are visible, as well as the maps.
- A “brush” is a small rectangle that can be moved over the data in a graph to investigate how changing subsets of observations have similar or different associations in a multivariate setting. Here, we will look at the regression slope in the scatterplots over time and how it changes with changing subsets of the data.
- In the boxplot window for 88–93, create a small rectangle around the uppermost observation (click on mouse in upper left corner, drag to lower right) and then press CTRL. The rectangle will flash, and then becomes a “brush”. Move the mouse down the box plot (i.e., drag the rectangle down) and observations will be highlighted in all

graphs that correspond to lower and lower homicide rates in 88–93. This provides a very intuitive and visual way to assess whether the same locations were “high” or “low” homicide locations in 84–88 compared to 88–93. If you also have the “Exclude Selected” option set in the scatterplots, you will see how the slope of the regression line changes as particular observations (or subsets of observations) are excluded from the analysis.

- You can also brush the scatterplots directly. Move to one of the two scatterplots and make a long and narrow rectangle around the highest observation on the x-axis (or, alternatively, on the y-axis). Click in the left-most corner and drag to form the rectangle, followed by pressing the CTRL key (the rectangle will flash for a brief moment). Now move the “brush” from high values of the covariate (or, alternatively, high values of the dependent variable) to low values, and observe how the slope of the regression line changes. See if you can identify two particularly “high influence” observations (and locate them on the map).

Exploring Spatial Heterogeneity

- The use of dynamically linked windows is particularly useful in exploring spatial heterogeneity (for its use in exploring spatial dependence, see exercise 20).
- In addition to the box plots and scatterplots you already have, create two histograms, one for each homicide variable.
- In the ArcView window, use the “circle” select button to select counties within 30 miles from St Louis city (if you are not familiar with Midwest geography, you may have to use a query first to locate St Louis city); check the distance in the “radius” variable on the lower left of the window (use “approximately” 30 miles). The selected counties will be highlighted in all the graphs. Assess the extent to which the homicide variable is different between the “core” subset and the periphery. Also assess the extent to which the slope of the association between rdac and the homicide variable changes across space. Focus in particular on the role of the two “high leverage” observations: how does the slope change when they are excluded? What would be a substantive interpretation?
- When an exploratory analysis suggests a hypothesis of spatial heterogeneity, a more formal test by means of Spatial Analysis of Variance may be in order.

Assignment

With the West Virginia housing data set (wvhouse.shp), assess the extent to which housing values in 80 and 90 (val80, val90) are different in the periphery from the “core”. Also assess the extent to which the regression between rental rates (rent80, rent90) and value in the same year is stable across space and over time.

Exercise 15 — Constructing Spatial Weights from ArcView Shape Files

Topics

- creating rook and queen contiguity weights based on ArcView shape files
- characteristics of spatial weights

Tutorial

- first start SpaceStat and then ArcView
- before you start the tutorial, make sure the *SpaceStat Extension* is loaded
- create a new View with the West Virginia housing data shape file

Create SpaceStat Data Set with ID Variable

- If you haven't already done so as part of an earlier exercise, create a SpaceStat data set for the West Virginia housing data that includes the housing value variables as well as the variable FIPSNO as the indicator variable.

Create a Sparse Contiguity Weights File

- In ArcView, make the View with the West Virginia map active and select Data>Rook Weights from Shape File; enter Fipsno for the ID variable.
- A contiguity matrix is created and saved in GAL format (wvhouse.gal); the contiguities are defined following the “rook” convention, i.e., only common boundaries are considered.
- Check the format of the gal file using any text editor (say notepad). The first couple of lines are header lines that contain the name of the data set with the ID variable as well as the ID variable name. Then follows the actual neighborhood information: the ID of the county, the number of neighbors and the Ids of the neighbors. You will see how the fips codes are used to identify the counties and their neighbors. Pick a county from the file and record its neighbor structure. In ArcView, select that same county and check its neighbors by means of the Theme>Select by theme>are within distance of command (keep the distance to the default of 0).
- To create a weights file following the “queen” convention (both common boundaries and common nodes), select Data>Queen Weights from Shape File and follow the same procedure.

Characteristics of Spatial Weights

- Check the connectedness characteristics of the spatial weights using Tools> 1 Weight characteristics> 1 Connectivity. Make sure you are in the correct directory. Also, you need to make sure that there is a SpaceStat data set with the same name as the gal file. If this is not the case (e.g., if you renamed the SpaceStat data set) then you need to edit the filename in the .gal file or the command will not work.
- Check the frequency distribution of the number of neighbors. Note that the counties are referred to by observation number, not by Fipsno. After you have checked the connectedness structure, the original .gal file (with the headers and fips codes) is saved with a .g00 extension and the new .gal file only contains observation numbers. Use a text editor to compare the structure of the two files.
- Set the Indicator variable option to Fipsno and repeat checking the connectedness structure. You will be asked to specify the data set that contains the indicator variable: enter Wvhouse. This time, the results are expressed in Fips codes. Identify the most and least connected counties and find them on the map (use a query with the fips code).

Assignment

Create spatial weights files for the North Carolina sides data sets. Make sure to create data sets with an indicator variable first. Check the characteristics of the contiguity structure using Tools>1 Weight characteristics>1 Connectivity in SpaceStat. Compare the overall connectedness structure between the West Virginia and North Carolina maps.

Note: the results of the weights characteristics are listed with polygons identified by their observation numbers. To have more useful identifiers, such as Fipsno for WV and NC, make sure to set the ID Variable option to the relevant variable with the Options (F1) key. The connectivity structure will then be given for polygons identified by the ID Variable.

Exercise 16 —Spatial Weights Based on Distance Metrics

Topics

- creating a distance matrix
- characteristics of a distance matrix
- creating contiguity weights based on distance criteria
- types of distance-based weights

Tutorial

- Make sure to start SpaceStat before ArcView.
- Use the Columbus data sets (columbus.dat and columbus.dht). If you don't have them, recreate them from ArcView and make sure they contain polyid, the X and Y from the original data set (or the centroids X_coord and Y_coord constructed with the SpaceStat extension).

Creating a Distance Matrix

- Create a distance matrix based on the Euclidean distance between the centroids of the Columbus neighborhoods.
- In SpaceStat, select Tools> 4 Distance weights> 1 Create distance matrix: enter the data set name (columbus), the name for the distance matrix file (e.g., coldis), and the variable name for X and Y (respectively X, Y, or X_coord, Y_coord).
- Before the distance matrix is calculated, you get a warning: smallest distance is < 1. While this is a simple scaling problem, in many instances having distances smaller than 1 may be problematic since it inflates the distance decay effect for small distances ($1/\text{distance}$ becomes > 1). Answer yes to the prompt. The distance matrix is saved in the FMT (Gauss full matrix) format as coldis.fmt in the current directory. Note: this is a matrix with the distances between each pair of observations, NOT a contiguity file.

Characteristics of a Distance Matrix

- Check on the characteristics of the distance matrix: Tools> 4 Distance weights> 3 Characteristics of distance matrix and specify coldis as the name of the distance file.

Some descriptive statistics are provided, such as the minimum and maximum distance. Note that the minimum is 1 (since you rescaled the distances).

- Pay particular attention to the minimum allowable distance cut-off (4.55 in the original data set). This is the smallest distance needed to ensure that all observations have at least one neighbor; choosing a distance cut-off less than this value will result in isolated (unconnected) observations or islands. This is often a problem when the units under consideration have very different areas (e.g., western US counties compared to eastern counties).

Creating Contiguity Weights based on Distance Criteria

- To build a binary contiguity weights file from the distance matrix use Tools> 4 Distance weights> 4 Distance to binary weights; enter the name of the distance file, a prefix for the weights files, such as wd (several weights files may be constructed from the same distance matrix)
- In response to the query about upper and lower bounds, enter 5, 0 and 4, 0, followed by return to end the sequence (note that 4 is less than the minimum cut off). Press **return** (NOT yes) in response to the question about row-standardization.
- The resulting contiguity files are saved to the current directory as files with a FMT extension (such as wd_1.fmt, wd_2.fmt).
- Check the characteristics of the weights using Tools > 1 Weight Characteristics > 1 Connectivity. Note whether the row-standardization is correctly indicated (if not, you probably did not press return to the prompt in question) and how wd_2 has several unconnected observations (islands).
- NOTE: make sure to turn off the indicator variable if it is still set to Fipsno from a previous exercise. Since there is no Fipsno variable in the Columbus data set, not adjusting the indicator variable will generate an error message.

Types of Distance-Based Weights

- In addition to straight Euclidean distances, you can also compute a distance matrix based on great circle distance (as in Exercise 12). With either of these distance matrices, a variety of spatial weights can be constructed. So far, only binary contiguity weights were used.

- To build a weights matrix with inverse distance weights (such as $1/\text{distance}$), use Tools > 4 Distance Weights > 5 Inverse Distance weights, enter the distance matrix name, the prefix for the distance weights, and an optional cut off distance (press return to make a full matrix). If you want to use an integer power for the distances (say 2, to reflect gravity weights) enter the number at the following prompt; simply pressing return enforces the default of inverse distance weights. Select No for row-standardization. The weights will be saved in the same directory.
- Check the characteristics of the weights just created. Note the lack of row-standardization and also that the least and most connected information is not informative since you built a full matrix.
- Sometimes it is useful to base spatial weights on the k nearest neighbors (say in real estate analysis). Build such a weights matrix for the Columbus data using Tools > 4 Distance Weights > 6 k Nearest Neighbors. Again, enter the name of the distance file and a prefix (e.g., nn) for the nearest neighbor weights. Specify 5 for the number of nearest neighbors and return for row-standardization. A weights file consisting of the prefix, followed by an underline and the number of neighbors will be saved as a fmt file (e.g., nn_5.fmt). Check the characteristics of these weights.

Assignment

Construct a set of distance-based spatial weights for the West Virginia coverage. Use at least two distance bands and compare the resulting weights in terms of their connectedness characteristics. Use the Theme>Select by theme>are within distance of command for the most connected county to check the neighbors (use the distance you specified as the upper bound in SpaceStat). Experiment with nearest neighbor distances and check for the most and least connected counties how different the neighborhood structure is from that for distance bands (set the indicator variable to fipsno). Also, use the Thiessen.shp file created in Exercise 8 to construct a rook spatial weights matrix and compare its properties to that of the distance based weights.

Exercise 17 — Spatial Weights and Spatial Lags

Topics

- formats for spatial weights in SpaceStat
- row-standardization of spatial weights
- higher order contiguity weights
- constructing spatial lags
- spatial lag bar charts and spatial lag pie charts

Tutorial

- start SpaceStat **before** you start ArcView
- before you start the tutorial, make sure the *SpaceStat Extension* is loaded
- create a new View with the Mississippi police expenditure data (police.shp in the police directory)
- create a SpaceStat data set police.dat for the Mississippi counties with at least the variables Fipsno, Police, Crime, Tax and White.

Formats for Spatial Weights in SpaceStat

- In ArcView, make the View with the Mississippi counties active and use Data>Rook weights from shape file to create a gal-format contiguity file with Fipsno as the ID variable. A file named police.gal will be created.
- The file police.gal contains only information on the neighborhood structure of the polygons in the View. To add quantitative information on the potential strength of interaction between two observations, so-called “general” weights can be used. In SpaceStat, general weights are stored in a “gwt” sparse format. Convert the police.gal file to a general format by means of Tools > 3 Weight Conversion > 6 FMT Weights to sparse general and call the new file wgmiss [NOTE: even though the menu specifies “FMT weights” it does work for GAL weights as well]. A file wgmiss.gwt will be created in the current directory; check its format with a text editor and its characteristics using Tools > 1 Weight Characteristics > 1 Connectivity.
- Sparse weights formats are used for all analyses in SpaceStat except maximum likelihood estimation of spatial regression models. For the latter, a “full” spatial

weights matrix is necessary. You convert the sparse formats to full format by means of a Tools > 3 Weight Conversion command. For gal files, the proper command is Tools > 3 Weight Conversion > 1 GAL file to matrix format (FMT), call the new file wms and press return in response to the prompt for row-standardization; and also for the symmetry check (this is good practice, but not necessary in this case). A file named wms.fmt will be created in the current directory. Check its characteristics in the usual fashion. Repeat the procedure, but now respond No for row-standardization; check the characteristics.

Row-Standardization of Spatial Weights

- For most analyses, you want the spatial weights to be in row-standardized form, i.e., such that the row elements for each observation sum to 1. Sparse gal files are row-standardized by default, but gwt files and fmt files are not. To row-standardize these, use Tools> 2 Weight transforms> 1 Row Standardization. Use this command to convert the full non-standardized matrix you constructed from the police.gal file to row-standardized form.

Higher Order Contiguity Weights

- Form second to fifth order contiguity weights for the Mississippi counties using the Tools> 2 Weight transform> 2 Higher Order Contiguity command. Enter wmiss (for the gal file) as the base file and wr as the prefix for the higher order files. Type 5 for the contiguity order. Four new files will be created on the current directory, wr_2 through wr_5 (with file extension either .gal or .fmt depending on what you used for the base file). Do NOT use the row-standardized wms file as the base file. Assess how the connectedness structure changes with the higher orders of contiguity [make sure to set Fipsno as the indicator variable].

Constructing Spatial Lags

- Use the first order contiguity file for the Mississippi counties to construct spatially lagged variables for Police, Crime, Tax and White. Make sure to set the Indicator Variable option to Fipsno and the Report Format to 2, Comma Delimited (use the Options - F1).
- Use Data > 5 Space Transform > 1 Spatial Lag, enter the data set name (police), spatial weights file (police), variable name (e.g., Police, Crime, Inc) and the name for

the spatial lag (e.g., w_police). Use option “0” for “raw” spatial lags (i.e., not standardized in any way). The new variables will be added to the data set. Also, the Report format option will generate an output file Sptran.txt in the current directory that contains the spatial lags you computed. You can check the contents of the file with a regular text editor.

Spatial Lag Bar Charts and Spatial Lag Pie Charts

- In ArcView, make the View of the Mississippi counties active and create a standard deviational map of the crime rates (Crime). Next, select SpaceStat> Spatial Lag Bar Chart from the menu. Select W_Police as the spatial lag variable. A new view will be created with a spatial bar chart. You may have to adjust the height and width of the bars for your view: use the Legend Editor, Properties and adjust the column chart properties if needed. Compare the relative magnitudes of per capita police expenditures in each county to those in the neighboring counties. Locate counties where own expenditure (blue) is much larger than for the neighbors (red), and vice versa, these are referred to as spatial outliers.
- Use SpaceStat> Spatial Lag Pie Chart to get a similar graph using pie charts to illustrate the relative magnitude of a variable for a county to that of its neighbors.

Assignment

Use the Africa shape files (afcon.shp and aflate.shp) and the techniques described above to describe the spatial distribution of military expenditures (mcge) and visually assess the presence of spatial outliers and/or spatial clusters. Do the same for any one of the conflict indicators (totcon, totcop, verhos, and versup). You will need to create a SpaceStat data set, a contiguity weights file (use the afcon theme with ID as the indicator variable) and spatial lags. Use the aflate theme only as a backdrop to see the full outline of the continent, do not use it to create spatial weights or SpaceStat data sets. Construct higher order contiguity weights for Africa and assess at what stage you start to get problems with “islands”.

Exercise 18 — Join Count Statistics

Topics

- unique value maps
- computation and inference for the join count statistic
- inference based on a permutation approach
- spatial association for categorical variables

Tutorial.

This exercise assumes that you already have a SpaceStat data set for the African conflict data with a matching spatial weights file. If not, go ahead and create these and make sure the data set contains at a minimum the Id variable and the variables Ind60, Islam and Govtyp.

Unique Value Maps

- In ArcView, create a view with the themes for the African conflict data (afcon.shp) and the African boundaries (aflin.shp).
- Make the afcon.shp theme active (click on it) and use the Legend Editor to create a map for the countries which were independent before 1960 (Ind60) using the “unique value” feature.
- In a second view for the African data, create a unique value map for predominantly Islamic countries (Islam).
- Compare the two patterns in terms of their spatial clustering. Would they be “clustered”? Would one be more clustered than the other? Which one?

Computation and Inference for the Join Count Statistics

- In SpaceStat, set the Output file option if you wish to save the results to a file.
- Select Explore> 2 Join Count> 1 Binary normal and choose the “2 > Interactive” approach (press return for the problem file). Next, specify the data set as africa (or afcon, if you created it directly from ArcView), the spatial weights file as africa (or afcon) and Ind60 and Islam as the variable names.
- Two join count results and their associated significance levels are given: the BB counts (coincidence of 1-1) and the BW counts (coincidence of 1-0). Note how Ind60

is only marginally significant ($p = 0.026$) while Islam is very strongly significant ($p = 0.0002$).

Join Count Statistics, Permutation Approach

- In SpaceStat, make sure the number of permutations is set to a higher value than the default of 99 (Use F1 and set the permutation option 7 to 999).
- Select Explore> 2 Join Count> 2 Binary Permute, and proceed as before (same data set, weights and variables as before — Ind60, Islam).
- Again, two join count results and their associated significance levels are given: the BB counts (coincidence of 1-1) and the BW counts (coincidence of 1-0). Note the different significance levels compared to the normal approach (the current ones are pseudo-significance levels), especially for Ind60, where the significance drops to $p = 0.04$
- Note that in order for these exact results to follow, the random seed in SpaceStat must be kept to the default. If the random seed is instead set to the system clock, the results for each simulation will be slightly different. You should normally not have to change the default unless you have a good reason to.

Spatial Association for Categorical Variables

- In ArcView, make a unique value map for the variable Govtyp, a classification of the type of government (see the data documentation for the precise definition and interpretation). Comment on the degree of patchiness of the data: does the map suggest a greater diversity than under randomness?
- To quantify the degree of spatial diversity, compute the Jtot statistic in SpaceStat. Use Explore> 2 Join Count > 3 Multinomial Normal and enter the data set, weights and Govtyp in response to the queries. The Jtot index of spatial diversity takes a z-value of -1.08, which is not significant to reject the null hypothesis of spatial randomness. This is confirmed by a permutation approach as well (Explore > 2 Join Count > 4 Multinomial Permute).

Assignment

Use the StLouis regional crime data set (stl.shp) to create a SpaceStat data set that contains dummy variables for those counties that had less or equal to 5 homicides in each of the periods (use de7984, de8488, de8893 and a query in ArcView to build the selection; create the dummies with Data > Add Selected Features Dummy). “Visually assess the degree of clustering in these counties. Now assess the clustering by means of join count statistics.

Exercise 19 — Spatial Correlogram for Moran's I and Geary's c

Topics

- visual impressions of spatial clustering
- computation and inference for Moran's I
- computation and inference for Geary's c
- spatial correlogram

Tutorial

This exercise assumes that you already have a SpaceStat data set and matching spatial weights file for the Nepal demographic data. If not, go ahead and create these, with the data set containing at a minimum the ID variable Obs and the variables Pgr7181, Pgr8191, Totlitr, Mallitr, Femlitr, Cpr91 and Cpr94.

Visual Impressions of Spatial Clustering

- In ArcView, open two views with the Nepal coverage (nepal.shp). In the first one, create a map for the population growth rate in the 1970s (Pgr7181), in the second, do the same for the growth rate in the 1980s (Pgr8191).
- Visually assess the extent of spatial clustering in the two maps. Do they appear to be spatially random or positively or negatively spatially correlated?

Moran's I

- In SpaceStat, make sure Output file option is set such that the results will be saved to a file.
- Use Explore > 1 Describe > 1 Descriptive Stats and create a problem file: enter moran.txt for the file name, followed by return. Next, enter nepal for the data set, nepal for the weights file and the two variables Pgr7181 and Pgr8191. Assess whether the null hypothesis of normality can be maintained for Pgr7181 (clearly rejected) and Pgr8191 (rejected at $p = 0.04$).
- Use your text editor to examine the contents of moran.txt: you will recognize the file names and variable names; don't change anything unless you are fully familiar with the structure of the problem files!

- Using the randomization option for the inference for Moran's I, select Explore> 2 Moran> 2 Randomization, followed by 1 Batch and enter the problem file name "moran.txt".
- After a summary of the weights (this is important to detect islands and other potential problem), the results for Moran's I, associated z-values and significance levels will be displayed. Notice that the value of Moran's I for Pgr7181 is negative, but the z-value indicates that the null hypothesis of spatial randomness cannot be rejected. By contrast, a clear pattern of significant positive spatial autocorrelation emerges for Pgr8191.

Geary's c

- Use the same procedure as before to assess spatial autocorrelation by means of Geary's c statistics. Use Explore> 4 Geary> 2 Randomization > 1 Batch and enter moran.txt for the problem file name. Again, the null hypothesis of spatial randomness cannot be rejected for Pgr7181, but Pgr8191 is significantly positively spatially autocorrelation (note that for positive autocorrelation Geary's c is *less than 1*, with a z-value of -3.09).

Spatial Correlogram

- Use Tools> 2 Weights transform> 2 Higher Order to construct spatial weights for the Nepal districts up to the fifth order.
- Reset the output file option to a new file that will only contain the results of this analysis, say high.txt.
- Select Explore> 3 Moran> 5 Correlogram Randomization to compute Moran's I for contiguity order 1 to 5 for the variable Pgr8191. You will have to use 2 Interactive, enter the data set, the five weight files (in order of contiguity) and the variable name. After the summary of the five weights, the results are listed.
- The results for the z-values in the output file can be used to construct a spatial correlogram as a graph in ArcView or in many other software packages. Open the high.txt file in any text editor. Add two header variables: Order and Z, separated by a comma (important!) on the first line. Edit the remainder of the file to obtain two columns with in each row the order and the corresponding z-values, with the values separated by a comma (you have to do this manually in the text editor).

- In ArcView, add the table to a project by clicking on Tables, Add and select delimited text as the type. After the table appears, select the “chart” button and the spatial correlogram will be visualized as a bar chart. Customize the bar chart to add horizontal lines for the z-values to highlight the significant results. You can also add this graph to a layout.

Assignment

Assess the degree of spatial autocorrelation using either Moran’s I or Geary’s c, using where appropriate either normality or randomization for the literacy and contraceptive prevalence variables. Comment on the “range” of autocorrelation. Try also with some alternative definitions for the spatial weights (distance bands, k nearest neighbors) and assess the sensitivity of the conclusions to the choice of the weights.

Exercise 20 — Visualizing Spatial Dependence, Moran Scatterplot

Topics

- decomposing global spatial association by means of a Moran scatterplot
- visualizing the four types of spatial association by means of a Moran scatterplot map
- identifying outliers and influential observations
- linking and brushing the Moran scatterplot

Tutorial

- Make sure to start SpaceStat before ArcView and have both the SpaceStat Extension and the DynESDA Extension active in ArcView.
- This exercise assumes that you already have a SpaceStat data set for the Mississippi police expenditure data. If not, go ahead and create one (police.dat) containing at a minimum the ID variable Fipsno and the variables Police, Crime, White and Tax.
- You will also need a spatial weights file for the Mississippi counties (such as police.gal).
- Make the Mississippi county theme active in ArcView (police.shp) and resize the ArcView window such that it only takes up the right hand side of your screen (you will be using the Dynamic ESDA features).

Moran Scatterplot

- In SpaceStat, use Options (F1) to set the ID variable option to Fipsno and the Report File option to comma delimited (2). Also set the Output to a File option to obtain a text file with the results of the computations.
- Compute the data for a Moran scatterplot by means of Explore> 3 Moran> 7 Moran scatterplot > 2 Interactive. Enter the data file name for Mississippi (police), the spatial weights (police) and the variable names Police, Crime, White and Tax.
- The Moran scatterplot results include Moran's I, the classification of spatial dependence into the four quadrants and the 10 extreme observations according to max normalized residuals, hat value and Cook's distance. These results will be written to a report file with name MS_police. This contains the id variable Fipsno, and, for each of the four variables, a standardized value (Z_varname), spatial lag (W_varname),

also standardized, quadrant in the scatterplot (Q_varname), normed residual (R_varname), hat value (H_varname) and Cook's distance (C_varname)

Moran Scatterplot Map

- In ArcView, create a view with a map of the Mississippi police expenditures (choose any classification type) and make sure the view is active.
- Select SpaceStat>Moran Scatterplot Map. In the dialog boxes that follow, select ms_polic.txt as the file and Q_police as the variable.
- A unique value map will be created with 1 for high-high observations, 2 for low-low, 3 for high-low and 4 for low-high. Notice how the scatterplot map smooths the original map and suggests broad “regions” of low and high values as well as potential spatial outliers (often located at the boundaries of the “clusters”).

Identifying Outliers and Influential Observations

- With the Moran scatterplot map as the active theme, start the DynESDA toolbar and click on the Moran scatterplot icon (the scatterplot with an M on it). Select Z_police as the variable. A plot will appear that decomposes the spatial autocorrelation into four categories and shows a regression line with a slope equal to the (global) Moran's I. For police expenditures, this line is very flat, indicating a low degree of spatial autocorrelation (Moran's I is 0.103).
- Notice the one value with extremely high expenditures pulling the line down (use the i-tool on the county to find out the leverage statistics). Also find the points in the scatterplot that correspond to the two counties in the south east corner of the state.
- In the Moran Scatterplot, use Tools > Exclude Selected to recompute the Moran's I without the value for Hinds county. Note how the statistic doubles in size.

Linking and Brushing the Moran Scatterplot

- The Moran scatterplot feature in the DynESDA extension can also be invoked without using output from SpaceStat.
- Close the current DynESDA toolbar, make the original map of Mississippi counties the active theme and start DynESDA up again (note that you need to close DynESDA every time you use a different theme). Select the Moran scatterplot button and take Police as the variable. A new Moran scatterplot opens, but the Moran statistic is 0.087 rather than 0.103. This is due to the choice of the “queen” criterion for the weights.

Go back to the toolbar and select Window > Properties > Criterion of Contiguity and change the check mark to “rook”. Now create a new Moran scatterplot for Police and you will see a figure identical to that generated by SpaceStat (note the /Q and /R after the scatterplot to designate which type of contiguity was used).

- Inference for Moran’s I is based on the permutation approach. In the Moran Scatterplot window, select Tools > Randomization > 499 Permutations. A graph is constructed that represents the density of the reference distribution of the simulated values. Note that there is only weak indication of spatial autocorrelation (pseudo significance of 0.028 for queen under the default). The yellow bar represents the observed Moran’s I, the other values are for simulated data sets. Close the randomization window in order to proceed.
- Use the DynESDA toolbar to open an additional Moran scatterplot for white. Notice the much stronger evidence for positive spatial association (Moran’s I of 0.56) and use the link tool (a box around the points in a respective quadrant) to locate the high-white (or, alternatively, the low-white) counties.
- Use the brush tool to assess the extent to which the spatial association in race matches the spatial association in police expenditures [in other words, are the high-high, low-low common to both variables]. In addition, you may want to add additional views to assess the relationship between police expenditures, race and crime.

Assignment

Use the African conflict data set and the DynESDA extension to decompose the pattern of spatial association for total conflict (totcon) and total cooperation (totcop) [use the afcon theme for the analysis and afile as a background]. What are some important clusters and outliers. How would you interpret the outliers? How does Moran’s I change after the high leverage points (Egypt and Sudan) are dropped?

Exercise 21 — LISA Maps

Topics

- computation and inference for Local Moran LISA statistic
- mapping significant LISA statistics in LISA maps
- combining a LISA map with a Moran scatterplot
- computation and inference for Getis-Ord G_i and G_i^* statistics
- mapping significant G_i and G_i^* statistics

Tutorial

- This exercise will use the Nepal data and assumes that you already have the necessary SpaceStat data set and weights files.

Local Moran

- In SpaceStat, set the ID option to the variable Obs, the Report file format to comma delimited, and the Output file option to save the results; change the number of permutations to 999.
- Compute the Moran's I statistics by selecting Explore> 3 Moran> 9 Local Moran permutation, specify nepal as the data set, the matching spatial weights file and Cpr91 as the variable name.
- Three Local Moran summary screens will be listed (and written to the output file, if specified): the mean value, standard deviation and outliers (beyond 2 standard deviations from the mean); the quartiles and outliers (a la box plot); and the 10 observations with the most extreme values.
- A report file is created that contains the observation ID, original variable in standardized form ($Z_{varname}$), spatial lag, also standardized ($W_{varname}$), quadrant in the Moran scatterplot ($Q_{varname}$), Local Moran statistic ($I_{varname}$), probability level ($P_{varname}$), and significance ($S_{varname}$; 3 at $p < 0.001$; 2 at $p < 0.01$; and 1 at $p < 0.05$).

LISA Maps

- Create a view for the Nepal districts with a choropleth map for Cpr91. In a second view, create a theme with the Nepal shape file (nepal.shp). Select SpaceStat > LISA Local Moran Map and click on the name of the LM_nepal.txt file and the variable S_Cpr91. A map will show the locations with significant Local Moran statistics. To find out whether these indicate local clusters or local outliers, use the identify button (check the sign of Z_var).
- Make the first View (choropleth map) active again and select SpaceStat > Moran Significance Map with lm_nepal.txt as the file and M_Cpr91 as the variable. A new view will be created in which the significant locations are categorized according to their location in the Moran Scatterplot (high-high, low-low, etc.). Note how the significant locations in the north west of the country are mostly associated with a low value cluster.

Combining a LISA Map with a Moran Scatterplot

- With the first view active, launch the DynESDA toolbar and select Local Moran (the boxplot icon with an L on it) [make sure to set the contiguity criterion to the same type as used in your SpaceStat analysis, most likely Rook]. Two new windows will be created, one with the Moran Scatterplot, the other a box plot of LISA statistics. Select the outliers in the box plot and note which locations on the LISA map they correspond to. You can assess significance of the “global” Moran’s I by selecting this option in the Moran Scatterplot. As before, you can also assess how the Moran’s I statistic changes in subsets of the data by using the Exclude Selected option.

Assignment

Compute the G_i^* local statistics for spatial association for Cpr91 [Explore > G-stats > 6 New G_i^*] and visualize the significant locations by means of a G-stat map in ArcView (SpaceStat> G-stat map, select the gi_rook.txt file and the S_Cpr91 variable). Assess the differences between the Local Moran map and the G-stat map (use the Moran scatterplot). How would you interpret these differences?

Exercise 22 — SpaceStat Regression Basics

Topics

- regression problem files
- standard OLS output
- heteroskedastic variables
- robust OLS

Tutorial

- All the exercises on spatial econometrics will use the Columbus data set so that the examples in Anselin's Spatial Econometrics book can be replicated.
- Make sure you have a complete Columbus data set, first order contiguity file (rook) in both gal and fnt format, as well as some other weights files.
- Also set the output file option to a file name and set the long output option (3) to yes.

Regression Problem Files

- As in the Explore module, a problem for regression analysis is constructed using either the interactive format or a batch problem file.
- Make sure your current directory is the one with the Columbus data. Start Regress > 1 Classic Model > 1 OLS and select 2 Interactive: enter reg1.txt as the file name for the problem file.
- The first prompt is for the data set: enter columbus. Next you are given a list of spatial weights files in the current directory. It is good practice always to include spatial weights in the problem file, even though in this exercise, you will not yet focus on the spatial aspects. Enter the name for two weight files (e.g., rook and a distance-based or k nearest neighbor weights matrix), and press return to continue.
- Next you are asked to select the type of regression specification. SpaceStat has special functions for trend surface, spatial expansion, spatial regimes and spatial anova. For now, select 1 Generic Regression. Note that all the special cases can also be handled by means of the generic form, as long as you know how to construct the special variables and tests.

- The next series of prompts pertain to the variables in the model. First, press return to select a constant term (unless you have very good reasons not to — e.g., the specific form of a theoretical model — you should always use a constant term in the regression). Next enter the variables for the model, with the dependent variable first: enter Crime (return), Inc (return), Hoval (return), and return to stop.
- The following prompt lets you choose a specific set of variables to model (linear or additive) heteroskedasticity. Again, unless you have good reason to do otherwise (see below), take the default (press return). The default will use a random coefficient specification in the tests against heteroskedasticity.
- Next follow the results. For now, simply press return a few times until you are back at the regression menu.
- The model specification was saved in the file reg1.txt. Use a text editor to take a look at that file. The first “1” means that there is one specification in the problem file. Typically you will edit this file to create multiple specifications: make sure that the first number correctly gives the number of specifications. The next “1” identifies the model as a “generic” regression. On the same line, the “2” indicates that there are two weights files (their file names are on the next line), the “1” is a constant term flag, the next “1” indicates one endogenous variable (in this case simply the dependent variable, listed on the next line), and the last “2” shows that there are two explanatory variables (also listed on the next line). If you edit this file, always make sure that the number of explanatory variables (the 2 in this example) matches the number of variable names on the next line (the same goes for the spatial weights). A more detailed description of the problem file can be found in the SpaceStat manual.

Standard OLS Output

- Start 1 Ols > 1 Batch and type reg1.txt as the name of the problem file. After you press return, the first page of results appears. This contains the usual measures of fit, the coefficient estimates, standard errors, t-test values and associated probability. Note that a log likelihood is listed as well, in order to allow for the comparison of the fit with the spatial models. The results for both the unbiased and the consistent estimators of the error variance are reported as well (again to facilitate comparison with the ML models).

- Press return to clear the screen. If you have specified an output file, the results will be written to that file. With the “long output” option set to yes, you next see the complete variance matrix for the coefficient estimates (with long option to no, this is skipped).
- The next screen shows a multicollinearity condition number (less than 30 is OK, larger could be problematic) and the results of the Jarque-Bera asymptotic test for normality. If normality is rejected, i.e., if the p-value of the test is very small (less than 0.05), the properties of the ML estimation and LM tests may be affected since they are based on an assumption of normality (although how much they are affected depends on the situation). Note that in the Columbus case, there is no problem of either form.
- The following screen deals with heteroskedasticity. Two test statistics are reported. The Breusch-Pagan statistic (or, when normality has been rejected, its robustified form, the Koenker-Bassett statistic) and the White statistic. The BP statistic is computed for a random coefficient specification as the default, or uses the specified heteroskedastic variables if these are set explicitly. For the Columbus data, both tests reject the null hypothesis (p value < 0.05).
- Next follow the diagnostics for spatial dependence, discussed in Exercise 23.
- Finally, with the long output set to yes, the observed and predicted values and the residuals are listed for every observation (using the indicator variable to identify the observation).

Heteroskedastic Variables

- Reset the option for Long Output to no (unless you want to keep seeing a list with predicted values and residuals). Start 1 OLS > 2 Interactive and either type in het1.txt as the new name for a problem file. Go through the interactive steps as above till you come to the heteroskedastic variables. Now enter EW in return to the prompt. This will use an East-West dummy variables in the test for additive heteroskedasticity. If you have good reasons to assume other variables cause the heteroskedasticity (other than in a standard random coefficient model), this is where you would specify them.
- The OLS results will be the same as before, except for the test on heteroskedasticity. Note the significance of the test, suggesting groupwise heteroskedasticity. Try some

other variables by editing the problem file (replace EW by CP — center/periphery — or by DISCBD — distance to CBD). None of the other results is affected by the test.

Robust OLS

- As an alternative to specifying a particular form for the additive heteroskedasticity, you can take a robust approach. This gives OLS estimates, but with adjusted standard errors that are robust to unspecified forms of heteroskedasticity. SpaceStat reports two forms, one based on the White approach, the other on a Jackknife. The only difference between the two are standard errors and t-values. There are no specification tests for spatial effects in the robust models. That doesn't mean that there are none, but that there is no test available, nor is there an OLS result that is robust to unspecified forms of spatial dependence.
- Run `Regress > 1 Classic Model > 2 OLS Robust` in batch mode with `reg1.txt` as the problem file. Note the slight differences in t-values for the estimates.

Assignment

Use the Mississippi police data set to replicate the police expenditure model in Kelejian and Robinson (1992). Regress `Police` on `Tax`, `Transfer`, `Inc`, `Crime`, `Unemp`, `Own`, `College`, `White` and `Commute`. Run both OLS and Robust regression and assess the degree of multicollinearity, normality of the errors, and heteroskedasticity (ignore spatial autocorrelation for now).

Exercise 23 — Heteroskedasticity

Topics

- specifying heteroskedastic models
- groupwise heteroskedasticity
- random coefficient models

Tutorial

- This exercise uses the Columbus data set
- Make sure you have a complete Columbus data set, spatial weights files are optional for now, but should be entered in the problem files for later use.
- Also set the output file option to a file name so you can keep your results.

Specifying Heteroskedastic Models

- Models with additive heteroskedasticity can be estimated with the commands in the Regress > 3 Heterosked Error Model menu. There are three types of models with FGLS and ML estimator options for each.
- The generic model allows you to specify the variables that enter into the specification of the heteroskedasticity. These variables are used without transformations, so if you want them to enter as squared values, you first need to create such a variable (use the Data transformations).
- Groupwise heteroskedasticity lets you specify a categorical variables to create regimes or groups of observations, whereas random coefficient estimation takes the specified exogenous (explanatory) variables to set up the specification for the heteroskedasticity (with the variables squared).

Groupwise Heteroskedasticity

- In the Heteroskedastic Error Menu, select > 3 Groupwise heteroskedasticity (FGLS), choose interactive model and enter group.txt as the problem file. Use columbus for the data set, rook for the spatial weights, 1 for generic regression, take the constant term default and enter crime, inc, hoval as the variables for the model. The categorical variables to determine the “groups” is EW (this is the same variable used in the test against groupwise heteroskedasticity in Exercise 22).

- The estimation results show measures of fit, and the usual set of estimates, standard errors, etc. Next comes a screen with the estimates for the groupwise variances (note the difference). This is followed by a Wald test on heteroskedasticity (a t-test on equality of groupwise variances) and a set of tests for spatial autocorrelation.
- Select the ML option > 4 Groupwise heteroskedasticity (ML) and enter group.txt as the batch file. Note the minor differences with the results for the FGLS approach. First the number of iterations is reported. Also, the standard errors are slightly smaller than for FGLS. In addition to a Wald test, the results of a Likelihood Ratio test for groupwise heteroskedasticity are reported as well. The spatial diagnostics are the same as for FGLS.
- Check if the same results are obtained when CP is used as the indicator for the groups; compare your conclusion to what you found in the tests against heteroskedasticity in Exercise 22.

Random Coefficient Models

- Select option 5 Random coefficients (FGLS) and create a new problem file with the name random.txt. Enter the same information as before, but press return when asked about the heteroskedastic variables (if you enter a variable name, there will be an error message, and whatever you entered will be replaced by the random coefficient specification; the squares of the explanatory variables).
- The results consist of a set of estimates, the estimates for the heteroskedastic coefficients (note the negative coefficient for inc), a Wald test against heteroskedasticity (joint significance of the heteroskedastic coefficients) and the usual spatial diagnostics.
- Repeat the estimation of this model, but now use > 6 Random coefficients (ML) and enter random.txt as the batch file. The program aborts with an error message. While a negative heteroskedastic coefficient (as for inc) is not in itself a problem, when the resulting variance matrix does not remain positive definite, there can be no estimation. Try using the generic approach (1 or 2) with a subset of the variables to see if the problem can be avoided (use hoval or hoval squared).

Assignment

Use the Africa conflict data set to assess the extent of and model the heteroskedasticity in a regression of totcon on socmob, mcge, area, trade and govtyp. Check and model groupwise heteroskedasticity using the islam and ind60 indicator variables. Alternatively, try to construct a “regional” indicator variable using the select tool in ArcView (and add the variable to the data set) to assess its role in modeling heteroskedasticity.

Exercise 24 — Discrete Spatial Heterogeneity

SANOVA and Regimes

Topics

- spatial analysis of variance
- spatial regimes

Tutorial

- This exercise uses the Columbus data set
- Make sure you have a complete Columbus data set, spatial weights files are optional for now, but should be entered in the problem files for later use.
- Also set the output file option to a file name so you can keep your results.

Spatial Analysis of Variance

- Start with Regress > 1 Classic Model > OLS. Use interactive mode and create a problem file with the name anova.txt (for later use). Enter columbus for the data set and rook for the weights, but choose > 5 ANOVA as the type of regression. Next enter the variables for the model, the dependent variable crime first, followed by the indicator variable EW. Also choose EW as the heteroskedastic variable.
- The results show the estimates in a dummy variable regression with a constant term and the EW_1 indicator. Check for heteroskedasticity. Repeat the process for the CP dummy variable and compare the results (both in terms of what it means for “crime”, as well as in terms of model diagnostics).
- If you want the results without a constant term, you have to run them as a “generic” regression, do not select the constant term and enter both dummy variables for both 0 and 1 values as explanatory variables (you can create such variables in the data module). Try this for the EW indicator.

Spatial Regimes

- Again, start in OLS and use the interactive mode. Specify regimes.txt as the problem file. Enter columbus as the data set, rook for the weights, but now select > 3 Spatial Regimes. Enter the variables as before, but after crime, inc and hoval, use EW as the indicator variable for structural stability.

- The regression results are presented by regime, with the variable names followed by an underline and the value for the indicator variable. There is also a test on structural stability, both on all coefficients joints (chow test), as well as on each coefficient in isolation. Repeat the model with CP as the indicator variable and compare the results to the EW case.
- Since the EW model shows a strong indication of heteroskedasticity (the groupwise indicator is used), re-estimate the model using both a robust approach (OLS) as well as a groupwise heteroskedastic approach (FGLS and/or ML). Compare the results.

Assignment

Use the Africa conflict data set to assess the extent to which regimes can be used to specify the heteroskedasticity in the same regression as in Exercise 23. Use either ind60 or islam, or a regional dummy variable to construct the regimes. Run the analysis for OLS as well as for robust OLS and/or the heteroskedastic models if warranted.

Exercise 25 — Continuous Spatial Heterogeneity

Trend Surface and Spatial Expansion

Topics

- trend surface analysis
- spatial expansion

Tutorial

- This exercise uses the Columbus data set
- Make sure you have a complete Columbus data set, spatial weights files are optional for now, but should be entered in the problem files for later use.
- Also set the output file option to a file name so you can keep your results, specify polyid as the indicator variable and set the report file to 2 comma delimited.
- Start ArcView as well with the SpaceStat extension loaded and the Columbus theme active.

Trend Surface Analysis

- Start with Regress > 1 Classic Model > OLS. Use interactive mode and create a problem file with the name trend.txt (for later use). Enter columbus for the data set and rook for the weights, but choose > 2 Trend Surface as the type of regression. Next enter the dependent variable for the model, followed by X and Y as the coordinates in the trend surface (do NOT enter any other explanatory variables).
- Next enter 2 for the order of the trend surface and press return for the default in the heteroskedastic specification.
- The results show estimated coefficients for X and Y, as well as for the squares and cross products of the coordinates. Next follow the usual set of diagnostics; pay particular attention to the multicollinearity condition number. You could try out a trend surface with a power of 3 and see the effect on the fit of the model and on the multicollinearity.
- For the quadratic surface, with the proper report file option, a file will be created that allows for the mapping of the predicted values. Make ArcView active and Select SpaceStat > Predicted Map. Enter pr_crime.txt as the file name. A bar chart map of

the observed and predicted values will be constructed. Change the legend of the map to graduated colors and select P_Crime as the Classification Field. Note the gradual change in the predicted values.

Spatial Expansion

- Again, start in OLS and use the interactive mode. Specify expand.txt as the problem file. Enter columbus as the data set, rook for the weights, but now select > 4 Spatial Expansion. Enter the variables as before, crime, inc and hoval. Next, select X and Y as the expansion variables and choose a linear expansion (1); take the default for the heteroskedastic specification.
- The regression results are presented with the variable names preceded by AA_ and BB_ etc. to indicate the expansion. In addition to the standard diagnostics, there is also a test on the expansion (slightly off in the screen output).
- Note the values for the coefficient of inc and its expansions. Go back to ArcView and in the attribute table for Columbus, create a new variable as a linear combination of the estimate for inc and its coefficients with the x and y coordinates (if the estimates are b_0 , b_1 and b_2 , the new value is $b_0 + b_1x + b_2y$). Create a graduated color map with the new variable: this is a representation of spatial drift in the regression coefficient. Try the same for the coefficient of hoval.

Assignment

Use the Africa conflict data set to create a trend surface for the four types of conflict (totcon, totcop, verhos and versup). Make sure to compute the centroids if you have not done so before. Map the trend surface for each variable and compare the overall pattern. Alternatively, assess the extent to which a spatial expansion in the variables of the model for totcon (socmob, mcge, area, trade, govtyp) yields a useful map of the “spatial drift” in these parameters.

Exercise 26 — Diagnostics for Spatial Effects

Topics

- spatial diagnostics in least squares regression
- visualizing residuals
- spatial dependence and spatial heterogeneity

Tutorial

- This exercise uses the Columbus crime data.
- In SpaceStat, set the ID option to the variable Polyid, the Report file format to comma delimited, and the Output file option to save the results.
- Load ArcView after SpaceStat, add the SpaceStat extension and make a theme of the Columbus crime data active.

Spatial Diagnostics in Least Squares Regression

- Run a simple regression of crime on inc and hoval, and make sure the spatial weights files are specified so that the diagnostics can be computed. Use `Regress> 1 Classic Model> 1 OLS`, and use batch mode with the problem file generated in an earlier exercise (reg1.txt in Exercise 22), or interactively enter the name for the Columbus data set and the spatial weights files, select 1 generic regression, press return for a constant term and enter the variables in order. As in the previous exercises, the OLS results will be listed (and written to a file if you set the proper option), followed by a series of diagnostics.
- In the diagnostics for spatial dependence, assess the extent to which dependence is present in the form of a lag or error alternative; focus on the difference between LM error and LM lag and their robust counterparts. Compare the indications given by Moran's I on the one hand, and K-R and LM tests.
- When considering the spatial diagnostics, pay attention to the extent of normality and heteroskedasticity. How would these affect your inference?

Visualizing Residuals

- If you had the report file format set to comma delimited, switch to ArcView (otherwise, re-run the analysis with the proper report format).

- In ArcView, with the Columbus theme active, select SpaceStat > Residual Map, and click on pr_crime.txt as the file name with the residuals. A standard deviational map of the regression residuals will be created. As seen earlier, you can also make a bar chart map of observed against predicted values. “Visually” inspect the map to see if you suspect spatial autocorrelation. Compare this with the formal assessment in the diagnostics.

Spatial Dependence and Spatial Heterogeneity

- Repeat the heteroskedastic regressions from Exercise 23 (use the same problem files) and focus on the spatial diagnostics for LM Error and LM Lag. Assess the extent to which modeling a form of heteroskedasticity has eliminated the suggestion of spatial autocorrelation.
- Apply the regime regression and spatial expansion to the Columbus data. Again, assess the extent to which modeling a form of spatial heterogeneity can eliminate spatial autocorrelation.

Assignment

Use the Africa conflict data set and the problem files you created earlier to assess the extent to which there is spatial autocorrelation of a lag or error variety in the basic regression. Proceed with various models for heteroskedasticity and spatial regimes and check if anything changes. Make sure to run the spatial regimes model with groupwise heteroskedastic errors.

Exercise 27 — Systems Models

Topics

- specifying systems equations
- spatial diagnostics in 2SLS
- heteroskedasticity and 2SLS

Tutorial

- This exercise uses the Columbus crime data.
- In SpaceStat, set the Output file option to save the results.

Specifying Systems Equations

- The Systems menu in the Regression module implements two stage least squares with diagnostics for spatial effects, as well as estimators for 2SLS with spatial error dependence. You will apply these techniques to the same simple crime model, but now considering that “hoval” might be endogenous, i.e., simultaneously determined with crime.
- Select Regress > 5 Systems Model and choose the first option 1 2SLS. In interactive mode, create a problem file (system1.txt) and enter columbus for the data set, rook (and some others) for the spatial weights and generic for the type of model.
- In the query for the variables, enter crime as the first endogenous variable (the dependent variable) and hoval for the second endogenous variable, press return to end the list of endogenous variables, and return again to select the default in the exogeneity test.
- For the exogenous variables, enter inc, and select “discbd” as instrument.
- Note how the results change drastically compared to OLS (2SLS is a large sample procedure that is not necessarily very efficient in small samples).
- The Durbin-Wu-Hausman test on exogeneity strongly rejects the null hypothesis, suggesting there is indeed simultaneity between crime and housing values.

Spatial Diagnostics in 2SLS

- The final screen of diagnostics reports a LM test for spatial error autocorrelation. This is an asymptotic test and it is only valid if the endogenous variables are NOT spatially

lagged dependent variables (for proper diagnostics in the spatial lag case, use SAR – IV 2SLS in the spatial lag menu). The test suggests a problem with spatial error autocorrelation.

Heteroskedasticity and 2SLS

- While SpaceStat does not (yet) contain diagnostics for heteroskedasticity, two estimators are implemented that allow some form of heteroskedasticity.
- Using interactive mode, respecify the model in Systems Model > 2 2SLS – Ghet and use EW as the indicator for groupwise heteroskedasticity (this is between the exogenous prompt and the prompt for instruments). Assess the effect of including this form of heteroskedasticity on the standard errors. The different group variances are treated as “nuisance parameters” and thus no standard errors (or t-tests, etc.) are computed. This is shown in the results by zeros in those columns (the zeros do NOT mean that the standard error is zero). Repeat the process, but now use CP as the group indicator.
- An alternative to specifying groupwise heteroskedasticity is to leave the form unspecified. Select Systems Model > 3 2SLS Robust to implement the White robustified two stage instrumental variables estimator. Enter the data set, skip the weights (no tests for spatial autocorrelation are computed), and enter the variables as before. Compare the results to those for the other estimators.

Assignment

Assess the effect of selecting more instruments on the precision of the various estimators. In addition to discbd, now also include x and y as instruments. Compare the standard errors to those when only discbd was used. Also assess the effect on the exogeneity test. How are the heteroskedastic and robust estimators affected?

Exercise 28 — ML Estimation of the Spatial Lag Model

Topics

- ML estimation of the spatial lag model
- diagnostics for spatial effects
- spatial lag and heteroskedasticity

Tutorial

- This exercise will use the Columbus crime data.
- In SpaceStat, set Output file option to save the results; also make sure to set Long Output to yes (at least the first time you run this type of model).
- Make sure that the weights file is in “full” (fmt) format and row-standardized. If this is not the case, use Tools > 3 Weights Conversion > 1 Gal file to matrix format and make sure to check the row-standardization option.

ML Estimation of the Spatial Lag Model

- Run ML estimation for the spatial lag model by selecting Regress> 4 Space lag model> 1 SAR-ML. Use interactive mode (create a new problem file; you can use a problem file from OLS if you had the full spatial weights listed first, otherwise the estimation will fail with an error message that ML is not implemented for sparse weights). As before, enter the data set name for Columbus, specify at least one row-standardized spatial weights file (in full format), select generic regression.
- Enter the same variables as for OLS (crime, inc, hoval) and skip the heteroskedastic variable (press return for the random coefficient default).
- With Long Output on, all details on the iterations will be listed to the screen (and written to the output file, if specified); the regression estimates appear.
- Compare the values and significance of the coefficients to those of the OLS estimation. The log likelihood listed can be compared to the one from OLS, but note that the R^2 are only pseudo R^2 and should not be used in a rigorous assessment of model fit.

Diagnostics for Spatial Effects

- Following the estimates is a screen with diagnostics. First comes a LM test against heteroskedasticity. Note how there is still (after the spatial lag) strong evidence of heteroskedasticity.
- The next test result is for a LR test on the spatial lag variable. This test statistic should be compared to the asymptotic t-test on rho (listed as the t-value with the coefficient estimates) and to the LM test on a spatial lag in OLS. This is not a test on remaining spatial error autocorrelation.
- The last test is an (asymptotic) LM test on remaining spatial error autocorrelation. In this case there is no such indication. If the null hypothesis were rejected, that may point to a higher order model or to a problem with the specification of the weights.

Spatial Lag and Heteroskedasticity

- Select `Regress > 4 Space Lag Model > 2 SAR ML GHET` to implement groupwise heteroskedasticity in conjunction with a spatial lag. This is particularly useful when estimating a spatial regime model (see assignment).
- Use interactive model (and create a problem file) and enter the problem definition as for the standard spatial lag model, except to specify EW as the indicator variable for the heteroskedasticity.
- Note the effect on the estimation results (the significance of rho) and their standard errors.
- Next follow the estimates for the groupwise variances and a LR test on the null of equal variances. In this case, this is clearly rejected (compare to a LM test on heteroskedasticity in the spatial lag model when you specify EW as the heteroskedastic variable). There are no further diagnostics (yet) for spatial autocorrelation in this model. Repeat the process using CP as the indicator.

Assignment

Estimate the spatial lag model for the crime data with a spatial regime specification.

Estimate the model without and with taking groupwise heteroskedasticity into account.

Compare using EW to CP as the regime variable. Alternatively, you can investigate the

African conflict model with spatial regimes and a spatial lag relative to a pure spatial lag.

Exercise 29 — IV Estimation of the Spatial Lag Model

Topics

- IV estimation of the spatial lag model
- IV spatial lag with endogeneity
- explicit 2SLS estimation
- spatial lag and heteroskedasticity

Tutorial

- This exercise will use the Columbus crime data.
- In SpaceStat, set Output file option to save the results.

IV Estimation of the Spatial Lag Model

- Run IV estimation for the spatial lag model by selecting Regress> 4 Space lag model> 1 SAR-IV (2SLS). Use interactive mode (create a new problem file; or you can use a problem file from OLS). As before, enter the data set name for Columbus, a spatial weights file (sparse is fine), and select generic regression.
- At the first screen of variable prompts, enter the dependent variable, but do NOT enter the exogenous variables at this point. If there were other endogenous variables (not the spatial lag), this is where you would enter them.
- At the next prompt, enter inc and hoval as exogenous variables.
- The estimates are listed. Compare the values and significance of the coefficients to those of the ML estimation. The only measures of fit are pseudo R^2 and a squared correlation between observed and predicted and they should not be used in a rigorous assessment of model fit.
- The only diagnostic provided is an asymptotic LM test for remaining spatial error autocorrelation. This test takes into account the fact that one of the endogenous variables is a spatial lag (the test in the systems 2SLS model does not). There is no evidence of remaining spatial dependence.

IV Spatial Lag with Endogeneity

- Rerun the spatial lag IV estimation, but this time enter hoval as an additional endogenous variable. Enter inc for the exogenous variable and discbd, x and y as the instruments.
- Compare the results to the previous ones as well as to the systems model in Exercise 27 (note the effect on the significance of hoval). Again, there is no evidence of remaining error autocorrelation.

Explicit 2SLS Estimation

- Before you can estimate the spatial lag model using the generic 2SLS functionality, you need to make sure you have lagged variables for the dependent as well as explanatory variables. If they are not yet part of the columbus data set, create those lags for crime, inc and hoval using Data > 5 Space Transform > 2 Spatial Lag (make sure to specify “raw”).
- Select Regress > 5 Systems Model > 1 2SLS and use interactive mode to enter the columbus data set, contiguity weights (sparse is fine) and generic regression.
- In response to the variable prompts, enter crime and w_crime as the endogenous variables (return for the default in the exogeneity test), inc and hoval as exogenous, and w_inc and w_hoval as the instruments.
- Compare the estimates to those in the IV spatial lag routine: they should be identical. Note that the exogeneity test does not reject the null hypothesis (illustrating the rather weak power of the test in this case). Also note that the value for the LM error statistic is different from the one obtained in the spatial lag routine. When a spatial lag is used as one of the endogenous variables in Systems 2SLS, the statistic is NOT correct.

Spatial Lag and Heteroskedasticity

- Select Regress > 4 Space Lag Model > 4 SAR IV GHET to estimate the spatial lag model with groupwise heteroskedasticity. Note that the heteroskedastic parameters are treated as nuisance parameters and no inference is possible. Run this model for the crime data with in turn EW and CP as the heteroskedastic variable. Compare the results to the ML case.

- Select Regress > 4 Space Lag Model > 5 SAR IV Robust to use the White robustified two stage estimator. Use this model with caution, since its properties are not yet well understood. Compare the results to those in the groupwise model.

Assignment

As in Exercise 28, estimate the spatial lag model for the crime data with a spatial regime specification, but now use the 2SLS estimator. Estimate the model without and with taking groupwise heteroskedasticity into account. Compare using EW to CP as the regime variable. Alternatively, you can investigate the African conflict model with spatial regimes and a spatial lag relative to a pure spatial lag.

Exercise 30 — ML Estimation of the Spatial Error Model

Topics

- ML estimation of the spatial error model
- diagnostics for spatial effects
- spatial error and heteroskedasticity

Tutorial

- This exercise will use the Columbus crime data.
- In SpaceStat, set Output file option to save the results; also make sure to set Long Output to yes (at least the first time you run this type of model).
- Make sure that the weights file is in “full” (fmt) format and row-standardized. If this is not the case, use Tools > 3 Weights Conversion > 1 Gal file to matrix format and make sure to check the row-standardization option.

ML Estimation of the Spatial Error Model

- Run ML estimation for the spatial error model by selecting Regress> 2 Space Err Model> 1 SAR-ML. Use interactive mode (create a new problem file; as for the spatial lag case, you can use a problem file from OLS if you had the full spatial weights listed first, otherwise the estimation will fail with an error message that ML is not implemented for sparse weights). As before, enter the data set name for Columbus, specify at least one row-standardized spatial weights file (in full format), select generic regression.
- Enter the same variables as for OLS (crime, inc, hoval) and skip the heteroskedastic variable (press return for the random coefficient default).
- With Long Output on, all details on the iterations will be listed to the screen (and written to the output file, if specified); the regression estimates appear.
- Compare the values and significance of the coefficients to those of the OLS estimation. The log likelihood listed can be compared to the one from OLS, but note that the R^2 are only pseudo R^2 and should not be used in a rigorous assessment of model fit.

- Use the value for the log-likelihood to compare the fit of the spatial error model to that of the spatial lag model. Does this match what the LM diagnostics suggested?

Diagnostics for Spatial Effects

- Following the estimates is a screen with diagnostics. First comes a LM test against heteroskedasticity. Note how there is still (after the spatial lag) strong evidence of heteroskedasticity.
- The next test result is for a LR test on the spatial error variable. This test statistic should be compared to the asymptotic t-test on lambda (listed as the t-value with the coefficient estimates) and to the LM test on a spatial error in OLS.
- Next is both a Wald and LR test on the common factor hypothesis. For the model truly to be a spatial error specification, this hypothesis should NOT be rejected. In this example, this is the case.
- The last test is an (asymptotic) LM test on remaining spatial lag autocorrelation. In this case there is no such indication. If the null hypothesis were rejected, that may point to a higher order model or to a problem with the specification of the weights.

Spatial Error and Heteroskedasticity

- Select `Regress > 2 Space Err Model > 2 SAR ML GHET` to implement groupwise heteroskedasticity in conjunction with spatial error autocorrelation. This is particularly useful when estimating a spatial regime model (see assignment).
- Use interactive model (and create a problem file) and enter the problem definition as for the standard spatial error model, except to specify EW as the indicator variable for the heteroskedasticity.
- Note the effect on the estimation results (the significance of hoval) and their standard errors. Also compare these results to the matching ones in the spatial lag model.
- Next follow the estimates for the groupwise variances and a LR test on the null of equal variances. In this case, this is clearly rejected (compare to a LM test on heteroskedasticity in the spatial lag model when you specify EW as the heteroskedastic variable). There are no further diagnostics (yet) for spatial autocorrelation in this model. Repeat the process using CP as the indicator.

Assignment

Estimate the spatial error model for the crime data with a spatial regime specification.

Estimate the model without and with taking groupwise heteroskedasticity into account.

Compare using EW to CP as the regime variable. Also, compare the inference and fit to that in the matching spatial lag models. Alternatively, you can investigate the African conflict model with spatial regimes and a spatial error relative to a pure spatial error.

Exercise 31 — GM Estimation of the Spatial Error Model

Topics

- GM estimation of the spatial error model
- spatial error and heteroskedasticity

Tutorial

- This exercise will use the Columbus crime data.
- In SpaceStat, set Output file option to save the results; also make sure to set Long Output to yes (at least the first time you run this type of model).

GM Estimation of the Spatial Error Model

- Run the Kelejian-Prucha GM estimator for the spatial error model by selecting Regress> 2 Space Err Model> 5 SAR-GM (two step). Use interactive mode (create a new problem file or use a problem file from OLS). As before, enter the data set name for Columbus, specify one spatial weights file (sparse is fine; if more than one is specified the other ones will be ignored), select generic regression.
- Enter the same variables as for OLS (crime, inc, hoval) and skip the heteroskedastic variable (press return for the random coefficient default).
- Note that the estimate for lambda has no inference (zeros for standard error, t-value and probability) since it is treated as a nuisance variable.
- Compare the values and significance of the coefficients to those of the ML Error. Note that there is quite a discrepancy in the value of lambda.
- Repeat this procedure, but now use 6 SAR-GM (iterated). Note how close the estimate for lambda is to its counterpart in ML error.

Spatial Error and Heteroskedasticity

- Select Regress> 2 Space Err Model > 6 SAR GM GHET to implement the generalized moment estimator for groupwise heteroskedasticity in conjunction with spatial error autocorrelation. This is particularly useful when estimating a spatial regime model (see assignment).

- Use interactive model (and create a problem file) and enter the problem definition as for the standard spatial error model, except to specify EW as the indicator variable for the heteroskedasticity.
- Note the effect on the estimation results (the significance of ρ) and their standard errors. Also compare these results to the matching ones in the spatial lag model.

Assignment

As in Exercise 30, estimate the spatial error model for the crime data with a spatial regime specification, but now using the GM estimator. Estimate the model without and with taking groupwise heteroskedasticity into account. Compare using EW to CP as the regime variable. Also, compare the inference and fit to that in the ML estimation.

Alternatively, you can investigate the African conflict model with spatial regimes and a spatial error relative to a pure spatial error.