Mixed Model Regression – Part 1: About the 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

None needed for this step

ESTIMATED TIME

5 minutes

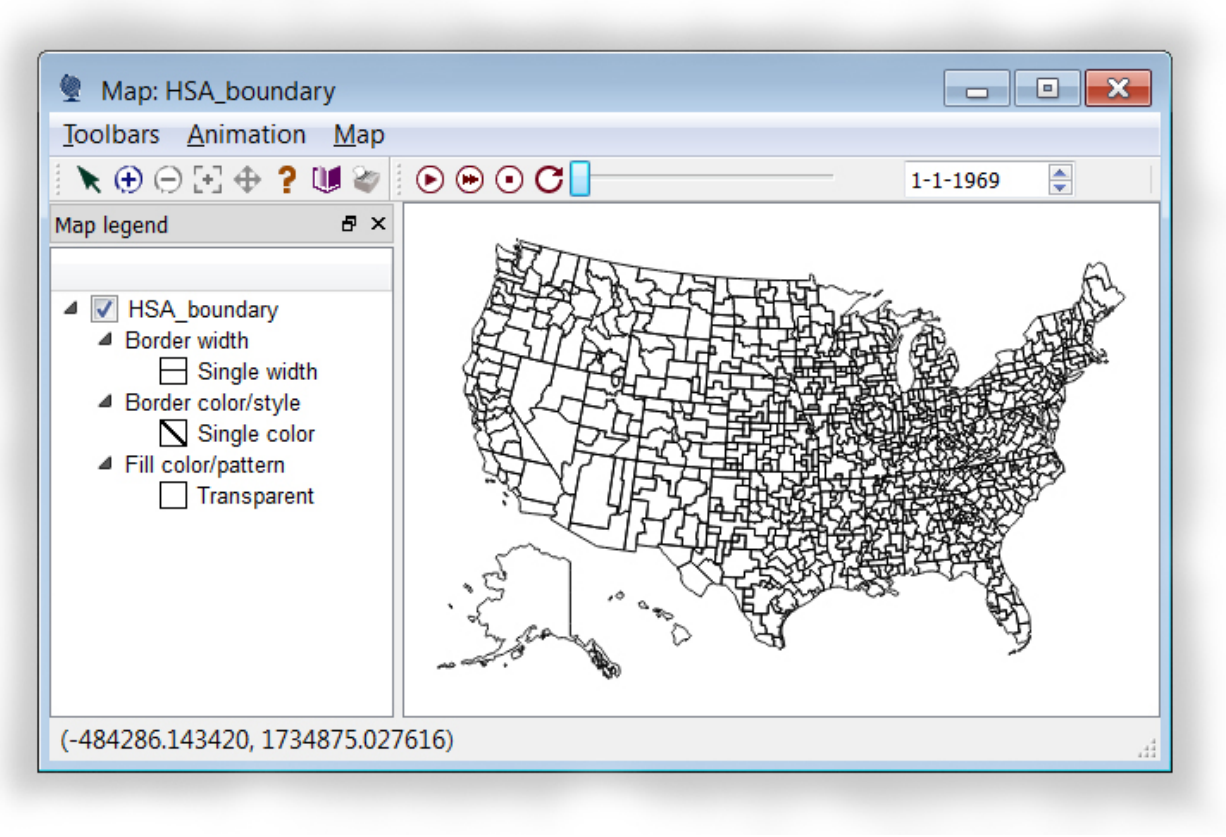
OBJECTIVE

This tutorial will use the Mixed Model Regression method in SpaceStat software to explore the patterns in all cancer mortality rates among women, 1969-2002. We will look for changing spatial patterns over time and see if some basic sociodemographic covariates can help to explain them. The mixed model approach will allow specified covariates to vary across broad U.S. regions.

In this step we will talk briefly about the data that you will analyze in the next 3 steps.

GEOGRAPHY

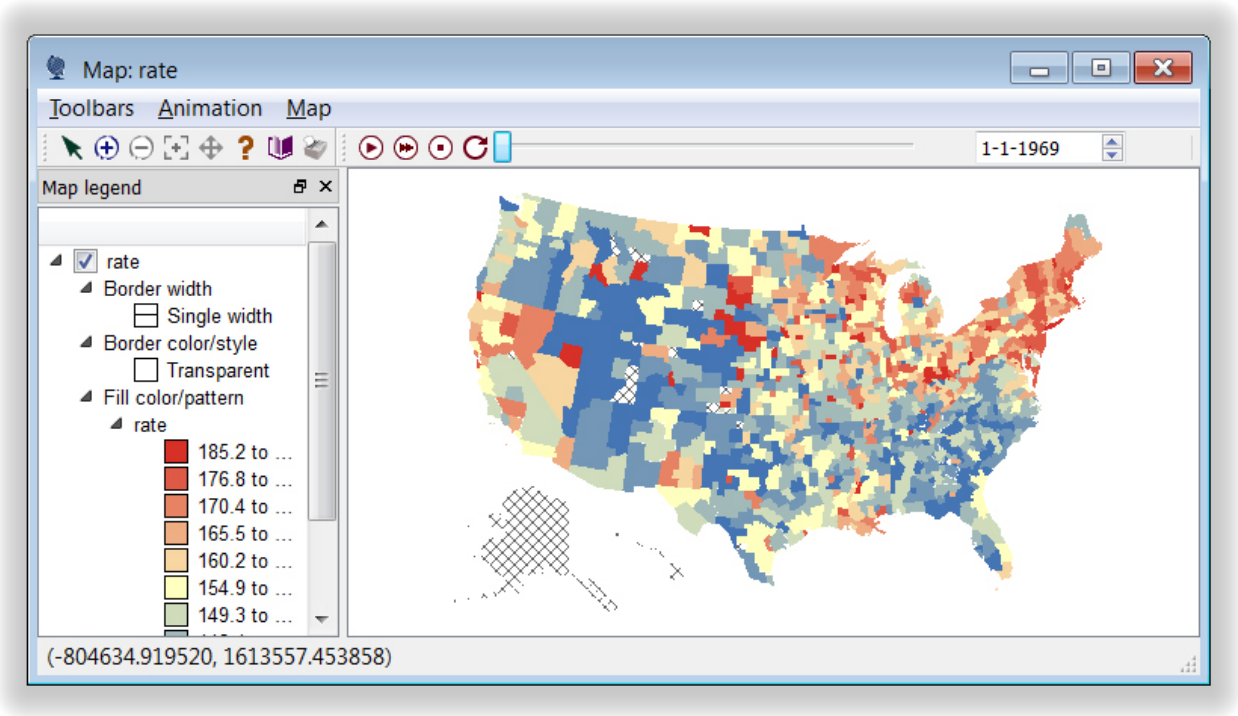
The underlying cause of death data is originally reported annually by the National Center for Health Statistics at the county level. We have aggregated counties to Health Service Areas (with an average of 3-4 counties each), with the original definitions modified by the National Cancer Institute (NCI) so that each HSA is within a single state. Changing county definitions over time have been accounted for, so that these boundaries are consistent over the period 1969-2007. The original shape file can be downloaded from http://gis.cancer.gov/tools/seerstat_bridge/fips_vars/#hsa_2000_2004_nci. (**You don't need to download this file if you are planning on using the SpaceStat project provided with Step 2 of the tutorial**)



A map of the geography rendered in SpaceStat

DEPENDENT VARIABLE

Rates, numbers of deaths and populations for death due to any malignant cancer among females were downloaded from NCI's SEER*Stat database (see http://seer.cancer.gov/seerstat/). To comply with NCHS confidentiality rules, rates and counts for HSA/year combinations that had fewer than 10 deaths are suppressed, as indicated by the missing value code -9999. Rates were directly age-adjusted to the 2000 standard million population. Mortality data cover the period 1969-2002, aggregated to 3-year periods: 1969-71, 1972-74, 1975-77, 1978-80, 1981-83, 1984-86, 1987-89, 1990-92, 1993-95, 1996-98, 1999-2002 (4 years). Alaska and Hawaii were excluded from this analysis. In the SpaceStat tutorial project, the datasets **rate**, **count**, and **pop** refer to the death rate, number of deaths, and populations for any malignant cancer among females.



A map of the death **rate** rendered in SpaceStat

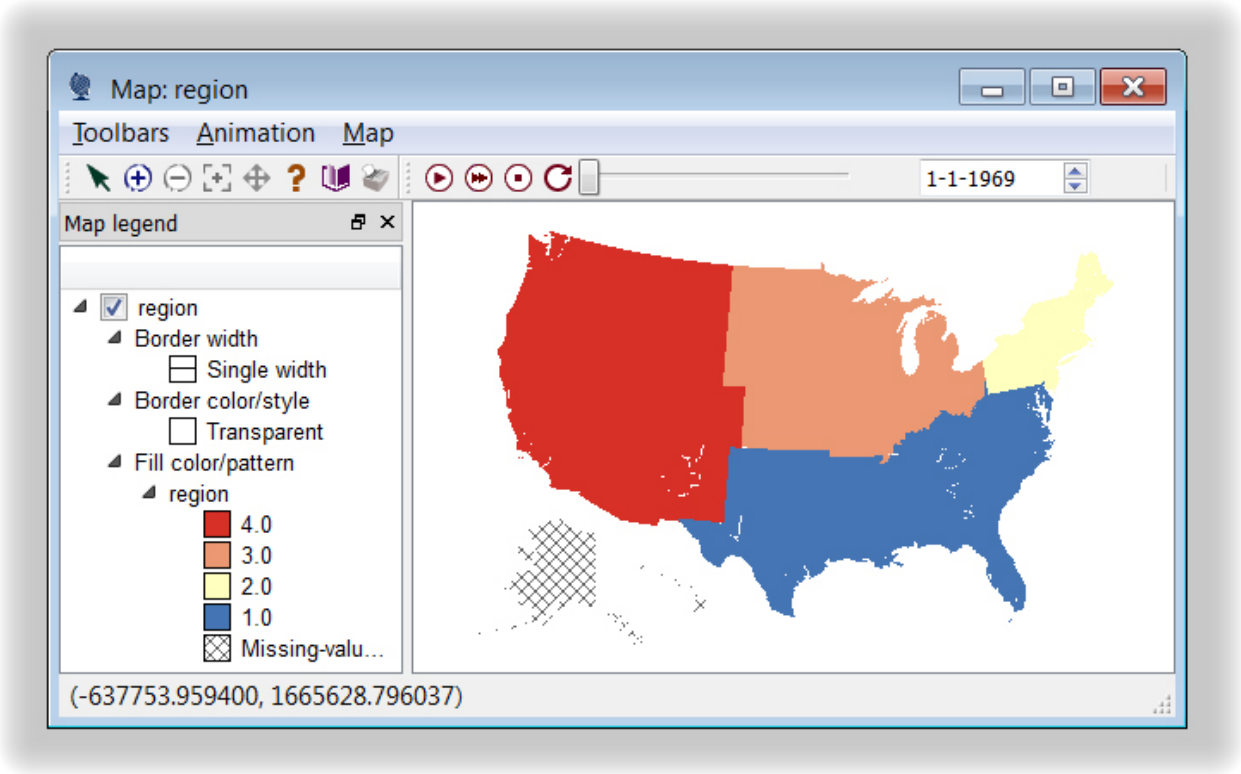
EXPLANATORY COVARIATES

Data for several potential explanatory covariates were obtained from several sources, primarily the Area Resource Files of 2002 and 2010 (http://arf.hrsa.gov/). Measures of other more direct risk factors, such as tobacco use, are not available at the small area level back to 1969. Covariate definitions (all percents are on scale of 0-100):

Pcturban: % residents living in an urban area
Pctfemhh: % of households headed by a female
MDratio: # of physicians per 1000 residents
Pcincome: per capita income
Pcincome10k: per capita income/10,000 (scaled for the model)
Pctpoor: % of individuals living below the federal poverty level
Pctcoled: % of adults (age 25+) who completed 4+ years of college
Crowded: % of households with an average of at least 1 person per room

Values of each covariate for individual years were linearly interpolated between available data points, then values for midpoint years of the mortality data aggregation were selected for use in the regression (e.g., 1970, 1973, etc); values for 2000 were used for the mortality period 1999-2002. County covariate values were aggregated to the HSA level as a weighted average of county values, with weights equal to the population of each component county for the corresponding year.

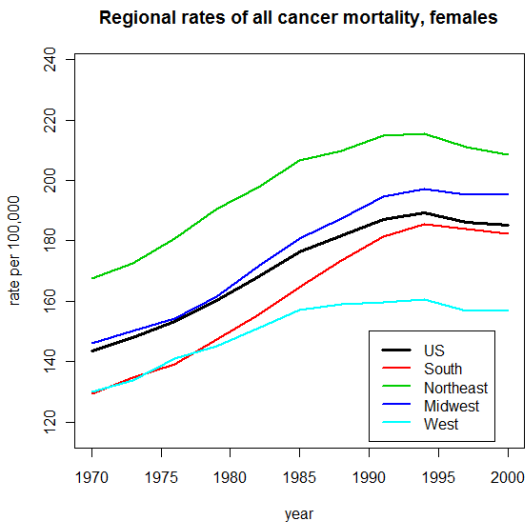
Populations for all counties within an HSA were added to obtain the total population for each year group. State FIPS # and Census Region codes were added to provide choices for the upper level geography. (Region: 1=South, 2=Northeast, 3=Midwest, 4=West. Note that we reformatted this as an integer within SpaceStat in order for it to appear in the drop-down list for upper level geography codes.).



A map of region codes rendered in SpaceStat

GOALS OF THE ANALYSIS

As we can see from the following plot, the U.S. rate (the black line) increases until the mid-1990s and then declines. However, the regional rate trends differ – the Northeast rates are parallel to the U.S. but higher over time, while rates in the West flattened out after 1985 and rates in the South rose dramatically. How much of these patterns can we explain by regressing the rates on some basic sociodemographic covariates? Is a simple fixed effects model good enough or do we need to account for heteroskedasticity (differing variances) by region? We will address these questions and others in the future parts of the Mixed Model Regression tutorial.



Regional rates of cancer mortality across time