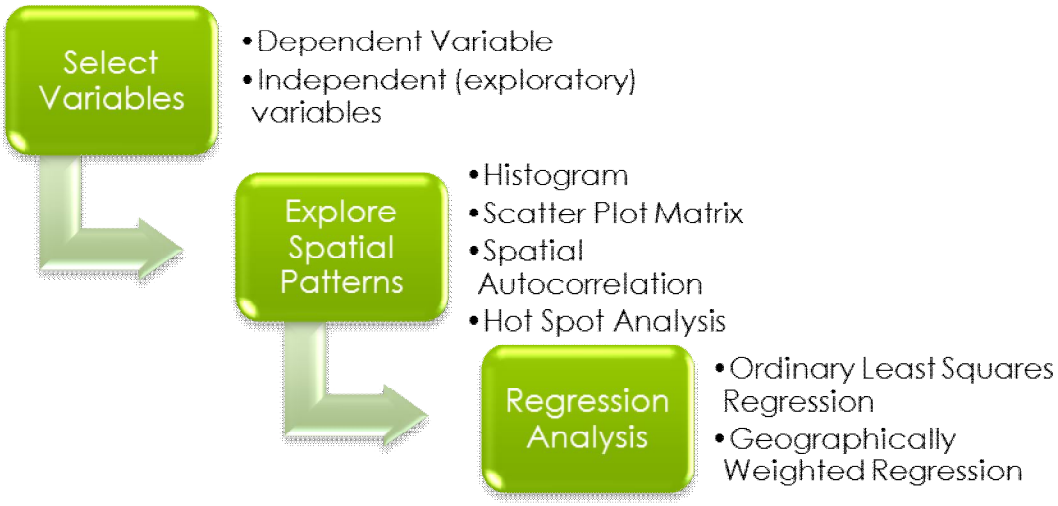
Exploratory Spatial Data Analysis (ESDA)

VANGHR's method of ESDA follows a typical geospatial framework of selecting variables, exploring spatial patterns, and regression analysis. The primary software tools used for ESDA are Arizona State University's GeoDa and Esri's ArcGIS (Spatial Statistics Toolbox). VANGHR selects the appropriate dependent and independent variables to examine. A combination of histograms and scatter plot matrices are used in visualizing the distribution of the data in space. Spatial autocorrelation is used to examine the nature of the spatial pattern of observation points (clustered vs. random vs. dispersed). Hot spot analysis is used to explore the clustering patterns of selected variables. Spatial autocorrelation is also used to explore the association between two variables. The next step in the process is to run an ordinary least square (OLS) regression. Based on the results of the OLS regression, VANGHR determines whether a geographically weighted regression analysis is necessary.

Figure 1: VANGHR Analytic Process



Spatial Autocorrelation in Bivariate Analysis

Using GeoDa's multivariate LISA (local indicator of spatial association) analysis, the correlation between the two variables is examined. The analysis produces both an overall Moran's I for the study area and a LISA value for each feature. Only those features with a statistically significant ($p = 0.05$) LISA value are mapped. Default permutation values are accepted. A graphic map output of the nature of the association is also produced. This "cluster" map indicates the

classification of the association of the two variables by a census tract as “High/High,” “Low/Low,” “Low/High,” and “High/Low.”

In GeoDa, a new project is opened with a polygon shapefile containing the appropriate variables. A weights matrix is then created for the polygon shapefile. The “Multivariate LISA” tool was selected from the “Space Toolbar” with the following inputs:

- 1st Variable (Y): Dependent Variable
- 2nd Variable (X): Independent Variable
- Choose to open “Cluster Map” and “Moran Scatter Plot”

The results of the analysis are saved by right-clicking the map output, choosing “Save Results,” and selecting “LISA indicies,” “Clusters,” and “Significances.” The shapefile table is then saved to a new shapefile. This new shapefile is then opened in ArcGIS and the cluster field is symbolized to indicate the association between the two variables at the local level.

Geospatial Regression Methods in Multivariate Analysis

Geographically Weighted Regression (GWR) is a technique for exploratory data analysis that provides estimates of regression coefficients for each geographical location, based on a weighting of other observations near that location (Mitchell, 2005). The basic assumption is that observations exhibit spatial dependency. This has its root from the first law of geography by Tobler which says that, "Everything is related to everything else, but near things are more related than distant things," (Tobler, 1970).

Ordinary Least Squares (OLS) regression serves as a starting point to build a well specified GWR and guides the researcher to select the key explanatory variables. Upon completion of the OLS, verification of the six tests for OLS is required before proceeding to GWR. Below are the six tests:

1. **Coefficients have the expected signs** – test to determine if the signs associated with each variable are appropriate. For example, in trying to model stroke hospitalization, when an analyst sees that the population 65 and over is negatively associated with stroke hospitalization, further investigation of the model is likely needed, since age (65 and over) is one of the risk factors for stroke hospitalization.
2. **AIC** (Akaike Information Criterion) – test to measure how the model performed and compare different regression models. The model with the lower AIC is held to be better. This indicator also addresses the benefit of moving from OLS regression (Global) to GWR (local regression).
3. **Variance Inflation Factor (VIF)** - test to verify if two or more variables are telling the same story (colinearity). The rule is that any variable with greater than 7.5 VIF should be removed.
4. **Jarque-Bera Statistic Test** – test to verify that the residuals are normally distributed and for model bias. Since the null hypothesis is that the residuals are normally distributed, test to make sure that this is not statistically significant. If this test has a value < or = 0.05, this means that the model is bias and cannot be trusted. Run spatial autocorrelation (Moran’s Index) to make sure the residuals are random. If the residuals cluster, that will be an indication that a critical explanatory variable is missing from our model.
5. **Adjusted R-Square** – the adjusted R-Square from the OLS is used to compare the ”goodness of fit”. This value is compared with the adjusted R-Square from GWR to see which model has a higher proportion of dependent variable variance accounted for by the regression model.

6. **Koenker Statistics** – test to determine which variables to select, based on the p-value. For example, if the Koenker test is significant, the analyst can only trust the robust probability. This assumes that the errors in the model are normally distributed. This is a test for non-stationarity and the robust probability will be an indicator of regional variation.

After the six tests for a properly specified model, the analyst selects the key variables from the OLS based on the p-value (statistically significant) to run the GWR. One aspect of GWR is that the estimated parameters are, in part, dependent on the weighting function or kernel selected.

In executing the GWR, the number of neighbors used for each local estimation becomes very important.
The shape and size of the bandwidth determines which features will be used to calibrate each local equation. It is always necessary to let the program choose a bandwidth or neighbor value that will identify an optimal fixed distance or optimal adaptive number of neighbors.

Select either Fixed or Adaptive as the kernel type and AICc or CV as bandwidth. If the observations are either reasonably or regularly positioned in the study area then the analyst will use Fixed kernel. This function uses the same distance as the coordinate system for feature class. The assumption is that each regression point is constant across the study area. At the regression point, the weight of a data point is unity and the weight decreases as distance from the regression point increases (Fotheringham, 2002). If the observations are clustered, then the analyst will apply the Adaptive kernel.

For the bandwidth parameter, VANHGR's approach is that any selection of number of neighbors should be based on theory by letting the data determine the number of neighbors to be included in the analysis. This is because as the bandwidth gets larger, the weights approach unity and the local GWR model approaches the global OLS model (Fotheringham, 2002). This is why it is imperative to let the data specify the distance or number of neighbors to be included in the model calibration. When features have a large variation in distribution (sparse in some areas, dense in others), adaptive kernel becomes the best choice.

Another technique VANGHR uses is optimization techniques in selecting appropriate value for the selected parameter by plotting the distance against the spatial autocorrelation Z-Score until an optimum distance is reached or at a point where the Z-score begins to decline. The optimum distance ensures that each feature has at least one neighbor.

An alternate technique utilized by VANGHR in selecting the number of neighbors parameter is to use a hot spot analysis based on the dependent variable. The p-value of the resulting Z scores in the hot spot analysis are evaluated and the number of features with a p-value of 0.05 or less is the number of neighbors to use as the parameter in the GWR analysis.

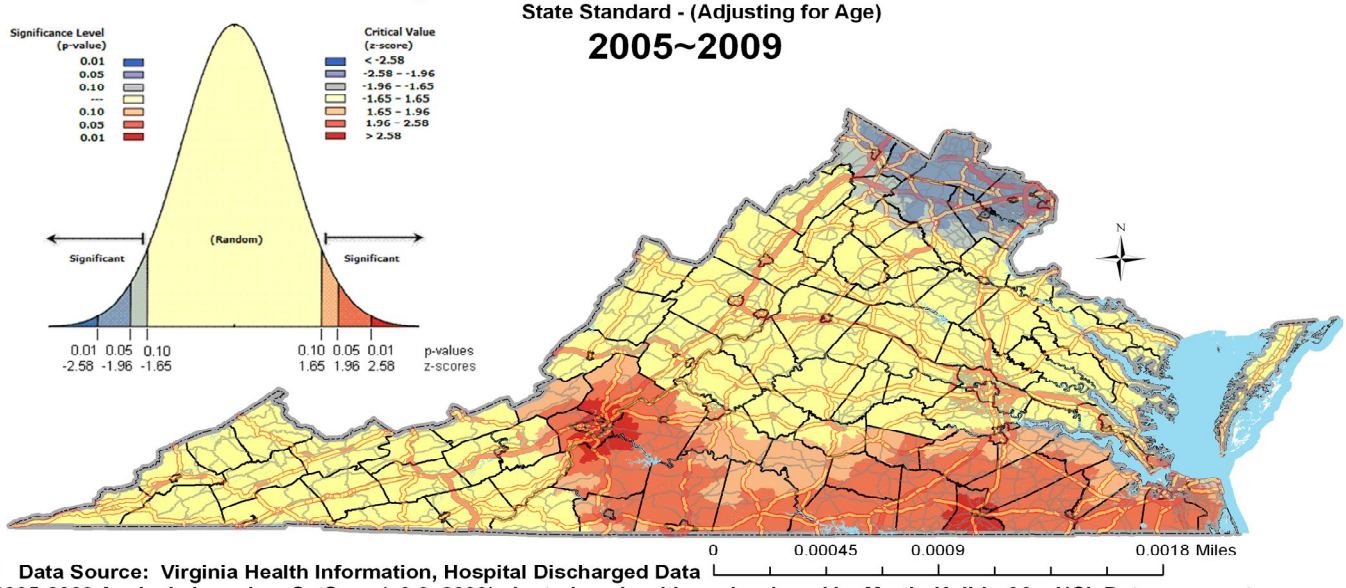
The output of the GWR is mapped (Local R-Square) to show where the model performed well and if it is answering the question being asked. The standardized residuals are also mapped to see if the residuals are random. The analyst will then run spatial autocorrelation to make sure the residuals are random. If they form clusters of high and low residuals, then the analyst takes a second look at the data to make sure a critical variable is not missing from the model.

Example Exploratory Spatial Data Analysis Maps

Virginia

Hot Spot Analysis ~ Relative Risk
Arterial Ischemic Stroke (AIS)
Hospitalization (Primary Diagnosis) Discharged Data
Ages 35 Years & Over by ZIP Code
State Standard - (Adjusting for Age)

2005~2009



* Data Source: Virginia Health Information, Hospital Discharged Data 2005-2009 Analysis based on SatScan (v9.0, 2009) clustering algorithms developed by Martin Kulldorf for NCI. Data represent Stroke Discharged Data which have been age-adjusted Viginia Standard Population. Relative Risks take into account SatScan adjustments based on distribution within contiguious area.

Virginia

Arterial Ischemic Stroke (AIS) ~ 35 Years & Over

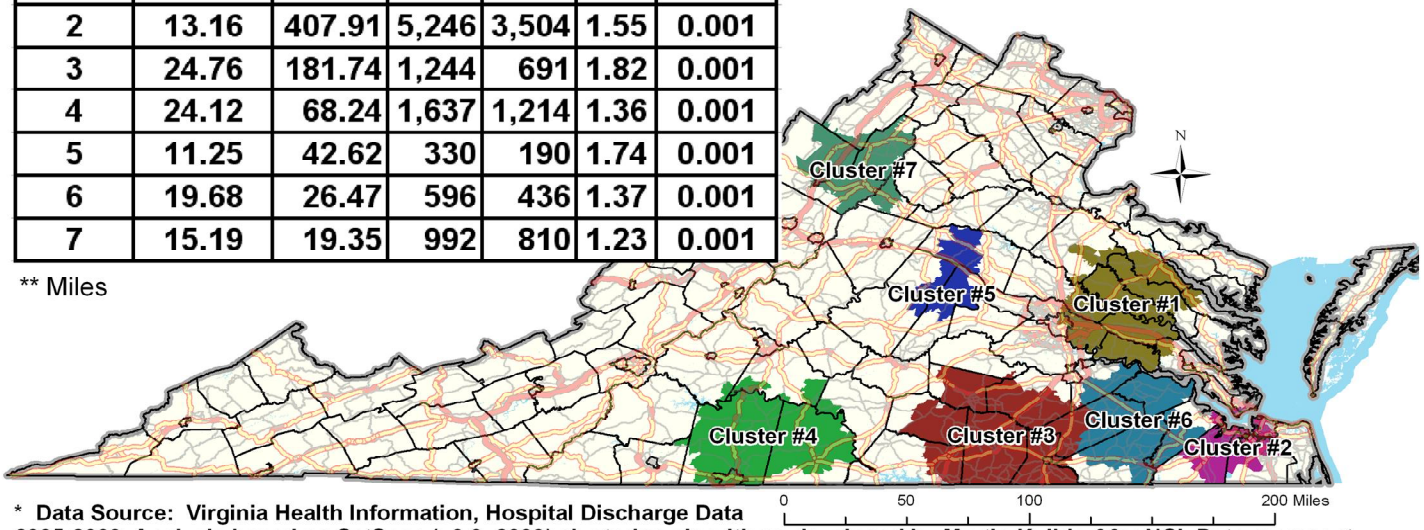
Hospitalization (Primary Diagnosis) Discharge Data *

SaTScan Cluster Analysis (25 Miles Radius)

2005~2009

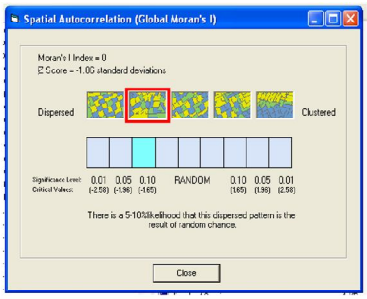
Cluster	Radius**	LLR	Obs.	Exp.	RR	P value
1	22.64	450.88	3,464	2,013	1.77	0.001
2	13.16	407.91	5,246	3,504	1.55	0.001
3	24.76	181.74	1,244	691	1.82	0.001
4	24.12	68.24	1,637	1,214	1.36	0.001
5	11.25	42.62	330	190	1.74	0.001
6	19.68	26.47	596	436	1.37	0.001
7	15.19	19.35	992	810	1.23	0.001

** Miles



* Data Source: Virginia Health Information, Hospital Discharge Data 2005-2009. Analysis based on SatScan (v9.0, 2009) clustering algorithms developed by Martin Kulldorf for NCI. Data represent Primary Diagnosis Stroke Discharges for ICD-9 Codes, 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11, 434.91, and 436. Data have been age-adjusted, to Viginia State Standard Population. Relative Risk Ratios take into account SatScan adjustments based on Poisson distributions within contiguious area.

Spatial Autocorrelation

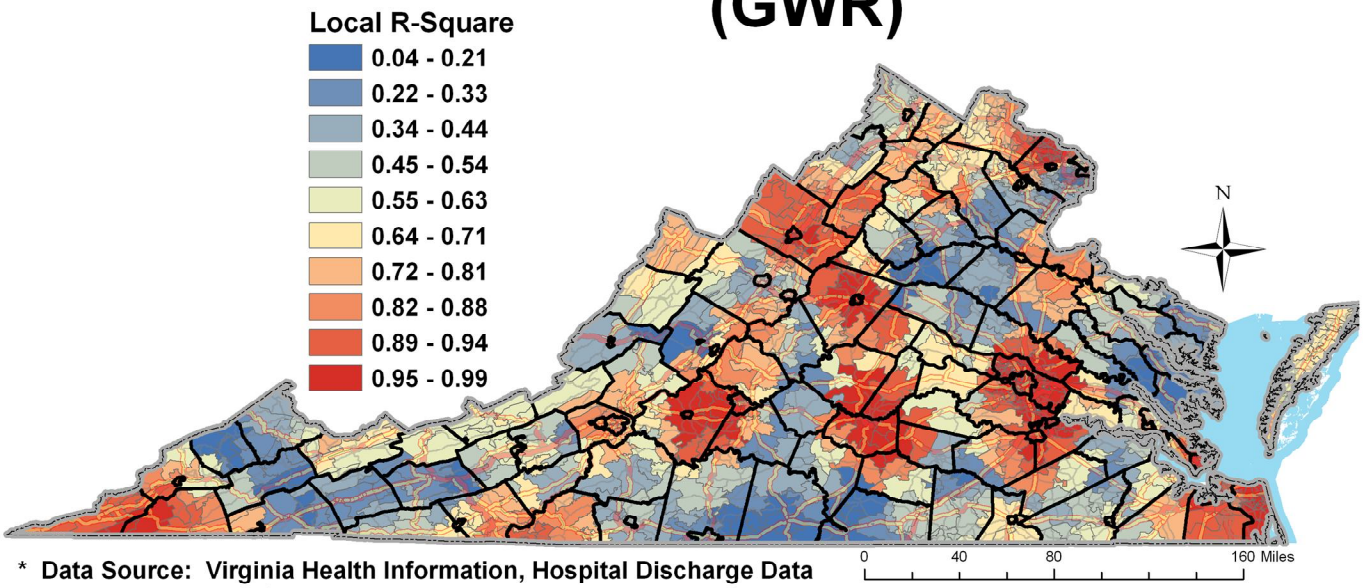


Virginia

Exploring Spatial Variation

Stroke Hospitalization Rate ~ 2005-2009

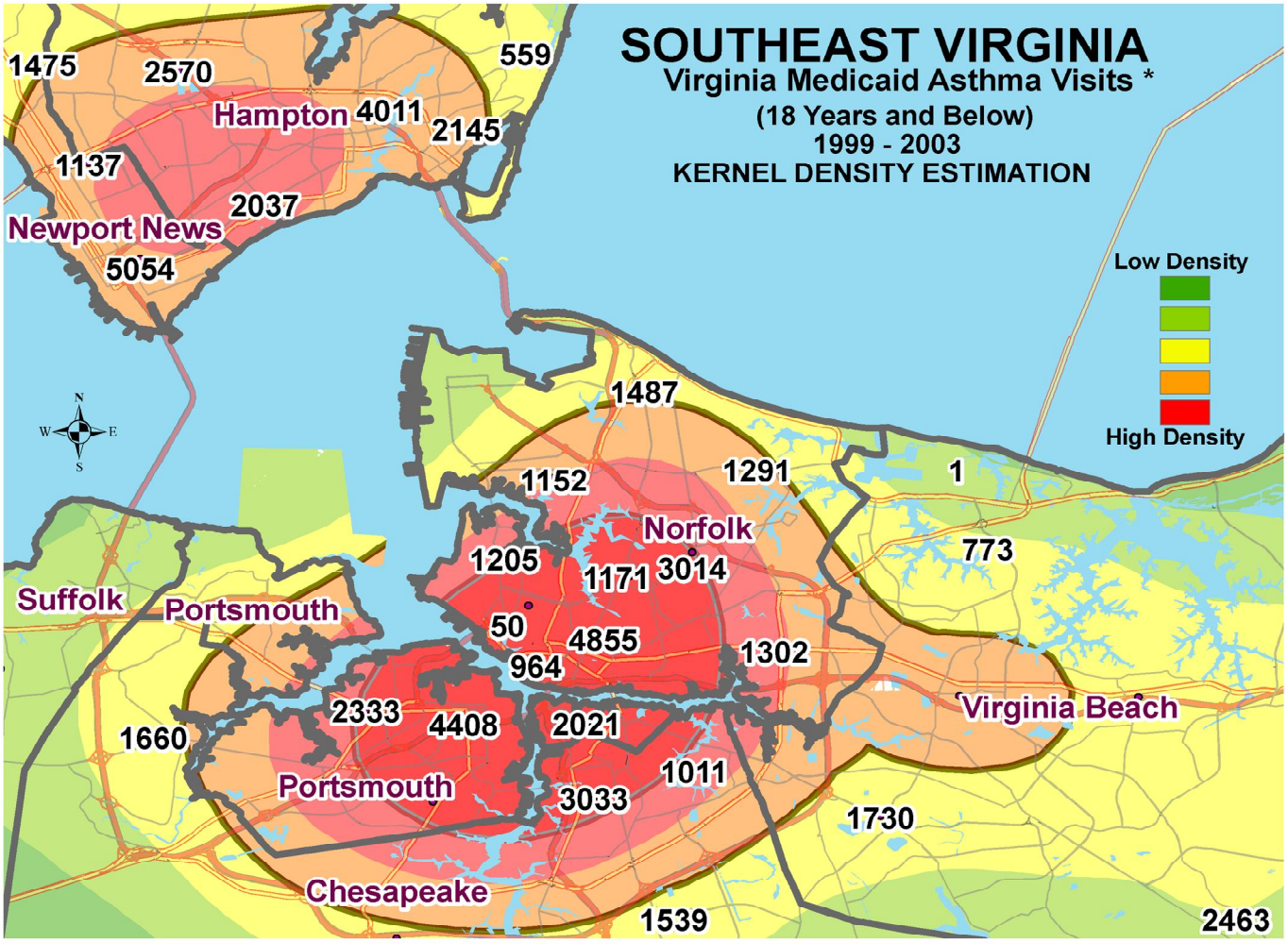
Geographically Weighted Regression (GWR)



* Data Source: Virginia Health Information, Hospital Discharge Data

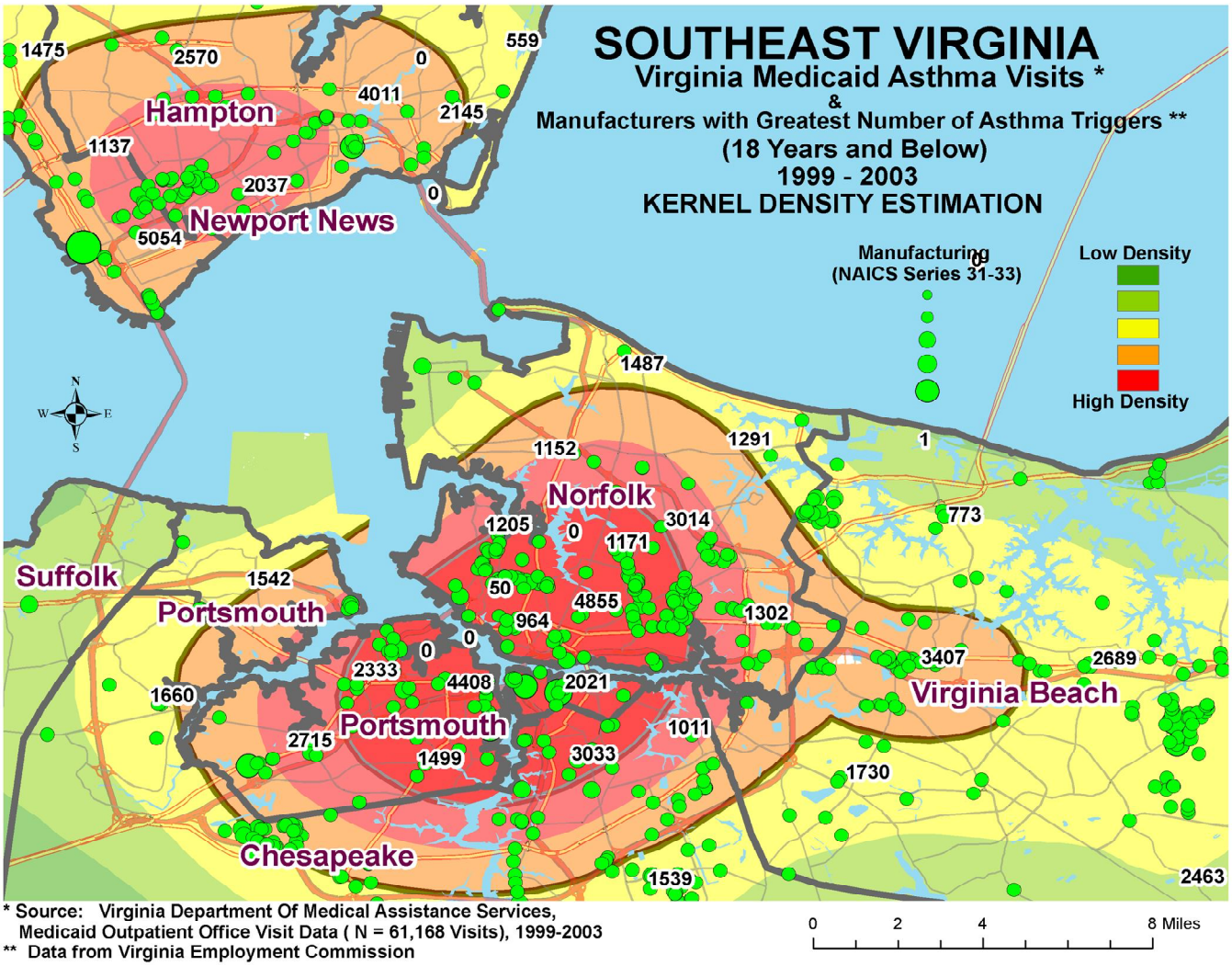
Dependent Variable: Stroke Hospitalization Rate

Independent Variables: PctPoverty, Pct 65 & Over, Average Distance to Care, PCP FTEs, Length of Stay



* Source: Virginia Department Of Medical Assistance Services,
Medicaid Outpatient Office Visit Data (N = 61,168 Visits), 1999-2003

0 2 4 8 Miles



References:

Fotheringham AS, Brunsdon C and Charlton M, 2002, *Geographically Weighted Regression: the analysis of spatially varying relationships*, Chichester: Wiley

Andy Mitchell, 2005, The ESRI Guide to GIS: Spatial Measurement & Statistics: ESRI Press.

Tobler W., (1970) "A computer movie simulating urban growth in the Detroit region". *Economic Geography*, 46(2): 234-240.

Contact Information:

- Steve Sedlock
  ssedlock@vnghr.org

Virginia Network for Geospatial Health Research, Inc.
PO Box 15818
Richmond, VA 23227
804.264.3325