1. Spatial Data and Software

There are several sources of spatial data (free of charge) – Annex 1. Additional auxiliary data (detailed at least by country) that could be used in spatial demography applications include FAOSTAT – Food and Agriculture Organization statistical datasets (http://faostat.fao.org/), World Population Prospects – United Nations Population Division – (http://esa.un.org/unpp/), the World Bank (http://www.worldbank.org/) – several databases on education, gender, health and nutrition, and Human Development Report (HDR) data (http://hdr.undp.org/statistics/data/), among others. Specifically for the US, data on income and migration by county and state is available upon the payment of a small fee at the IRS – Internal Revenue Service website (http://www.irs.gov/taxstats/indtaxstats/article/0,,id=98123,00.html).

Different routines and software can be used to perform spatial data analysis. A good reference for users searching for appropriate software is the Center for Spatially Integrated Social Science (CSISS) website (http://www.csiss.org/). A list of freely distributed software is presented in Annex 2. Also, there are several commercial software for spatial analysis, including major statistical packages that incorporate special modules or specific functions to run spatial analysis. These include S-Plus®®, SAS, STATA, and MATLAB®®. The capabilities of performing spatial analysis are expanded through sharing (often free of charge) of codes written by users of these programs. Finally, GIS software, such as ArcView®® and ArcMap®®, also offer the capability of performing spatial analysis through specific modules that accompany the software or through scripts written by users and made available free of charge.

2. Hands-on

In this lab will cover two tasks: (i) perform a clustering analysis in ArcMap9, and (ii) take you through the steps to perform a spatial estimation in ArcMap9 (Geostatistical extension). We will use the malaria dataset shown in the lecture as an example. Afterwards, you are highly encouraged to use your own data.

(i) Spatial Patterns (spatial autocorrelation)

This exercise demonstrates the calculation of the $G_i^*(d)$ statistic presented in the morning lecture. The main idea is that points with extreme values (either high or low) are important, but points
with extreme values surrounded by other points with extreme values are even more so. The \( G_i^*(d) \) statistic facilitates the identification of such clusters of high and low values.

Your data should have one variable that has counts of an event or rates. This will be your target variable for the identification of clusters. The routine for the \( G_i^*(d) \) statistic is available in ArcToolbox.

1. **Opening ArcToolbox:**
   - You can open ArcToolbox simply by clicking on the icon in your menu bar. Alternatively, you can go to Windows menu, and click on ArcToolbox.

2. **Loading your data:**
   - Open the \textit{loca-95.shp}.

3. **Using ArcToolbox and inputting the parameters:**
   - In ArcToolbox, select \textit{Spatial Statistical Tools / Mapping Clusters / Hot Spot Analysis (Getis-Ord \( G_i^* \))}, as shown in Figure 1.
   - Double click on \textit{Hot Spot Analysis (Getis-Ord \( G_i^* \))}. This will open a window that allows you to select the parameters of the clustering analysis you are about to perform (Figure 1). Input parameters are:

   ![Figure 1 – ArcToolbox functions and input parameters for Hot Spot Analysis using \( G_i^*(d) \) statistic](image)
- **Input Feature Class** – the name of the layer with your data (*loca-95*).
- **Input Field** – the variable that contains the information you want to use to check for spatial patterns (for example, disease rates, number of events, registered crimes, etc). Choose *TX12*.
- **Output Feature Class** – the name of the file that will contain the results of the statistic.
- **Conceptualization of Spatial Relationships** – specifies how the weights will be computed. The options are:
  a. **Inverse distance** – the impact of one location on another decreases with distance
  b. **Inverse distance squared** – similar to inverse distance with a more sharply impact
  c. **Fixed distance band** – all locations within a specified critical distance receive a weight 1, and all those outside the critical distance are assigned a weight zero
  d. **Zone of indifference** – a combination of inverse distance and fixed distance band
  e. **Get spatial weights from file** – allows you to upload a pre-defined weight matrix
- **Distance method** – specifies how distances are calculated. There are two options:
  a. **Euclidian Distance** – straight line between two points
  b. **Manhattan Distance** – distance measured along axes at right angles (path)
- **Standardization** – this option allows the user to standardize the weights
- **Distance Band or Threshold Distance** – a cutoff value for distance
- Input your parameters with special attention to the **Distance Band**. The distance cannot be too small, so that each location ends up with no neighbors; and it cannot be too big, so that the defined neighborhood is so large that carries no meaning for the purposes of identifying spatial correlations. Ideally, you should try a distance that can be justified by your data. For example, in the case of malaria transmission I used a combination of three factors to define distance: the flight range of the mosquito, the average size of the plots, and the usual spatial configuration of local health interventions, resulting in 3,500m.
- Click **OK**. While the script is running, a pop-up window shows the status of the process (figure 2).
- Click **Close**, so that you can go back to your data frame window and check the results.
4. Checking your results:

- After you close the pop-up window, a new layer with the results (Output Feature Class) is automatically added to your data frame.

- Open the attribute table and note that the last column contains the z-scores associated to the Gi*(d) statistic. Display the z-scores with graduated colors, using critical values from a normal distribution as the cutoffs in your legend, so that you can easily access the significance for clustering. If you assume a 5% level, the cutoffs should be ± 1.96. All values equal or higher than 1.96 are significant for a clustering pattern of high values, while those equal or lower than – 1.96 are significant for a clustering of low values. All other values in between are not significant for a clustering pattern.

**PS1:** Remember that this is the interpretation assuming no correction for multiple and dependent tests.

You can correct for multiple testing using the False Discovery Rate (FDR) procedure, using an Excel spreadsheet and the following stepwise procedure (assuming a level $\alpha$):

a. order the test statistics $p$-values ($p_i$) in ascending order ($p_1 \leq p_2 \leq ... \leq p_m$);

b. starting from $p_m$ find the first $p_i$ for which $p_i \leq (i/m)\alpha$; and

c. regard all tests as significant for which $p_i \leq p_{critical} = (i/m)\alpha = p_{FDR}$. Therefore, if $p_{critical}$ equals 5% it means that, on average, among the rejected null hypothesis, 5% were truly null.
**PS2:** If you do not have point data, you might need to create X,Y coordinates that represent the centroids of the spatial unit of your data. This can be done using the following steps:

- Open ArcToolbox by clicking on the icon in your menu bar. Alternatively, you can go to Windows menu, and click on ArcToolbox.
- In Arctool box, select Data Management Tools / Features / Features to Point
- Input the name of your shapefile, and choose a name for the output file that will contain your point data.

**PS3:** If your data is in decimal degrees you need to re-project it before applying the indicators of spatial association, since decimal degrees are not meaningful units to work with (ideally you should work with meters, kilometers, or miles, for example). Each data requires different projection parameters. In ArcMap9 the projection routines are available in ArcToolbox:

- Open ArcToolbox by clicking on the icon in your menu bar. Alternatively, you can go to Windows menu, and click on ArcToolbox.
- In Arctool box, select Data Management Tools / Projections and Transformations / Define Projection

**PS4:** Instead of using ArcMap, you could run the same analysis in PPA – Point Patter Analysis: [http://www.nku.edu/~longa/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi](http://www.nku.edu/~longa/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi), with a DOS version provided at: [http://www-rohan.sdsu.edu/~aldstadt/tools.htm](http://www-rohan.sdsu.edu/~aldstadt/tools.htm).

**Suggested readings:**


(ii) Spatial estimation (variogram and kriging)

The spatial estimation, also known as kriging, is a minimum-mean-squared error spatial prediction method based on the second-order structure of a spatial process. It is the best spatial estimation that you can perform. The structure of the spatial process is given by the semivariogram, which describes how the spatial data varies as distance increases. ArcMap9 has an extension that performs kriging, named Geostatistical Analyst.

This exercise is only an initial guide of the basic steps for performing kriging on ArcMap9. For future reference, keep in mind that ArcMap has a very comprehensive help for that extension (both on-line and as a separate manual), which explains the different input parameters, and the indicators that should be considered when checking the performance of the model.

A more detailed exercise must address two issues. First, not all datasets are suited for kriging given the assumptions of the model. Second, a thorough kriging analysis takes time.

1. Loading your data:
   - Open the shapefile that contains your spatial and attribute information.

2. Loading the Geostatistical Analyst extension:
   - Go to Tools menu, and click on Extensions. Then activate the Geostatistical Analyst extension.
   - Go to View menu, and click on Toolbars. Then activate Geostatistical Analyst. The following menu box will appear in your screen.

3. Performing the kriging analysis:
   - Go to Geostatistical Analyst menu, and select Geostatistical Wizard. This will lead you through all the steps necessary to model a variogram and to perform kriging. ArcMap9 allows the use of different types of kriging, such as:
     - Ordinary, for data with unknown constant mean value
     - Simple, for data with known mean value
     - Universal, for data with mean value as a function on coordinates
     - Indicator, for discrete or data transformed to discrete
     - Probability, for discrete data as primary variable and continuous data as secondary variables
     - Disjunctive, for nonlinear predictions
     - Cokriging (multivariate version of the above-mentioned kriging models)
   - Choosing all the parameters requires deep knowledge of the process being analyzed, and of the technique itself. As mentioned before, kriging will not be the best method for all of
you, but it might be a good avenue to pursue for some. I’d be happy to talk with each one of you, and then give you some guidance regarding the appropriateness of kriging for your research.

**Suggested readings:**


GeoDa conducts ordinary least squares with diagnostics for spatial effects. If spatial effects are identified, the user has the option to perform maximum likelihood estimation of spatial lag and spatial error models:

1. Spatial lag model, also called a real contagion model, assumes that the value of the dependent variable in one area will be influenced by the values of that variable in the surrounding neighborhood; therefore, a weighted average of the dependent value for the neighborhood location is introduced as an additional covariate.

2. Spatial error model, also called a false contagion model, suggests omitted covariates; therefore an autoregressive error term is included.

Suggested readings (spatial autoregressive models):


### Annex 1 – Freely distributed datasets containing spatial information

<table>
<thead>
<tr>
<th>Data name/source</th>
<th>Description and Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Geographically Based Economic Data – G-Econ</td>
<td>A dataset that has information on total population and gross domestic product measured at a 1-degree longitude by 1-degree latitude resolution at a global scale (Nordhaus, 2006). <a href="http://gecon.yale.edu/">http://gecon.yale.edu/</a></td>
</tr>
<tr>
<td>7. US Census Bureau</td>
<td>Distributes geographic data that facilitates the use and analysis of census information on a spatial dimension, as well as census summary files, and public use microdata sample (PUMS) files. <a href="http://www.census.gov/">http://www.census.gov/</a></td>
</tr>
<tr>
<td>9. Demographic and Health Survey – DHS (Distributed by Macro International Inc.)</td>
<td>Nationally-representative household surveys conducted in developing countries, typically every 5 years, collecting information on population, health, and nutrition. Since 1996, DHS surveys have been collecting locational data (Rutstein, 2000) that not only allows the use of spatial analytic techniques, but also facilitates linking DHS data with other spatial datasets. <a href="http://www.measuredhs.com/">http://www.measuredhs.com/</a></td>
</tr>
<tr>
<td>10. Latin America &amp; Caribbean Demographic Centre – CELADE/CEPAL</td>
<td>Census data from selected countries of Latin American. Data is made available through REDATAM (an acronym for REtrieval of DATa for small Areas by Microcomputer), a software that facilitates tabulation and mapping of data. <a href="http://www.eclac.cl/cgi-bin/getProd.asp?xml=/redatam/noticias/paginas/7/13277/P13277.xm1&amp;xsi=/redatam/tp/p18f.xsl&amp;base=/redatam/tp/top-bottom.xsl">http://www.eclac.cl/cgi-bin/getProd.asp?xml=/redatam/noticias/paginas/7/13277/P13277.xm1&amp;xsi=/redatam/tp/p18f.xsl&amp;base=/redatam/tp/top-bottom.xsl</a></td>
</tr>
</tbody>
</table>
### Annex 2 – Freely distributed software for spatial statistical analysis

<table>
<thead>
<tr>
<th>Software</th>
<th>Description and website</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GeoDa™</td>
<td>Developed at the Spatial Analysis Laboratory (SAL) at the University of Illinois, Urbana-Champaign. Includes techniques of spatial autocorrelation (local and global indicators) and spatial regression (OLS, spatial lag and spatial error models) (Anselin, Syabri and Kho, 2006). <a href="https://geoda.uiuc.edu">https://geoda.uiuc.edu</a></td>
</tr>
<tr>
<td>2. CrimeStat</td>
<td>Distributed by the National Institute of Justice (NIJ). Software developed for crime mapping and analysis, although suitable for other applications. Main features include spatial autocorrelation, hot spot analysis, spatial interpolation, space-time analysis, and models of travel behavior (Levine, 2006). <a href="http://www.icpsr.umich.edu/CRIMESTAT/">http://www.icpsr.umich.edu/CRIMESTAT/</a></td>
</tr>
<tr>
<td>3. STARS – Space-Time Analysis of Regional Systems</td>
<td>An open source software for the analysis of areal data measured over time. Includes techniques for exploratory spatial data analysis (global and local indicators), and for space-time analysis (quantification and decomposition of inequality, Markov models, and mobility dynamics), combined with a powerful and interactive geovisualization component (Rey and Janikas, 2006). <a href="http://regal.sdsu.edu/stars/">http://regal.sdsu.edu/stars/</a></td>
</tr>
<tr>
<td>4. PPA – Point Pattern Analysis</td>
<td>Includes several techniques, such as nearest neighbor methods, Ripley’s K, local and global spatial autocorrelation statistics, Knox statistics for space-time clustering, and spatially filtered regression. <a href="http://www.nku.edu/~longa/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi">http://www.nku.edu/~longa/cgi-bin/cgi-tcl-examples/generic/ppa/ppa.cgi</a></td>
</tr>
<tr>
<td>5. SANET – Spatial Analysis on a NETwork</td>
<td>Designed to analyze spatial network phenomena. This program includes 13 different tools, and, among other capabilities, facilitates the assessment of proper distribution/location of services (e.g. hospitals) (Okabe, Okunuki and Shiode, 2006). <a href="http://okabe.t.u-tokyo.ac.jp/okabelab/atsu/sanet/sanet-index.html">http://okabe.t.u-tokyo.ac.jp/okabelab/atsu/sanet/sanet-index.html</a></td>
</tr>
<tr>
<td>6. R language</td>
<td>An open source program. Several packages for spatial data analysis are made available, which can be applied to point and areal data (geostatistics – package <code>geoR</code>, point pattern analysis – package <code>spatstat</code>, empirical Bayes estimation – package <code>spBayes</code>, clustering analysis – package <code>Dcluster</code>, etc) (Bivand, 2006). <a href="http://cran.r-project.org">http://cran.r-project.org</a> and <a href="http://sal.uiuc.edu/csiss/Rgeo/">http://sal.uiuc.edu/csiss/Rgeo/</a></td>
</tr>
<tr>
<td>8. Geostatistical code library</td>
<td>Hosted at Stanford University, offers several codes to run numerous types of geostatistical models (Deutsch and Journel, 1992). <a href="http://ekofisk.stanford.edu/SCRFweb/supporting/index.html">http://ekofisk.stanford.edu/SCRFweb/supporting/index.html</a></td>
</tr>
<tr>
<td>9. SaTScan™</td>
<td>Analyzes spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics. Has a wide application on epidemiology, demography, criminology, and history, among other fields. Developed and distributed by Martin Kulldorff (Kulldorff, 1997). <a href="http://www.satscan.org/">http://www.satscan.org/</a></td>
</tr>
</tbody>
</table>
References (Annex):


