Tutorial: Exploratory Data Analysis with GeoDa

D G Rossiter
Cornell University, Section of Soil & Crop Sciences

March 7, 2024

Contents

This exercise is an adaptation and extension of one prepared by Dr. Diana Sinton for Cornell course PLSCS/NTRES 6200 in 2016.

1 Introduction

GeoDa is an open-source program, cross-platform program designed as a simple tool for exploratory spatial data analysis (ESDA) and some spatial modelling of **spatial polygon** data, that is, maps of polygon units such as census tracts or political divisions with a set of **attributes** measured on each one.

GeoDa was first developed at Arizona State University and is now hosted at the University of Chicago[1]. The GeoDa program, documentation and sample data is freely available for download from the Geodata Center's GitHub[2]. A recent journal paper [1] explains the history of GeoDa and its applications.

GeoDa allows users to experiment with visualization functionality such as linking and brushing across windows. This can be very helpful both for interpretation of and communication about these spatial patterns. It also incorporates several spatial statistical models.

Task 1 : Download and install GeoDa. •

Task 2 : Install example datasets in a convenient location. •

GeoDa comes with some sample data, but many GIS file formats[3], including ESRI Shapefiles with associated databases, can be used. This tutorial should come with two compressed files (1) `NY_data.zip` and a subset of this, (2) `Syr.zip`. Uncompress these to find the Shapefiles, with extension `.shp`.

This is the "New York leukemia dataset" developed by Waller and Gotway [3] and adapted by Bivand et al. [2]. It contains information on the census tracts (the spatial units) in an eight county area including Syracuse (NY) city, relating possible causes to the incidence of leukemia, in particular, exposure to the industrial chemical TCE[4].

> **Note:** In the USA census tracts have 1 500–8 000 people (optimum size 4 000). They are designed to be socio-economically and demographically fairly homogeneous. Each tract has several block groups; these are made up of 20–40 individual blocks. The tract is usually large enough to compile reliable statistics.[5].

I have reduced this to just Syracuse city in the second example to reduce the size of maps and graphs for this tutorial, but you may prefer to work with the full dataset, which includes both urban and rural areas.

Task 3 : Start GeoDa. •

[1] https://spatial.uchicago.edu/geoda
[2] http://geodacenter.github.io
[3] https://geodacenter.github.io/formats.html
[4] Trichloroethylene, an industrial solvent often found in groundwater
[5] https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf

2 Dataset

Task 4 : Load the Syracuse leukemia incidence dataset into GeoDa, using the `File | New ...` menu item or the file open icon. This is a shapefile with base name `Syr`, so select `Syr.shp` •

You will see a plain map of the polygons (Fig. 1).

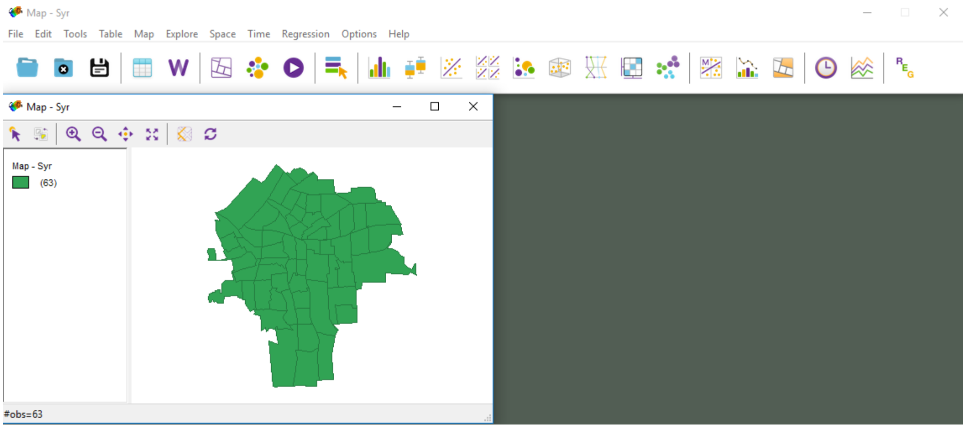


Figure 1: Base map and GeoDa tool bar

Task 5 : Open the data table by clicking the Table icon. Examine the rows and columns. •

We will work with these variables (fields):

These are all reported on the basis of 1980 census tracts. First, the **response** (dependent) variables:

`Cases` : the number of leukemia cases 1978–1982; some cases had insufficient georeference, these were added proportionally to tracts, so some "counts" are not integers.

`Z` : log-transformed rate, i.e., normalized by census tract population: $Z_i = \log(1000[\text{Cases} + 1]/n)$, where n is the population of the tract.

Second, possible **predictors**:

`PEXPOSURE` : potential exposure, computed as the logarithm of 100 times the inverse of the distance between a census tract centroid and the nearest TCE-producing site;

`PCTAGE65P` : percent older than 65 years; this could represent long-term exposure to any environmental factor;

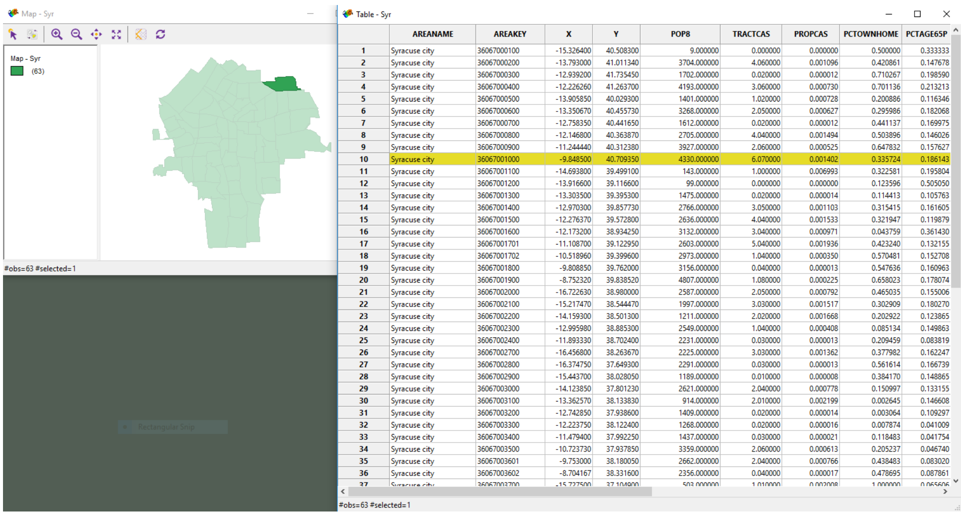
`PCTOWNHOME` : percent home ownership; this could indicate lifestyle or economic level.

Task 6 : Rearrange the Table and the Map so that you can view both. •

The basic GIS “linking” functionality is in place; you can click on polygons in the map and their associated records in the table will highlight, and vice versa. To unselect objects, click anywhere in the white area surrounding the map or at the upper-left of the table. You can select multiple polygons (on the map) or tracts (in the attribute table) with Shift-click for a set or Ctrl-click to add one-by-one. You can also “brush” over the map by holding down the left mouse button, to select in a window,

---

Q1 : *Click on the northeaternermost census tract. What is its AREAKEY? What is its population? What percent of its homes are owned rather than rented?* •



3 Exploratory Data Analysis

3.1 Univariate

Task 7 : Display some themed maps in the `Map` menu, for one or more of the variables, for example `PCTSGE65P`. •

Compare quantile, percentile, box, and natural breaks maps. Examine how they present the same theme in different ways (Fig. 2).

Q2 : *Look at the southeasternmost census tract in these four maps. How do they describe its proportion of older residents, compared to the entire City? Which map(s) best show(s) whether it is unusual?* •

Q3 : *Which map is best for assessing spatial autocorrelation of this variable? Why? Does there appear to be autocorrelation? Across how many neighbouring census tracts?* •

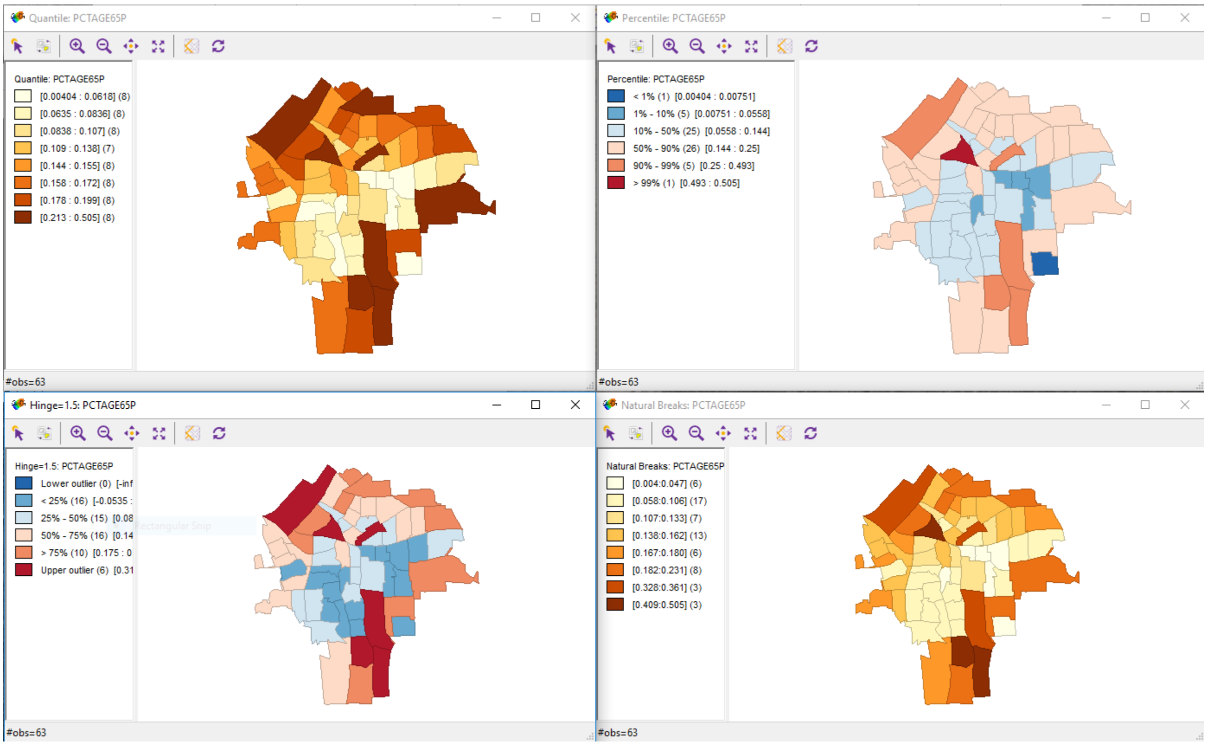


Figure 2: Thematic maps

3.2 Bivariate

Now we explore some feature-space plots.

Task 8 : Under the Explore menu, create a Histogram of PCTAGE65P. Also create a Scatter Plot of the proportion of residents over the age of 65 PCTAGE65P (Y variable) and the proportion of homes that are owned rather than rented. PCTOWNHOME (X variable) •

See Figure 3.

Q4 : *Which tract has the highest proportion of older residents?* •

Q5 : *Describe the relation between these two attributes.* •

Task 9 : Find an unusual tract (not fitting the overall pattern for the city) and click on its point in the scatterplot. •

See Figure 4.

Because all of the individual graphic elements for each of the 63 polygons are linked, any time that one or more are selected in one window

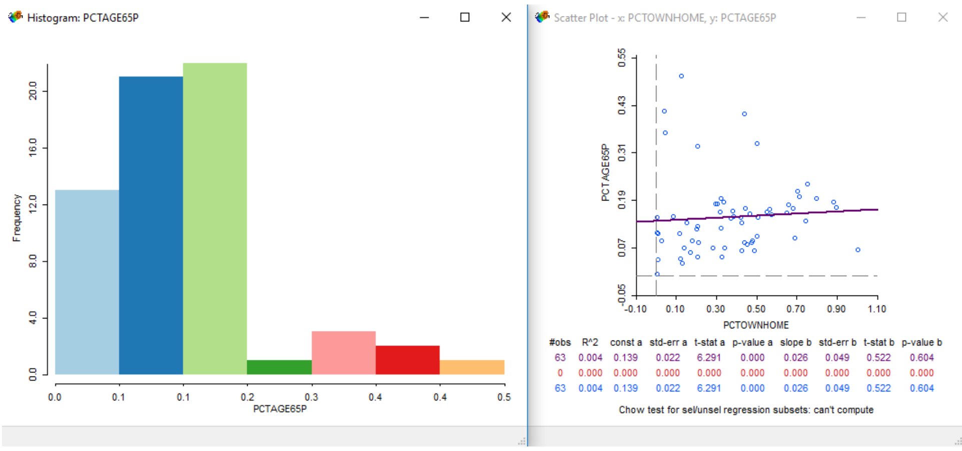


Figure 3: Histogram and scatterplot

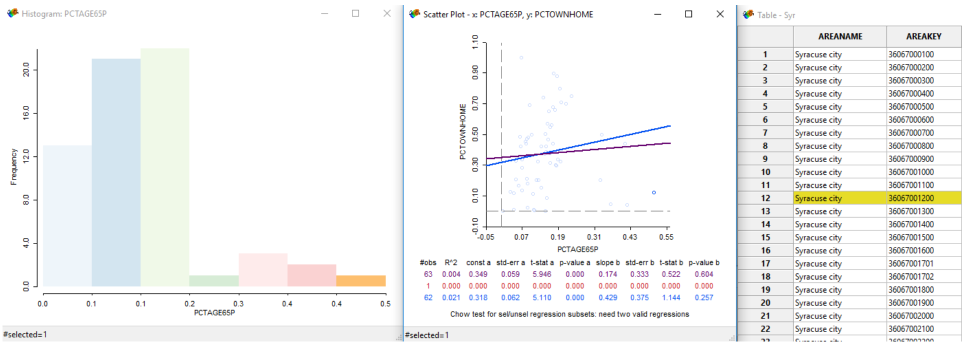


Figure 4: An unusual tract

or in one of the exploration plots, their linked highlighted display will activate in all other windows.

Q6 : *Which tract did you select? Where is it located?* •

TASK 10 : Brush over the southern few tracts by holding down the left mouse button as you define a rectangular window. •

Q7 : *What is the overall relation between home ownership and proportion of over-65 residents? What is this relation for the four southernmost tracts? How do you explain this?* •

Another interesting plot is the Cartogram.

TASK 11 : Make a cartogram (`Map | Cartogram`) of the proportion over 65 years old `PCTAGE65P` as the circle *size*, with the disease incidence `Z` as the circle *colour*. •

See Figure 5.

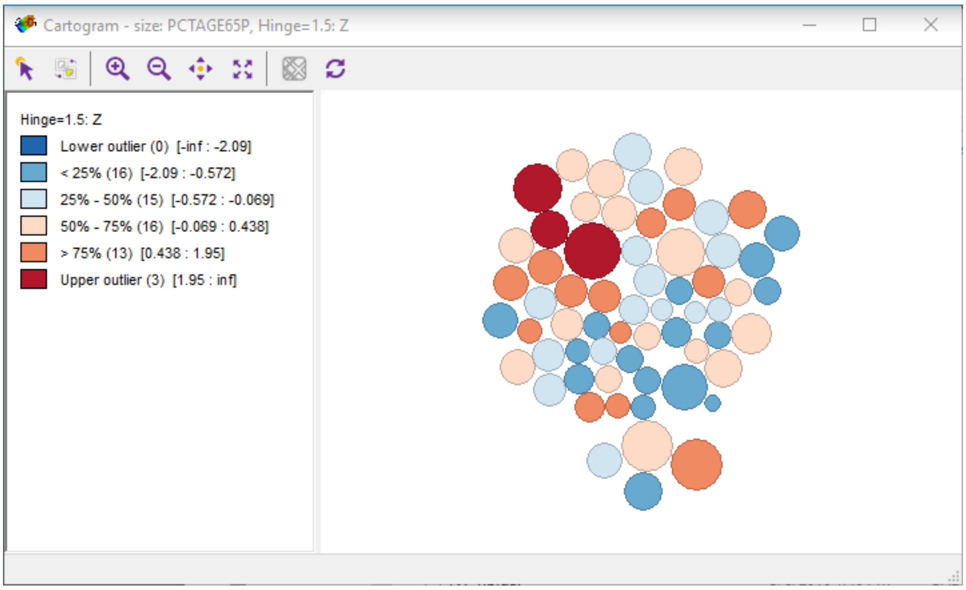


Figure 5: Cartogram of leukemia incidence vs. older residents

---

Q8 : *How are the circles placed in the plot? What insight does this give you into the relation between disease incidence and older residents?* •

3.3 Multivariate

Task 12 : Open a scatterplot matrix (`Explore | Scatter Plot Matrix` of the three `PCTOWNHOME`, `PCTAGE65P`, and `PEXPOSURE`, as well as the response variable `Z`. •

See Figure 6.

Q9 : *Describe the feature-space distributions of the four variables. Looking at the proposed bivariate linear regressions, which have tracts with high leverage, i.e., that greatly influence the line?* •

Task 13 : In the matrix, select the histogram bar for the lowest proportion of home ownership, i.e., where more households rent. •

Note how the linked maps highlight these tracts. See Figure 7.

Task 14 : With the scatterplot matrix displayed, select menu option `Options | View | Regime Regression`. This will then show the separate regression lines and statistics for the overall, selected, and non-selected census tracts. •

Q10 : *Look at the red proposed regression lines – which bivariate corre-*

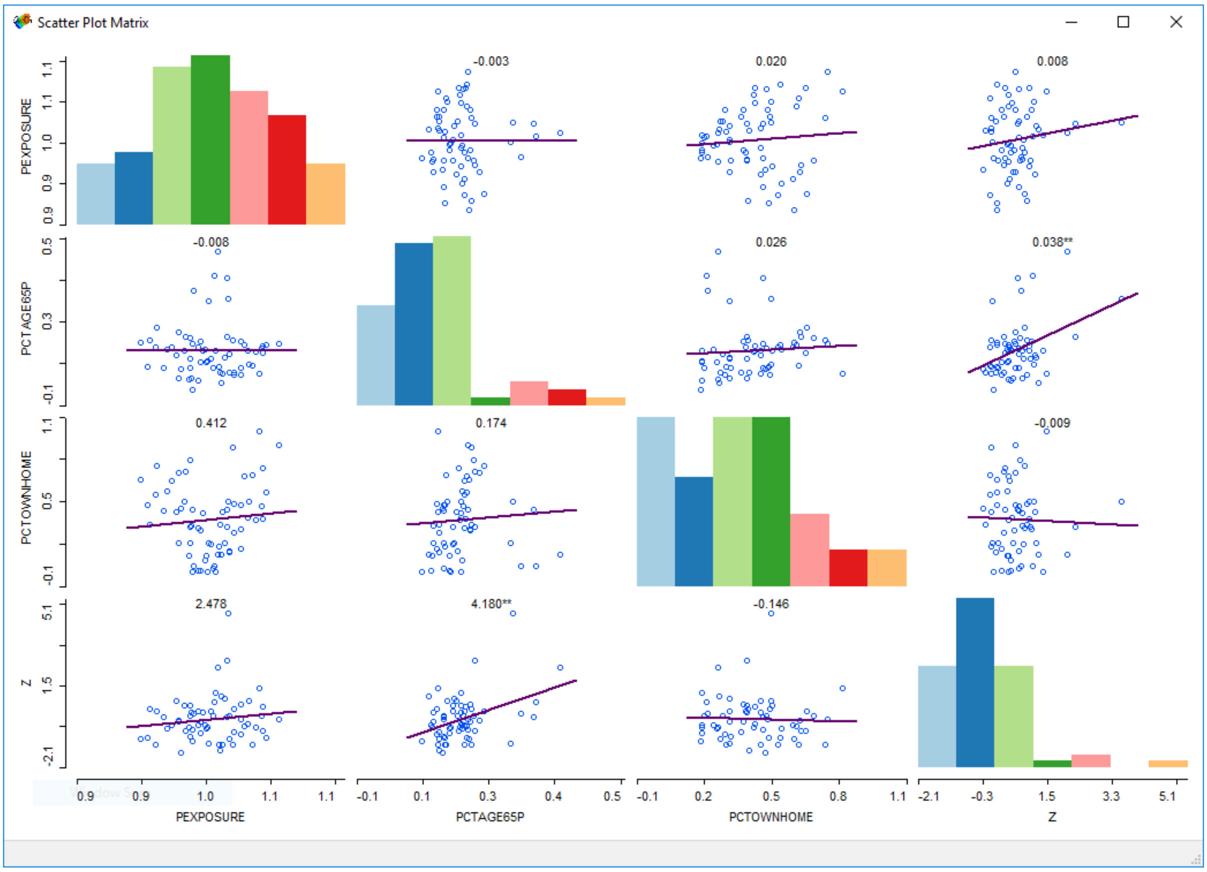


Figure 6: Scatterplot matrix

lations are substantially different from the overall correlations if we only consider these tracts? •

TASK 15 : Open a Parallel Coordinate Plot (PCP) and Include the three possible predictors `PCTOWNHOME`, `PCTAGE65P`, and `PEXPOSURE`, as well as the response variable `Z`. •

Q11 : *What is their overall inter-relation?* •

TASK 16 : Click on the line to the highest response. •

See Figure 8.

Q12 : *Which tract is this? How is this response related to the three predictors?* •

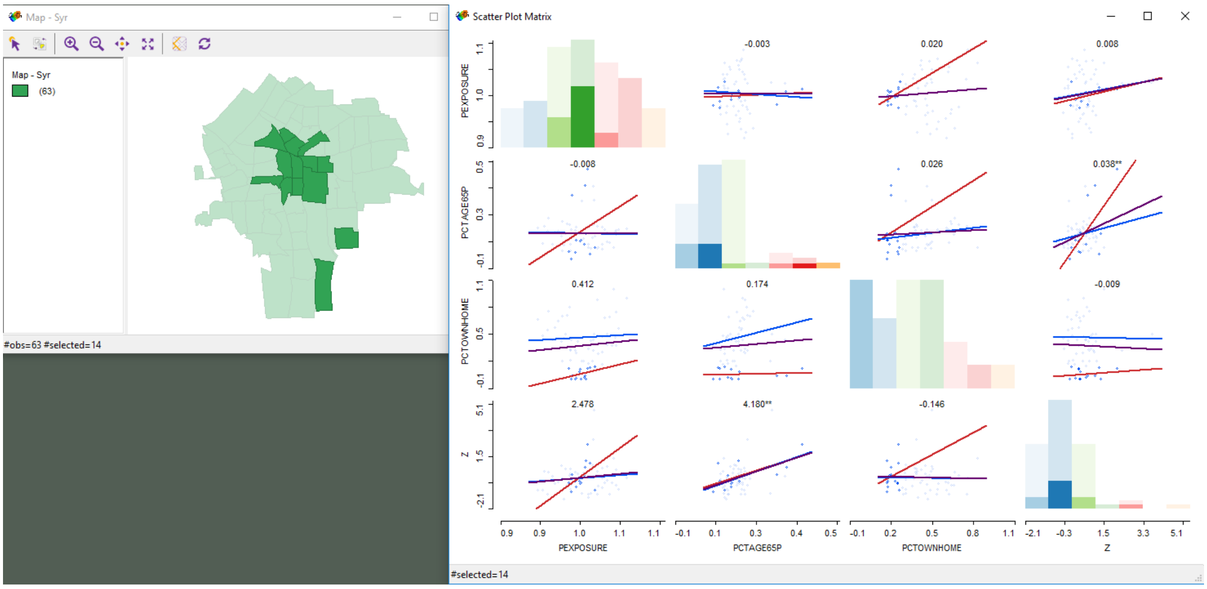


Figure 7: Scatterplot matrix with low home ownership tracts selected

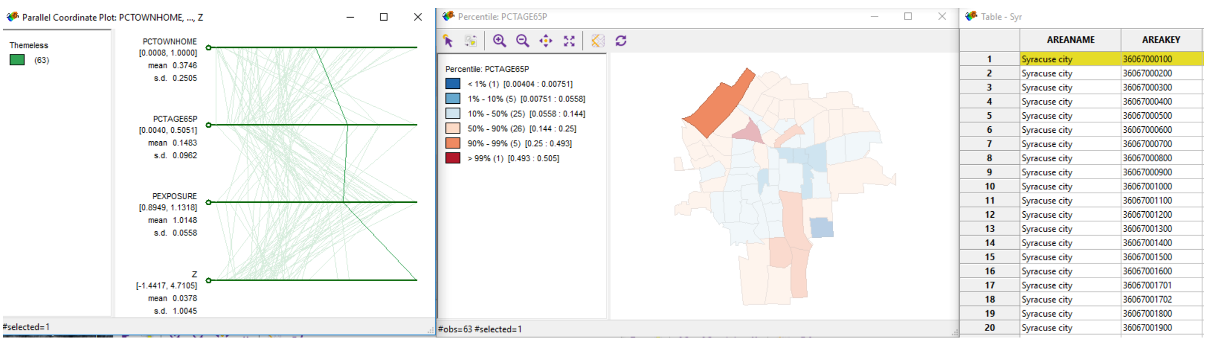


Figure 8: Parallel coordinate plot, highest response selected

4 Clustering

We may want to see which areas go together in **clusters**, based on one or more factors. These are groupings with *minimum* **pooled within-class** differences, and thus *maximum* **between-class** differences. These variances can be **parametric**, i.e., based on **means**, or **non-parametric**, i.e., based on **medians**.

4.1 k-means

One popular method is **k-means**, with a user-specific number of clusters k. There are many methods to find an "optimum" number of clusters, but we will not investigate those here, rather, we will decide on a number. This could be, for example, to stratify the area for a survey or marketing campaign.

We begin with clustering based on a single variable, home ownership.

Task 17: Select menu item `Clusters | K means` and in the dialog box select `PCTOWNHOME` and 4 clusters. Run the analysis. See Figure 9. This should produce the map in Figure 10 •

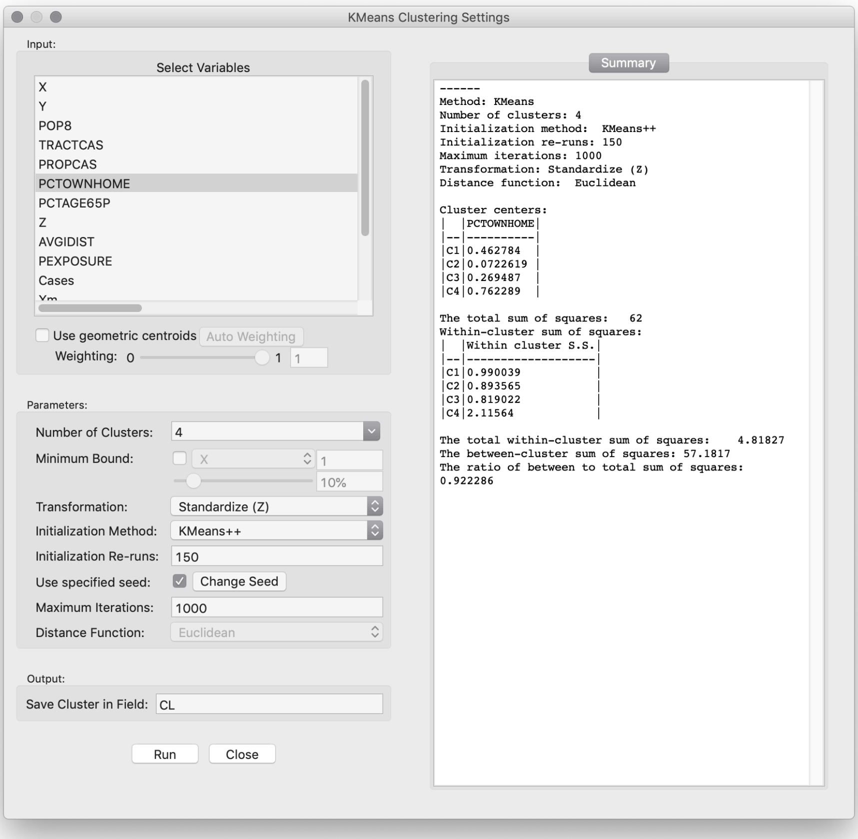


Figure 9: Creating a k-means cluster

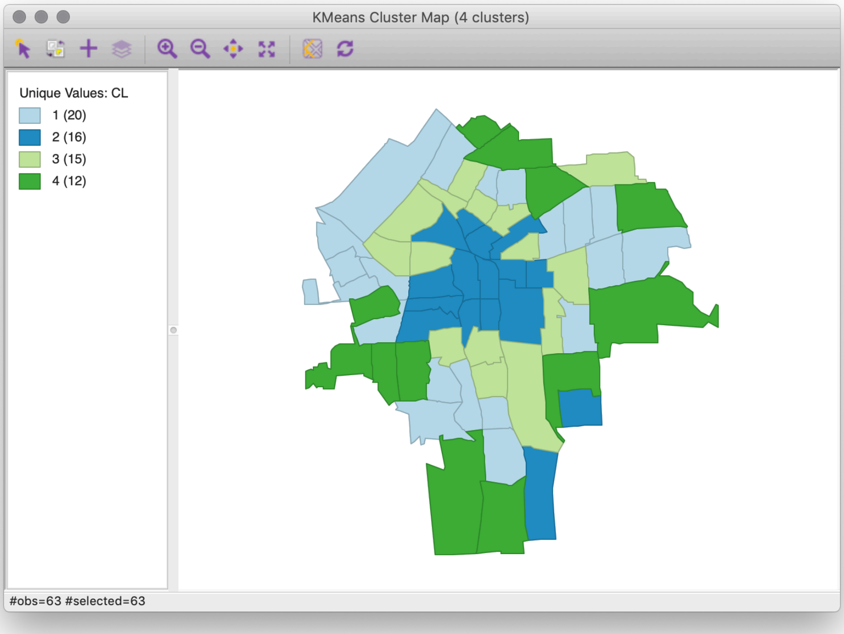


Figure 10: Four k-means clusters

Q13 : *Describe the spatial pattern of the clusters. How is home ownership distributed across Syracuse?*

•

Q14 : *Do the clusters have the same number of census districts? Should they?*

•

Q15 : *How much of the total sum of squares of the clustering variable **PCTOWNHOME** is explained by the clustering?*

•

Q16 : *Which cluster is most homogeneous in feature space, i.e., has the least within-cluster sum of squares? The most?*

•

Q17 : *The cluster centers are the mean of the target variable for the census blocks within the cluster. How different are they?*

•

4.2 Geographically-compact k-means

What if we required **geographically-compact** clusters? For example, most voting district laws require at least contiguous areas, and some prefer as compact as possible.

TASK 18 : Repeat the k-means clustering for four clusters, but select the "Use geometric centroids" option, with **auto-weighting**. This forces the clusters to be geographially-contiguous, while giving the least weight possible to that aspect, keeping as much feature-space homogeneity as possible. See Figure 11. This should produce the map in Figure 12. •

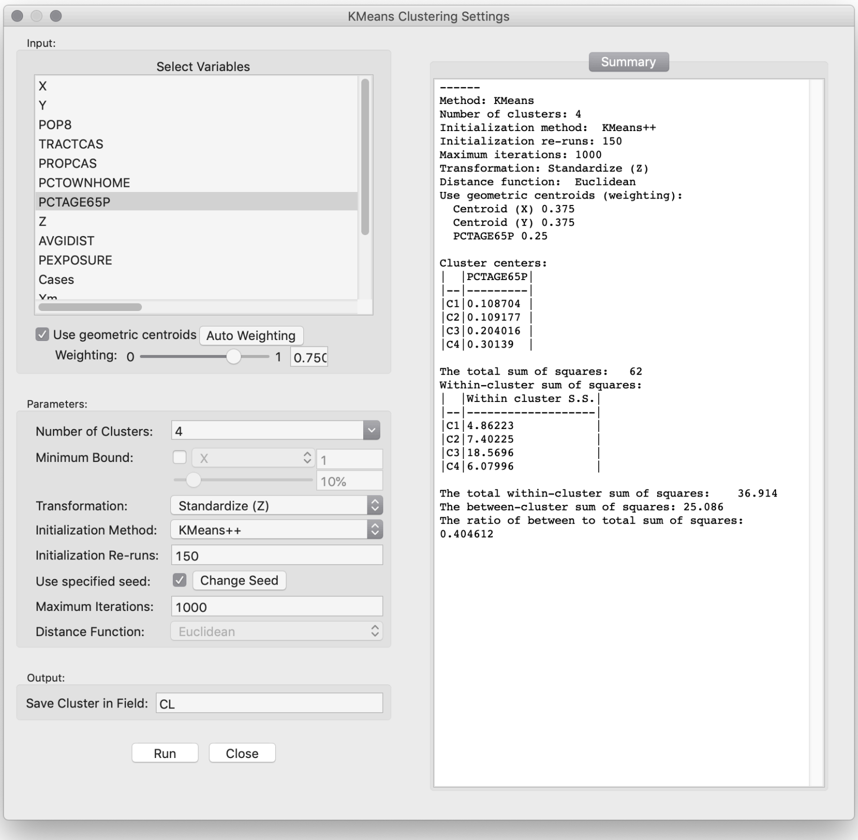


Figure 11: Creating a compact k-means cluster

Q18 : *How much of the total sum of squares of the clustering variable* ***PCTOWNHOME*** *is explained by the clustering? Why is this so much lower than for the k-means without forcing compact clusters?* •

Q19 : *Which cluster is most homogeneous in feature space, i.e., has the least within-cluster sum of squares? The most?* •

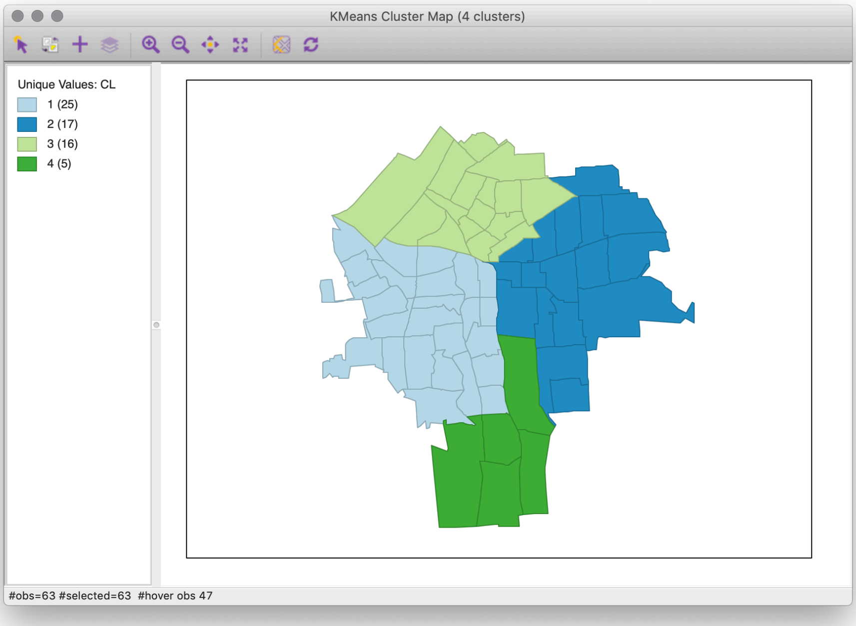


Figure 12: Four k-means compact clusters

Q20 : *The cluster centers are the mean of the target variable for the census blocks within the cluster. How different are they?* •

Q21 : *Why might a policy maker prefer the compact clusters?* •

Optional: You can balance compactness in geographic and feature space by changing the weighting given to the geometric centroids.

4.3 k-medians

Optional: You can used the **medians** of the clusters, rather than the means, to define the within- and between-class sum-of-squares. That is, rather than calculating the mean for each cluster to determine its centroid, instead calculate the median. This is **k-medians**. Run a 4-cluster k-median analysis and compare with the 4-cluster k-means.

4.4 k-mediods

Optional: You can used the **mediods** of the clusters, rather than the means, to define the within- and between-class sum-of-squares. A **mediod** is one of the observations, also called the **exemplar** of the cluster. This allows the identification of the most "typical" observation in a cluster. This is **k-mediods**. Run a 4-cluster k-mediod analysis and compare with the 4-cluster k-means.

See Figure 13. Here the cluster center values are actual values of one of the census blocks in the cluster. This should produce the map in Figure 14; notice the exemplars.

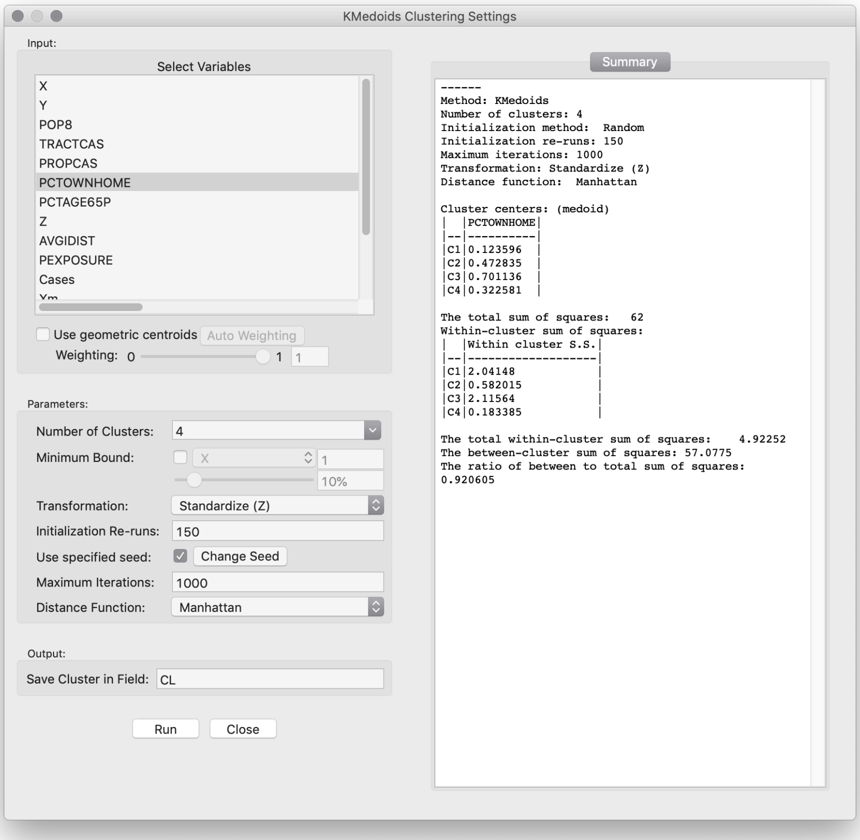


Figure 13: Creating a k-mediods cluster

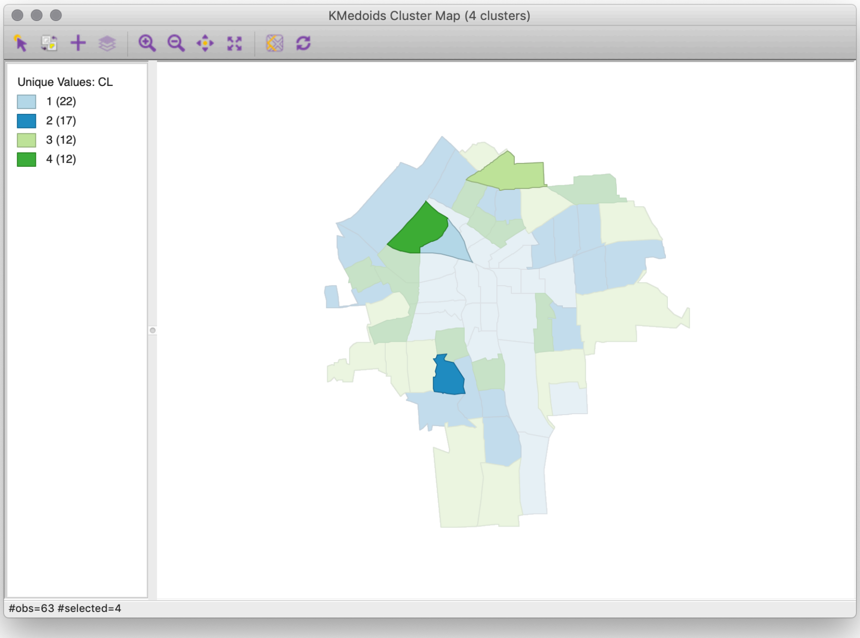


Figure 14: Four k-mediods clusters, exemplars, i.e., “typical” census blocks, highlighted

5 Neighbors and Distances

For spatial models, we must impose some **spatial structure** on the 63 polygons, that is, how they are related in space. Then we can assess this statistically.

One way is by **distance between polygon centroids**, as in point geostatistics; the spatial weights are based on separation, typically as inverse distance. This considers that distance is the only factor driving any spatial correlation.

However, there are other ways to build a **weights matrix** that relates neighbours; these relate to different hypotheses about how space affects the response. For example, a binary neighbours weighting considers that all first-order neighbours contribute equally to any spatial effect, i.e., it is averaged across the neighbours. With this weighting every tract is influenced equally (on average) by neighbours, and this influence is divded among the neighbours.

Task 19 : Generate two weights files: (1) Distance Weights and indicate X and Y coordinates, (2) order-1 Contiguity with Queen neighbours (i.e., tracts meeting only at a point are also considered to be neighbours). •

To generate a Weights File, choose `Tools | Weights Manager | Create`. Every shape must have its own unique ID, so check the `Add ID Variable` and use the existing `AREAKEY` variable. By default, the new ID variable will be named `POLY_ID`, or you can choose otherwise. You can then select a type of weighting methods

For the distance weighting, the Threshold distance will automatically be calculated at the minimum distance to ensure that every polygon has at least one neighbor, but you can set any distance you desire. The distance units will be the units associated with the shapefile. For example, if we think that the phenomenon might be spatially-correlated (after accounting for the feature-space regression) to 2.4 km, set the threshold distance to 2400 m.

See Figure 15.

Q22 : *What is the maximum number of neighbours considered for any tract in this distance weighting?* •

> **Note:** You could also use the k-Nearest Neighbors option to specific a set number of its closest neighbors that you wish each polygon to use.

When your choices are set, Create the file and name the file with a label indicating the approach used to calculate the neighbors. For example, `Queen1.gal` would indicate a Queen directionality with 1 order of contiguity; `Dist24.gwt` would indicate a 2400 m radius inverse-distance weighting.

As you are deciding which method derives the most valid weights file for your question of interest, you can visually compare the results by using

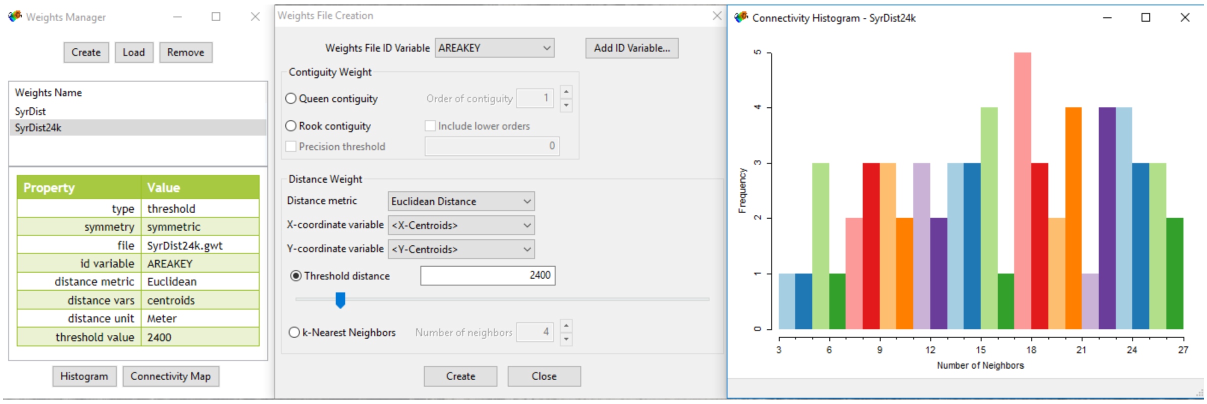


Figure 15: Creating a distance weighting

the Connectivity Histogram button in the Weights Manager window, each time after you select your different weights tables. See Figure 16.

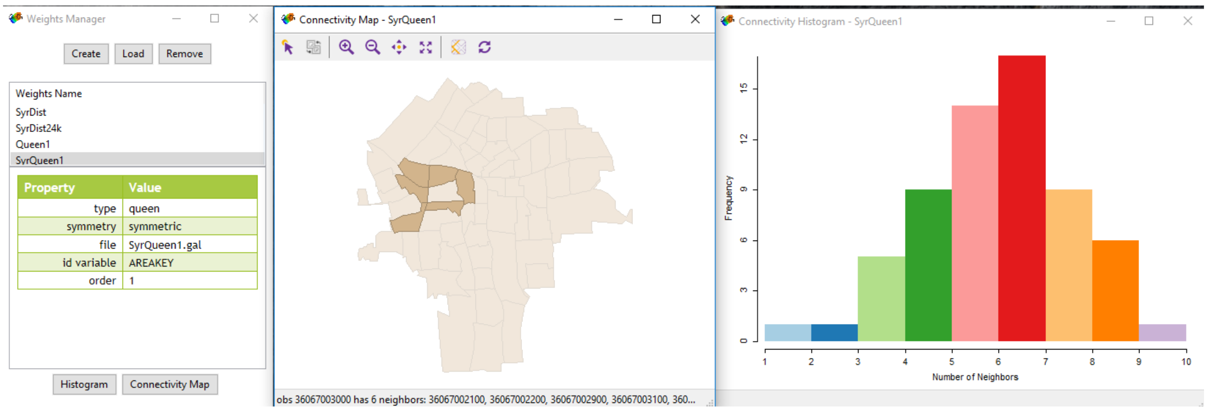


Figure 16: Connectivity map and histogram, Queen lag-1 neigbours

Q23 : *What is the most common number of neighbours using the Queen lag-1 neighbours?* •

6 Assessing Global Spatial Autocorrelation

Here we evaluate whether the rates of leukemia across the study area may be spatially auto-correlated, without considering any predictors.

Task 20 : Use `Space | Univariate Moran's I`, with **Z** as the variable, to produce a global Moran's I plot. Do this for all the weighting schemes you defined. •

See Figure 17.

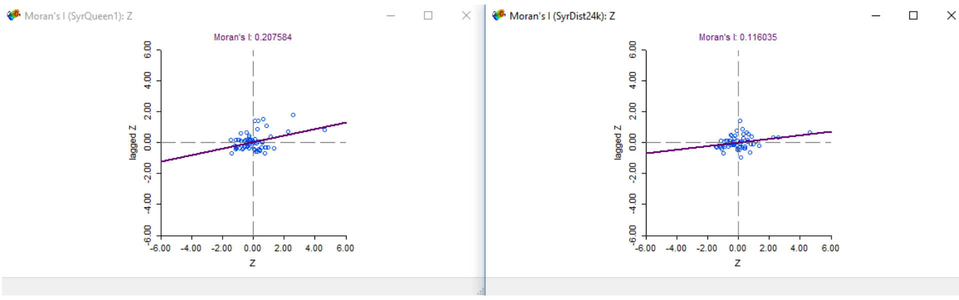


Figure 17: Global Moran's I

Q24 : *Do the weighting schemes all give the same value of global Moran's I? If not, which implies stronger spatial correlation? Why?* •

Task 21 : Open two themed maps: decile (10-quantile) of `Z` (incidence) and `PEXPOSURE` (exposure).

In the Moran's I scatterplot, click on the highest positive `Z` (incidence) and highest weighted lag `Z`. •

See Figure 18.

Q25 : *Where is this tract located? Does it also have a high exposure? Do its neighbours have high incidences? Do they have high exposures?* •

7 Assessing Local Spatial Autocorrelation

Task 22 : Examine where in the map are the hotspots of local autocorrelation. •

First, make sure that your desired Weights file is set as the default, i.e., highlighted in the Weights Manager window.

Task 23 : Use `Space | Univariate Local Moran's I`, with **Z** as the variable, to produce a local Moran's I plot. Do this for all the weighting schemes you defined. Generate two output windows: the Significance Map and the Cluster Map. •

The Significance map shows where there are leukemia values that are statistically significantly higher **or** lower than the neighboring values would have predicted. With the Cluster Map, you can see where the higher-than-expected and lower-then-expected values vary.

See Figure 19.

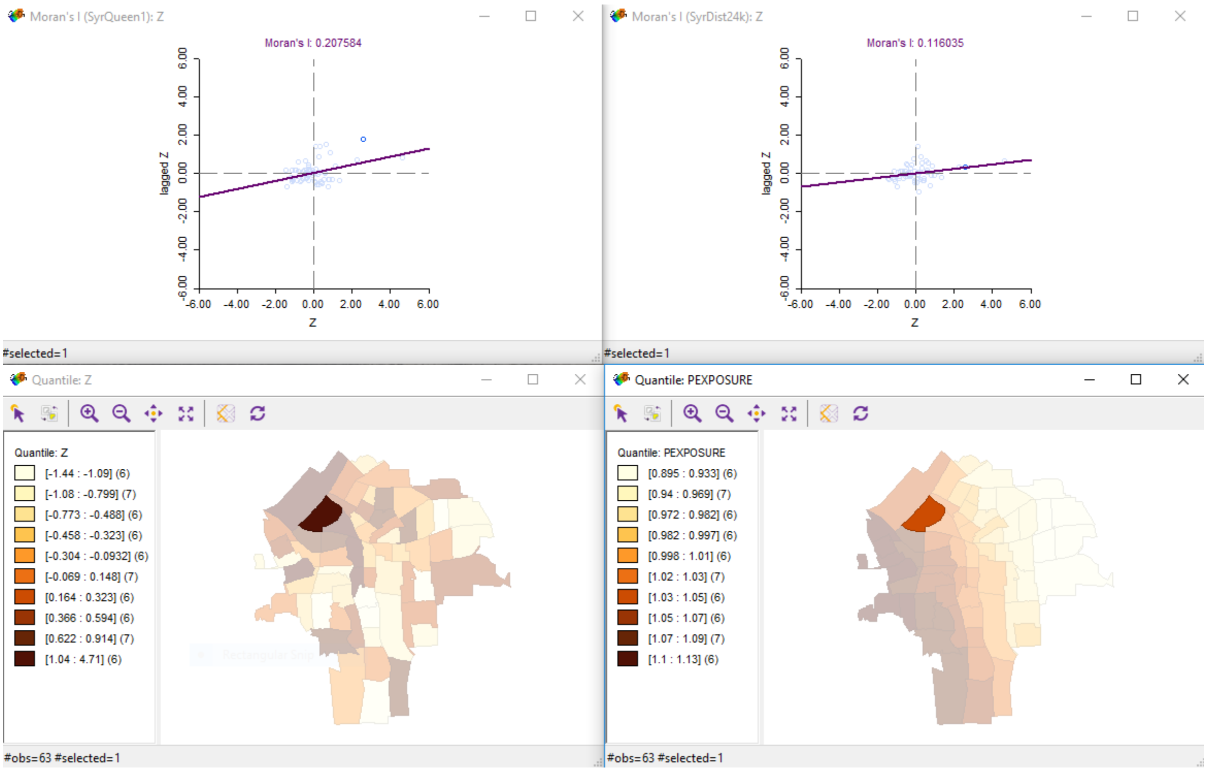


Figure 18: Global Moran's I

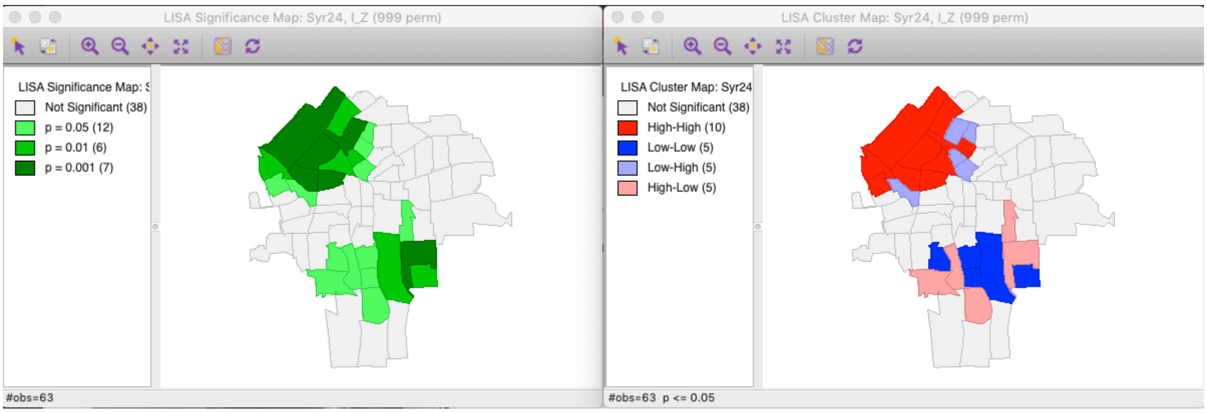


Figure 19: Local Moran's I signficance and clusters

Q26 : *Which areas of the city are clusters of high leukemia incidence? Are there any tracts that have high incidence, but are surrounded by tracts with low incidence?* •

Another way to find hot spots is with Geary's G or G*; if you want you can experiment with these.

8 Spatial Regression

Here we try to find the covariates ("predictors") correlated (which maybe cause) leukemia. Of course, we can do this non-spatially, i.e., all in attribute space, not taking spatial relations into account.

TASK 24 : Compute a multivariate linear regression model of leukemia incidence (response) as predicted by the three possible causitive factors (predictors). This is with the `Regression` menu item. Select `Z` as the dependent variable, and `PCTOWNHOME`, `PCTAGE65P`, and `PEXPOSURE` as the covariates. This is the Classic linear model, i.e., not taking spatial correlation into account. •

Non-spatial linear model: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

Figure 20 shows how to specify the regression; Figure 21 shows the results.

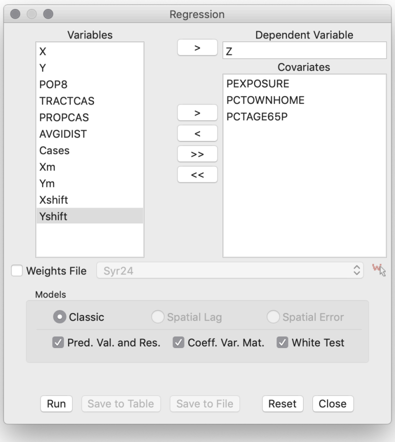
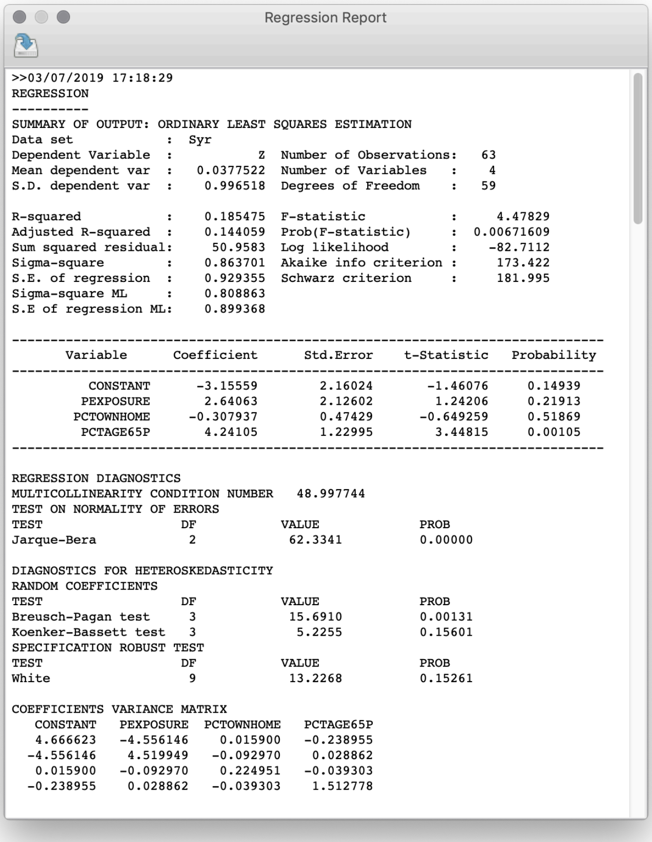


Figure 20: Specifying a "classic" linear regression

Q27 : *What is the adjusted R^2 of this model? What are the signs of the slopes for each predictor? What is the interpretation? Which (if any) predictors are significantly different from zero?* •

The model summary shows many problems with the linear model:

1. The **multicolilinearity** (or multiple) **condition number** represents the sensitivity of the model to small changes in the design matrix, i.e., the values of the covariables. A high value (often > 30) indicates high colinearity in one or more predictors; here we see that is the case.




```
>>03/07/2019 17:18:29
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set             :  Syr
Dependent Variable   :           Z  Number of Observations:   63
Mean dependent var   :   0.0377522  Number of Variables   :    4
S.D. dependent var   :    0.996518  Degrees of Freedom    :   59

R-squared            :    0.185475  F-statistic           :     4.47829
Adjusted R-squared   :    0.144059  Prob(F-statistic)     :  0.00671609
Sum squared residual:      50.9583  Log likelihood        :    -82.7112
Sigma-square         :    0.863701  Akaike info criterion :     173.422
S.E. of regression   :    0.929355  Schwarz criterion     :     181.995
Sigma-square ML      :    0.808863
S.E of regression ML:     0.899368

-----------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT      -3.15559        2.16024      -1.46076      0.14939
         PEXPOSURE       2.64063        2.12602       1.24206      0.21913
        PCTOWNHOME     -0.307937        0.47429     -0.649259      0.51869
         PCTAGE65P       4.24105        1.22995       3.44815      0.00105
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   48.997744
TEST ON NORMALITY OF ERRORS
TEST                  DF           VALUE             PROB
Jarque-Bera            2            62.3341          0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF           VALUE             PROB
Breusch-Pagan test     3            15.6910          0.00131
Koenker-Bassett test   3             5.2255          0.15601
SPECIFICATION ROBUST TEST
TEST                  DF           VALUE             PROB
White                  9            13.2268          0.15261

COEFFICIENTS VARIANCE MATRIX
   CONSTANT   PEXPOSURE  PCTOWNHOME   PCTAGE65P
   4.666623   -4.556146    0.015900   -0.238955
  -4.556146    4.519949   -0.092970    0.028862
   0.015900   -0.092970    0.224951   -0.039303
  -0.238955    0.028862   -0.039303    1.512778
```

Figure 21: Multiple linear regression results

2. The Jarque–Bera test is whether the residuals have the skewness and kurtosis matching a normal distribution. Here we see a high value, quite unlikely to be normal.

However we will not fix up this model, we proceed to compare it to models which do take into account spatial correlation.

8.1 Spatial Error model

The first model with a spatial component we will consider is the **spatial error model**. This allows **resduals** of the linear model to be spatially-correlated, and quantifies to what extent they are included in the model.

This typically occurs when there is some spatially-correlated covariate that (1) affects the response and (2) we do not know, or maybe even suspect – otherwise we would identify it, measure it, and include in the linear model. However, we may suspect a factor that we have not, or can not, measure, and this factor has spatial correlation.

For example, this database does not report the proportion of different ethnic groups, nor of different occupational groups (factory workers, office workers, service workers . . .). These may be (1) related to leukemia (genetic susceptibility, occupational exposure), (2) spatially-correlated. If such factors influence leukemia, they will be represented in the residuals, and thus the spatial error model will be provably better than the feature space-only model.

The spatial error model is:

- formula: $\mathbf{Y} = \mathbf{X}\beta + \lambda\mathbf{W}\mathbf{u} + \varepsilon$
- $\mathbf{W}$ is a matrix representing the spatial structure (e.g., neighbour weights)
- $\mathbf{u} = (\mathbf{Y} - \mathbf{X}\beta)$ are the spatially-correlated **residuals**
- λ is the strength of autoregression of the errors
- ε is the independent error (not autoregressive)

Task 25 : Compute a multivariate linear regression model of leukemia incidence (response) as predicted by the three possible causitive factors (predictors). This time (1) select a Weights file (one you created above), and then you can specify the SAR **Spatial Error** linear model. This takes spatial correlation of the **linear model residuals** into account, considering the values of the model **residual** in each tract's neighbourhood, as defined by the weights. •

Figure 22 shows the results.

Q28 : *What is the pseudo-R^2 of this model?*[6]*. Is it higher or lower than that for the feature-space only model? Is this expected? How can it be explained?* •

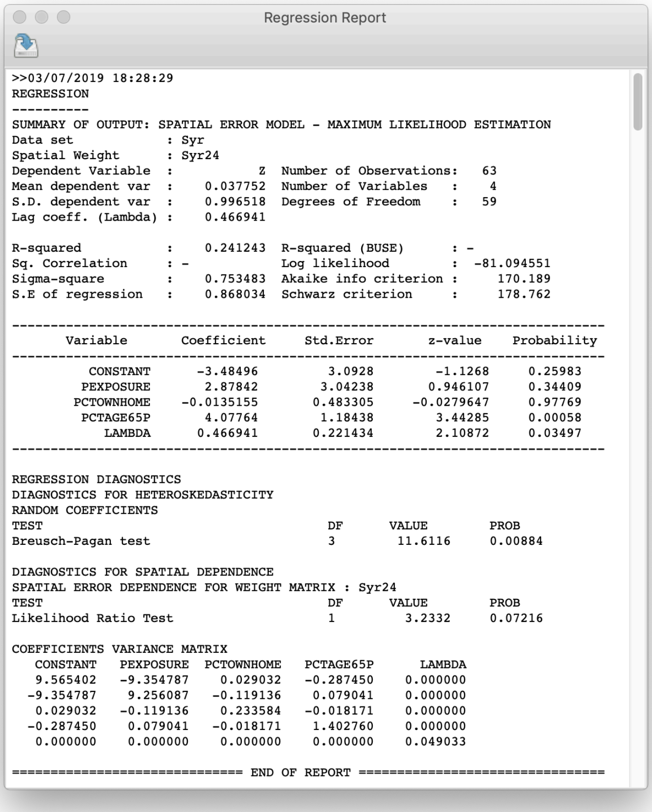
Q29 : *What are the signs of the slopes for each predictor? What is the interpretation? Which (if any) predictors are significantly different from zero? What changes in this model compared to the feature-space multiple regression? I.e., which predictors become more or less important and/or significant?* •

Q30 : *What is the strength of the autocorrelation parameter λ?* •

The **likelihood ratio test** gives the probability that the SAR spatial error model is *not* better than the feature-space-only multiple regression.

Q31 : *What is the probability that the SAR spatial error model is* not *better than the feature-space-only multiple regression? What does this imply about the possible causes of leukemia?* •

[6] This is not, strictly speaking, an R^2 but does express the proportion of variance explained



```
>>03/07/2019 18:28:29
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Syr
Spatial Weight      : Syr24
Dependent Variable  :           Z  Number of Observations:   63
Mean dependent var  :    0.037752  Number of Variables   :    4
S.D. dependent var  :    0.996518  Degrees of Freedom    :   59
Lag coeff. (Lambda) :    0.466941

R-squared           :    0.241243  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :  -81.094551
Sigma-square        :    0.753483  Akaike info criterion :     170.189
S.E of regression   :    0.868034  Schwarz criterion     :     178.762

-----------------------------------------------------------------------------
       Variable       Coefficient      Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT      -3.48496         3.0928        -1.1268       0.25983
         PEXPOSURE       2.87842        3.04238       0.946107       0.34409
        PCTOWNHOME    -0.0135155       0.483305     -0.0279647       0.97769
         PCTAGE65P       4.07764        1.18438        3.44285       0.00058
            LAMBDA      0.466941       0.221434        2.10872       0.03497
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                     DF      VALUE        PROB
Breusch-Pagan test                       3       11.6116      0.00884

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : Syr24
TEST                                     DF      VALUE        PROB
Likelihood Ratio Test                    1        3.2332      0.07216

COEFFICIENTS VARIANCE MATRIX
   CONSTANT   PEXPOSURE  PCTOWNHOME   PCTAGE65P      LAMBDA
   9.565402   -9.354787    0.029032   -0.287450    0.000000
  -9.354787    9.256087   -0.119136    0.079041    0.000000
   0.029032   -0.119136    0.233584   -0.018171    0.000000
  -0.287450    0.079041   -0.018171    1.402760    0.000000
   0.000000    0.000000    0.000000    0.000000    0.049033

============================== END OF REPORT ================================
```

Figure 22: SAR spatial error model regression results

8.2 Spatial Lag model

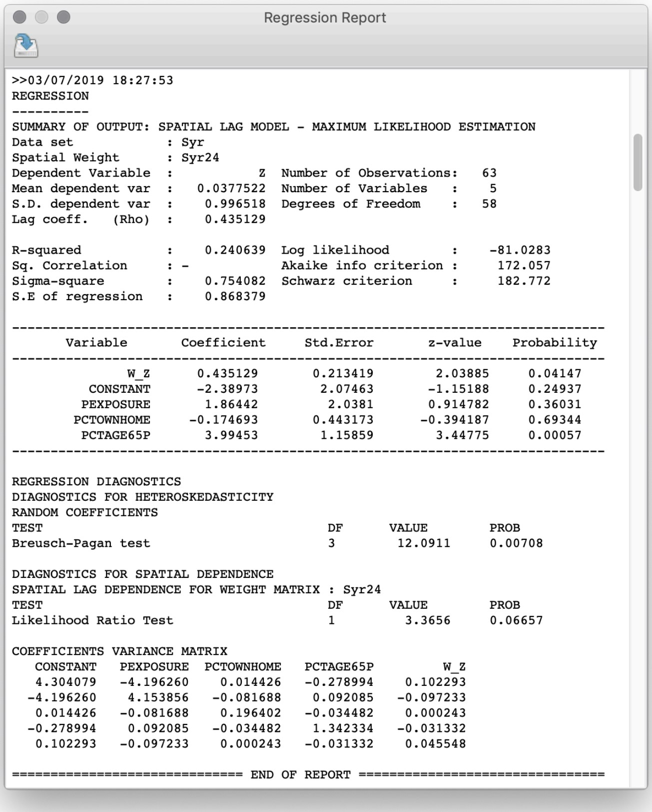
Another possible effect of spatial autocorrelation is in the response, that is, the values of the response in a tract's neighbours directly influence the response in the tract, after taking into account the feature-space prediction. This measures "contagion", which seems unlikely for human leukemia[7], however we still evaluate this.

Task 26 : Compute a multivariate linear regression model of leukemia incidence (response) as predicted by the three possible causitive factors (predictors). This time (1) select a Weights file (one you created above), and then you can specify the SAR **Spatial Lag** linear model. This takes spatial correlation into account, considering the values of the **response** variable in each tract's neighbourhood, as defined by the weights. •

The spatial lag model is: $\mathbf{Y} = \rho\mathbf{WY} + \mathbf{X}\beta + \varepsilon$, where ρ is the strength of autoregression of the response; this multiplies the weights matrix times the response $\mathbf{WY}$ on the right-hand (predictor) side of the equation.

Figure 23 shows the results.

[7] although quite likely for feline leukemia, if infected cats travel across tract boundaries



```
Regression Report
>>03/07/2019 18:27:53
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Syr
Spatial Weight      : Syr24
Dependent Variable  :           Z  Number of Observations:   63
Mean dependent var  :   0.0377522  Number of Variables   :    5
S.D. dependent var  :    0.996518  Degrees of Freedom    :   58
Lag coeff.   (Rho)  :    0.435129

R-squared           :    0.240639  Log likelihood        :    -81.0283
Sq. Correlation     : -            Akaike info criterion :     172.057
Sigma-square        :    0.754082  Schwarz criterion     :     182.772
S.E of regression   :    0.868379

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
                W_Z      0.435129      0.213419       2.03885     0.04147
           CONSTANT      -2.38973       2.07463      -1.15188     0.24937
          PEXPOSURE       1.86442        2.0381      0.914782     0.36031
         PCTOWNHOME     -0.174693      0.443173     -0.394187     0.69344
          PCTAGE65P       3.99453       1.15859       3.44775     0.00057
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                        DF      VALUE        PROB
Breusch-Pagan test                          3       12.0911      0.00708

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : Syr24
TEST                                        DF      VALUE        PROB
Likelihood Ratio Test                       1        3.3656      0.06657

COEFFICIENTS VARIANCE MATRIX
   CONSTANT   PEXPOSURE  PCTOWNHOME   PCTAGE65P         W_Z
   4.304079   -4.196260    0.014426   -0.278994    0.102293
  -4.196260    4.153856   -0.081688    0.092085   -0.097233
   0.014426   -0.081688    0.196402   -0.034482    0.000243
  -0.278994    0.092085   -0.034482    1.342334   -0.031332
   0.102293   -0.097233    0.000243   -0.031332    0.045548

============================== END OF REPORT ================================
```

Figure 23: SAR spatial lag model regression results

Q32 : *Is the lag coefficient ρ significant in the regression? What is the probability that the SAR spatial lag model is* not *better than the feature-space-only multiple regression? What does this imply about the possible causes of leukemia?* •

9 Finishing with GeoDa

GeoDa provides several opportunities to save images or export newly derived data in tabular form. If while using GeoDa you have derived values and added variables to the attribute table of your shapefile, you will be prompted to Save these as you Exit. Otherwise, the program can simply be closed.

References

[1] Luc Anselin, Xun Li, and Julia Koschinsky. GeoDa, from the desktop to an ecosystem for exploring spatial data. *Geographical Analysis*, 54(3):439–466, November 2021. ISSN 1538-4632. doi: 10.1111/gean.12311.

[2] Roger S. Bivand, Edzer J. Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer, 2nd edition, 2013. ISBN 978-1-4614-7617-7; 978-1-4614-7618-4 (e-book). URL http://www.asdar-book.org/.

[3] L. A. Waller and C. A. Gotway. *Applied spatial statistics for public health data*. Wiley-Interscience, Hoboken, N.J., 2004.