GW-SELECT

WESLEY BROOKS

1. Introduction

Varying-coefficient models are a technique of local regression used in geostatistical analysis when one suspects that the effect of some covariate is not constant across the domain of a model. One method of fitting a varying-coefficient model is Geographically Weighted Regression (GWR) [1]. In GWR, a local regression model is fit at each point of interest (often, the points of interest are the locations where the data was collected). The technique of fitting the local regression model is to give a weight between zero and one to each observation based on its geographical distance from the point of interest, then do a weighted regression analysis to get the local model.

This paper discusses ongoing work in the area of GWR, focusing on an effort to establish a method of variable selection for GWR models. Variable selection can mean several things:

- What are the predictor variables that have no effect on the measured output, anywhere in the model's domain?
- Which coefficients are constant throughout the model's domain?
- Which coefficients are zero in some regions but non-zero in others?

The goal of this project is to develop a method that can answer these questions and work out the conditions under which the answers are reliable.

2. Data

2.1. **Observational data.**

Observational data comes from four sources:

- Poverty data:
- Soybean aphid data:

- Mountain Pine Beetle Data:

- Land-cover data:

2.2. **Simulated data.** Simulation studies are used to validate the method of analysis.

3. Methods

The basic operation of GWR is to build a regression model at each of a set of pre-specified locations. At each model location, all of the observations in the data set are weighted based on their distance from the model location (the weights are uniquely specified by the combination of a kernel function and a bandwidth). The weighted observations are then used to build a regression model that is only meaningful at that location. The regression model can be quite generic - for gaussian data something like `R`'s `lm` function is used, and for the binomial case, something like the `glm` function is used. Prior weights can be specified in case over-dispersion (or heteroskedasticity) is expected. This came into play with the poverty data: each county's entry was weighted by its population.

3.1. **Kernels.** At this time there are two kernel functions working: the bisquare kernel:

$$\mathrm{W(dist,\ bw)} = (1 - (\frac{\mathrm{dist}}{\mathrm{bw}})^2)^2$$

...And the gaussian kernel:

$$\mathrm{W(dist,\ bw)} = \phi(\frac{\mathrm{dist}}{\mathrm{bw}})$$

3.2. **Bandwidth selection.** The bandwidth is selected by minimizing the sum of absolute cross-validation (CV) errors from the model-fitting. Each of the unique observation locations is used as a model location, and the observations from that point are reserved as the test set. The model is built over the other data and used to predict the test set. The sum of the absolute errors from this prediction step are added to the total. The `optimize` function in R uses a golden section search to find the minimum of the total absolute CV error.

3.3. **Gaussian case (linear model).** test

3.4. **Generalized linear model.** test

4. Variable selection

References

[1] A.S. Fotheringham, C. Brunsdon, and M. Charlton. *Geographically weighted regression: the analysis of spatially varying relationships.* Wiley, 2002.