Peer Areas Cookbook

I- Data Needed: When investigating small areas all of the data need to be at the same level of geography, for example, if the analysis is at the ZIP Code level then the following data need to be available for each ZIP Code.

A- Population estimates for each area (by age and sex if possible).

B- Demographic variables for each area (Median income, percent in poverty, population over 65 years of age).

C- Data on measures of health status for each area (case counts for cancer or heart disease). These are found in vital records, hospitalization records, national surveys (BRFSS) and other sources.

II- Choosing the Demographic variables

To begin we need to make some decisions on which variables we will use to determine when our areas are demographically similar. We are interested in demographic variables that have a known association with health outcomes but are not health outcomes themselves. We begin with ZIP Code level information on a variety of demographic variables such as income, education, employment, and poverty obtained from the 2000 US Census. Many of the variables are highly correlated so we need to be careful in how we select those that appear in our model since the variables we choose need to be independent of one another. We used Factor Analysis to help us identify important variables since the factors created by this procedure are linearly independent. We selected the demographic variable with the highest factor loading to represent the factor established by the Factor Analysis procedure. That is we did not use the factors themselves but instead chose a single variable to represent the factor. In this way we reduced the number of demographic variables to the following five for Utah; percent of population 65 and over, percent of population with a college degree, percent of children in poverty, percent of population that are Hispanic, and percentage of population living in owner occupied housing.

III- Defining Demographic Distance

Create the Variance /Covariance matrix for the demographic variables, this is done with statistical computing software such as SAS, Stata, SPSS, or with macros within Excel.

Let $\mathbf{S}$ be the variance/covariance matrix of the n demographic variables. The variance for each variable is on the diagonal of $\mathbf{S}$ and the covariances are the off diagonal entries. The inverse of this matrix, $\mathbf{S}^{-1}$, is used in the calculation for the distance between each area and all other areas, we do this in order to weight the distance by the inverse of the variance/covariance in the variables. With this method the variables with a small variance will receive high weights and variables with a large variance will receive low weights.

This will reflect the precision of measurement within the demographic variable. This also gives a higher weight to the demographic variables that are measured with greater precision.

Each area is described by an ordered n-tuple of values from the "n" selected demographic variables. If we let $A_i = (x_{i1},x_{i2},x_{i3},\ldots,x_{in})$ represent the values for the n demographic variables in area i and $A_j = (x_{j1},x_{j2},x_{j3},\ldots,x_{jn})$ represent the values for the n demographic variables in area j then the "statistical distance" between each area is defined to be:

$$\mathbf{d(Ai,Aj)= ([Ai-Aj]S^{-1}[Ai-Aj]')^{1/2}}$$

IV- Using nearest neighbors to smooth health status outcome data.

We want to create an algorithm that will capture information from a few statistically close neighbors and weight these areas more heavily than information from dissimilar areas. Each area receives a weight according to its distance from the index area using the following function: $\exp(-a*d_{ij}^2)$, where d_{ij} = distance from the index area "i" to the neighbor area "j" calculated previously and "a" is a positive number that determines the amount of smoothing, values near zero produce high smoothing and values near one produce low smoothing . We chose this weighting scheme because we did not want to arbitrarily select a fixed number of neighbors for smoothing. This function will give the index area a full weight of one and apply progressively diminishing weights to areas as the demographic distance increases.
We begin with the crude rate for each area, this is the ratio of the number of events to the population for each small area. The smoothed rate is then calculated as the sum of weighted rates relative to the sum of the weights and is referred to as a weighted average.

The smoothed rate for area "i" is now calculated as:

$$[\sum_j R_j * \exp(-a*d_{ij}^2)] / [\sum_j \exp(-a*d_{ij}^2)]$$

where j = {1,2,3,..,i,…,K} the set of peers, and Rj = rate in area j

All areas are involved in the smoothing process but only the demographically close areas have any meaningful influence on the smoothed rate.

We have recently included an additional smoothing algorithm that is based on regression, the results appear in the Lest Squares Smoothing page. This procedure will be described in more detail in a later publication. It is important to note that in the Least Squares procedure the analyst chooses the number of nearest neighbors used for smoothing.

III- Assessing Smoothness

The smoothed rate is an additional rate that will be presented to local health officials and community leaders. The principal question will be “how do we interpret this?” so care needs to be taken when explaining what the rate means, how it is different than the crude rate, and when it should be used. The smoothed rate is most useful in areas with small populations or for low probability events and should be used as a reference to guide decisions when the crude or age specific rate is unstable.
The smoothed rate for an area without close neighbors will likely retain most of its own crude rate whereas the smoothed rate for an index area with many close neighbors will be more heavily influenced by the rates in those areas. It could be the case that an area with an unusually high or low rate has no near neighbors. In this case there may not be much modifying influence from the peer group and so the smoothed rate remains close to the crude rate.

The Intraclass Correlation Coefficient measures reliability/consistency across different raters. In our case we compare the variance between small areas to the variance within small areas across years. The ICC will measure how similar (smooth) the rates are across the years. Higher values correspond to smaller variability across the years. We should note that the ICC is an overall measure, it will be the same value for each small area.

In the case of competing smoothing algorithms we can create a Sum of Squared differences between the smoothed and crude rate to see which estimate is most similar to the crude. You may be familiar with this from regression as this is the fitted line that has the smallest sum of squared differences from the given data.

The “Peer Area Add-in” and the “Peer Area Spreadsheet” will help you get familiar with the smoothing algorithms. You can also enter your own data into an Excel spreadsheet and use the “Peer Area Add-in” to create smoothed rates for your own small areas.

For further reading on exponential functions you can reference the paper:
Poggio, Tomaso; Smale, Steve
The Mathematics of Learning: Dealing with Data. Notices Amer. Math. Soc. 50 (2003), no. 5, 537—544