Principal Components Analysis with Spatial Data

A SpaceStat Software Tutorial

Principal Investigators: Pierre Goovaerts and Geoffrey Jacquez . SpaceStat Team: Eve Do, Pierre Goovaerts, Sue Hinton, Geoff Jacquez, Andy Kaufmann, Sharon Matthews, Susan Maxwell, Kristin Michael, Yanna Pallicaris, Jawaid Rasul, and Robert Rommel.

SpaceStat was supported by grant CA92669 from the **National Cancer Institute** (NCI) and grant ES10220 from the **National Institute for Environmental Health Sciences** (NIEHS) to BioMedware, Inc. The software and help contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCI or NIEHS.

BioMedware
Geospatial Research and Software

MATERIALS

PCA Tutorial.SPT SpaceStat project

ESTIMATED TIME

20 minutes

OBJECTIVE

This tutorial will undertake a Principal Components Analysis (PCA) of geographically distributed data in SpaceStat. The data are homeownership and socioeconomic data for the state of Michigan at the Census Tract level. You will undertake a PCA, project the results back into the geography, and interpret them.

WHY PCA?

PCA will allow you to construct orthogonal (e.g. independent) synthetic variables that are linear combinations of the predictors. This is useful for (1) identifying combinations of variables that best explain the variance in the data (2) reducing *dimensionality* so that fewer synthetic variables explain the observed variance, (3) defining suites of variables that may be functionally related. The synthetic variables are more commonly known as principal components, which are constructed as linear combinations of the original variables. The term "synthetic variable" makes explicit the idea that a common mechanism may be driving co-variation among the variables that load on the synthetic variables. We'll use "synthetic variable" and "principal component" interchangeably in this tutorial. Because the synthetic variables resulting from the PCA are orthogonal, they may be used as predictors in regression models without having to worry about colinearity. Hence when you have problems with many potential predictors PCA may be used both to reduce dimensionality and colinearity, resulting in more robust and simpler regression models. One problem with PCA is that, because they are constructed from observed variables, it can be difficult to interpret the synthetic variables.

ASSUMPTIONS AND CAVEATS

PCA assumes the variables analyzed are jointly normally distributed, and it is sensitive to the relative scaling of the original variables. When these assumptions are violated the resulting principal components may not be independent. For this reason variables to be analyzed are often centered to 0

and scaled by their variance, using a z-score transformation. If the variables are highly non-normal one can use the Normal score transform in SpaceStat.

In this example we skip the variable transformation step, but we do inspect the correlations among the principal components and find them to be independent. Usually you will want to transform the variables before analysis so that they are normally distributed.

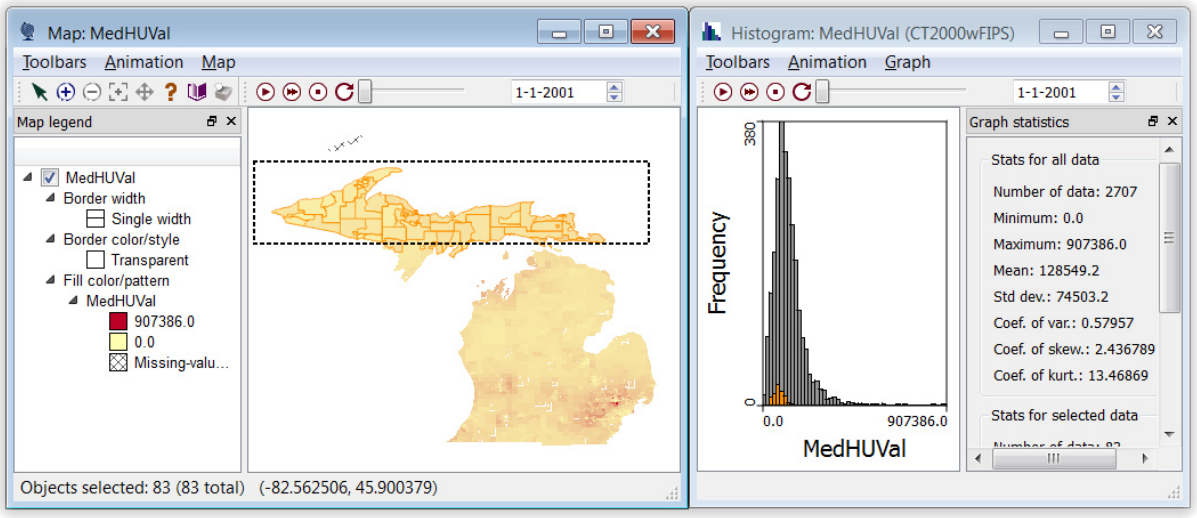
STEP 1: LOAD THE PROJECT

To open this project, **go to File -> Open**, and then **browse to "PCA Tutorial"**. **Click the "OK" button** and the project will load. Look at the Datasets listed in the Data view. For this exercise, you are interested in constructing predictive models of median household unit value ("MedHUVal"). You have census variables describing socio-economic characteristics of the population in each census tract (e.g. Occupation, ethnicity, education, employment, and housing). You are concerned that some of these variables might be correlated; resulting in problems of multi-colinearity should they be incorporated into the regression model.

STEP 2: MAPS AND DESCRIPTIVE STATISTICS

For this exercise focus your attention on the ethnicity variables PCT_Asn, PCT_Blk, PCT_Hisp and PCT_Wht, describing the percentage Asian, Black, Hispanic and White. Construct histograms and maps for each of these potential predictors, as well as MedHUVal. To create a map, **click on the Map button** on the toolbar (the one that looks like a globe) and **select your variable as the fill dataset**. To create a histogram, **click on the Histogram button** on the toolbar (the one that looks like a histogram) and **select your variable as the dataset**. Are these data time-dynamic?

You can create a map-histogram pair for each of the variables by arranging the respective map and histogram next to one another. Now use brush selection on the maps and on the histograms to obtain an understanding of geographic variability in each of the variables.



Brush-selecting on the map-histogram pair for MedHUVal

Use the maps and histograms to answer the following questions:

What is the mean of the median housing prices in the upper peninsula of Michigan?

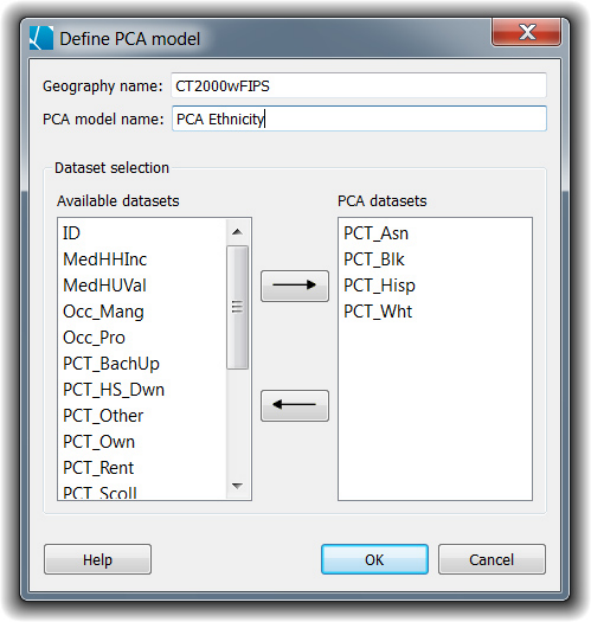
Where are the largest percentages of non-white populations found?

Inspect the histograms. Are any of the distributions bimodal? What does a bimodal distribution of an ethnic population mean? What would these distributions look like should there be no racial segregation?

STEP 3: PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis will transform a number of possibly correlated variables into a group of uncorrelated variables. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much remaining variability as possible.

Go to Methods -> Principal Components. Select "CT2000wFIPS" as your geography. Then **click the "Create" button** to begin setting up the analysis. In this case, **select PCT_Asn, PCT_Black, PCT_Hisp, PCT_Wht** from the Available datasets. **Click the arrow that points to the right** to add each variable to the "PCA variables" box. **Change the PCA model name to a descriptive title** such as "PCA Ethnicity". Click the OK button to finalize your model.



PCA model settings

Click the 'Next' button to proceed to the PCA settings tab. You can name the output folder where the Principal Components will appear and you have the ability to specify the time range over which the PCA will be run. This will create a PCA analysis with time-dynamic synthetic variables for each time period in which at least one of the variables values change. Our dataset is static, so only one PCA will be conducted. PCA may be calculated from the correlation matrix or from the variance-covariance matrix. The results should be almost identical. **Use the correlation matrix** in this case. **Press the 'Run' button** to start the analysis.

STEP 4: CORRELATION COEFFICIENTS

After the run is complete the principal components will appear in the 'Data view' under the folder name you chose. Maps of the first two components will be created, as well as various tables in the Log. The table of correlation coefficients is shown below:

Variable	PCT_Asn	PCT_Blk	PCT_Hisp	PCT_Wht
PCT_Asn	1.0	-0.1086	-0.00436	-0.02795
PCT_Blk	-0.1086	1.0	-0.08354	-0.96235
PCT_Hisp	-0.00436	-0.08354	1.0	-0.14417
PCT_Wht	-0.02795	-0.96235	-0.14417	1.0

We can use this table to interpret correlations among the variables. There is a strong negative correlation between PCT_Wht and PCT_Blk, meaning as the white population increases, the black population decreases. There's a weak positive correlation between PCT_Wht and PCT_Asn, and no correlation between PCT_Wht and PCT_Hisp. There is a weak negative correlation between PCT_Asn and PCT_Blk, and no correlation between PCT_Asn and PCT_Hisp. Finally, there is a weak negative correlation between PCT_Blk and PCT_Hisp.

STEP 5: PRINCIPAL COMPONENTS

The next two tables in the log show the proportion of variance explained by the 4 principal components (this table is titled "Principal Components"), followed by the correlations between each principal component and the original variables (this table is titled "Eigenvectors of the Pearson correlation matrix".

Index	Eigenvalue	Variance proportion	Cumulative variance
1	1.967635	0.491909	0.491909
2	1.031624	0.257906	0.749815
3	0.998554	0.249639	0.999454
4	0.002186	0.000546	1.0

The first row is for the first principal component, the second is for the second principal component and so on. There are four Eigenvalues presented, one for each principal component. These define the new orthogonal data space. The third column is the proportion of the variance the original variables that is explained by that principal component. For example, the first principal component (PC) explains 49% of the overall variance. The last column is the cumulative variance explained. Together, the first two synthetic variables explain almost 75% of the variance. But how are they constructed?

Variable	PC_1	PC_2	PC_3	PC_4
PCT_Asn	0.058904	-0.46713	-0.87701	-0.09574
PCT_Blk	-0.70463	0.139023	-0.04558	-0.69433
PCT_Hisp	-0.04458	-0.86446	0.474805	-0.15901
PCT_Wht	0.705721	0.123189	0.057688	-0.69531

The above table shows the correlations between the variables and each of the principal components. You can use this to obtain an understanding of how the PC's are constructed. For example, the first principal component is correlated 0.7057 with PCT_Wht and -0.7046 with PCT_Blk. It thus has something to do with the presence of Whites and the absence of Blacks. The second PC is loaded most heavily by the absence of Hispanics and Asians.

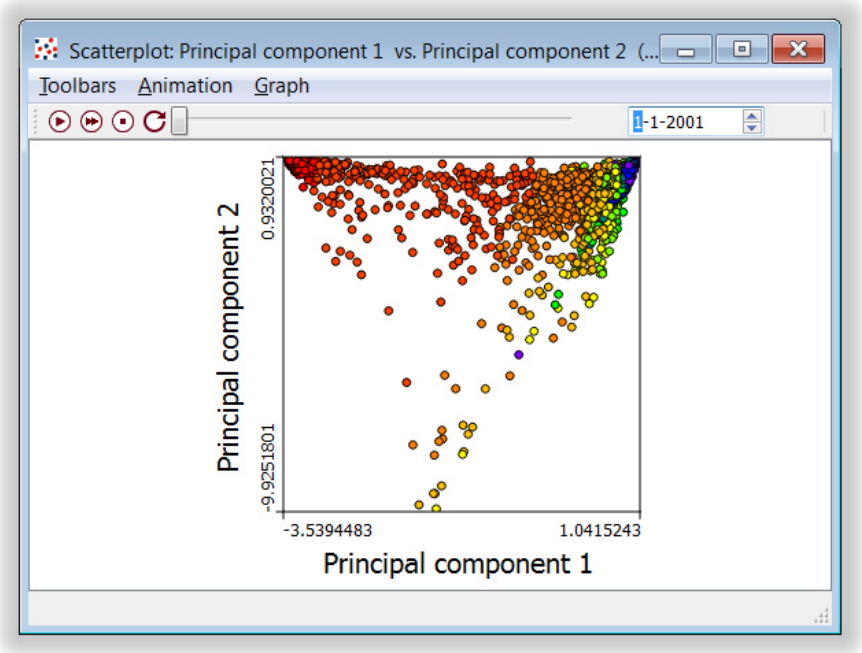
STEP 6: SCATTERPLOTS OF THE PRINCIPAL COMPONENTS

You can create a scatterplot of the first two components. **Click the 'Scatterplot' button** on the toolbar. **Select 'Principal Component 1' from the X dataset dropdown** and **select 'Principal Component 2' from the Y dataset dropdown**. **Click the 'OK' button** to create the scatterplot.

This graphic often displays groupings within the data. A good way to tease out various structures within the data is to color the observations based on the values of one of the variables.

Select Graph->Properties at the top of the scatterplot. **Click the 'Point fill properties' tab** and **select PCT_Blk** from the Datasets. **Choose Classified from the Color mode dropdown** and leave the

defaults for Number of Classes and Classification Method (10, Quantiles). Then **select the Reverse Spectrum Color palette** to clearly distinguish the different values. We suggest this palette because the colors are very saturated and high values of PCT_Blk will show in red. Click the 'OK' button to modify the scatterplot.



Scatterplot of PC 1 and 2 with the fill color set to PCT_Blk

The bands of color reveal the internal structure of the data. The blue dots are the tracts with the lowest percentage of black population and red dots are the tracts with the highest percent of black population. You can brush select on the histograms of ethnicity, and then inspect the PC scatterplot, to obtain a better understanding of how the original variables load on the PC's.

STEP 7: PRINCIPAL COMPONENTS MAPS

SpaceStat creates maps for the first two principal components. Look at these maps alongside histograms of the PC's (you will need to create histograms for the PC's by pressing the Histogram button and choosing the PC as the dataset). The first PC captures a good deal of the variability in PCT_Wht and PCT_Blk, and we would expect there to be similarities between maps of Principal component 1 and these two variables. Brush select on the maps and histograms for Principal component 1, PCT_Wht and PCT_Blk to confirm that their spatial patterns of variation are similar. You may need to zoom in on the map to see the variation around locations such as Detroit.

THINK ABOUT YOUR RESULTS

You now understand how you can use PCA in dimensionality reduction and to construct independent synthetic variables. Recall from Step 1 that you are interested in constructing predictive models of median household unit value ("MedHUVal"). *How might you use the synthetic variables in such a regression?*