PROJECT SPATIAL ECONOMETRICS

Christine Thomas

Home to work commuting flows modeling

M2 Statistics and Econometrics - UT1 - February 2015

1 Instructions

The project is done by teams of two. Please announce your teams on Moodle next week. I will not accept a project by a single person, if there is an odd number of students in the elective, please do a team of three (a single one !). You should turn in by mail **no later than Sunday March 22nd.**

1. A set of slides (no more than 20) presenting your results
2. A commented code file

There will be an oral presentation of the projects on **Tuesday March 24th.** During the oral presentation, each student will present part of his team slides and will be asked questions about the project, the two papers and/or the course.

2 Accompanying files

1. Data set of explanatory variables :

   `Xfile_district.xls`

2. Target variable flows :

   `flux_district.txt`

3. Map :

   `map_district_region.dbf, map_district_region.prj`
   `map_district_region.shp, map_district_region.shx`

4. Two papers to read (Chun et al., LeSage and Pace)

3 Objective

We consider the problem of analyzing the determinants of home to work commuting in the region of Toulouse based on a data set of the 1999 census from INSEE (the original data has been disturbed by a small noise for confidentiality reasons). The study region has been designed so that as few people as possible coming from outside the region and working inside and symmetrically as few people as possible living inside and working outside. The target variable contains "theoretical flows"(between the districts around Toulouse) in the sense that they have not been measured on a given day nor averaged on

a given period, but hypothesized from the addresses of home and work of the active population declarations. The explanatory variables comprise characteristics of the population such as socio-professional category, sector of activity, age class (coming from the same file as the flow data), housing characteristics, area of the spatial units, latitude and longitude of the centroids of the spatial units. The socio-professional category, sector of activity and age class are available for the active population living in a zone (active residents) as well as for the active people working in the zone (workers). You will compute the employment coverage rate as the ratio between the total number of jobs in a zone to the active population of the zone. When this ratio is larger than 1, one expects to observe incoming flows, and it is the case for zones in the center of our study area, whereas when it is smaller than 1, one expects outgoing flows and these zones are in the outskirts of this region.

4 Work to do

4.1 Exploratory phase

After loading the map files and the data files

1. Compute and map the incoming flows (to a district) and outgoing flows (from a district).
2. Construct a neighborhood matrix for the district centers locations W and analyze it
3. Do a spatial exploratory analysis of the incoming and outgoing flows (contrast them with the employment coverage rate)

4.2 Modeling phase

1. Fit a traditional gravity model to explain the district to district flows using the following explanatory variables : labour force (at origin and destination), employment (at origin and destination), activity and unemployment rates (at origin and destination), housing units (at origin and destination)and distance. Select a smaller number of explanatory variables (up to seven). You will use specific coefficients for diagonal flows. Analyze the residuals.
2. Construct an origin-based dependence and destination based dependence matrices W_O and W_D as in LeSage and Pace (2008) from your matrix W. Using the matrix $W_O + W_D$ as neighborhood structure for flows, analyze the spatial autocorrelation of the residuals of the traditional gravity model.
3. Fit a LAG model to the flows and present the results.
4. Fit a spatial Durbin model to the flows and present the results (compare the two models).

4.3 Prediction phase

Eliminate 10% of the flows by random selection and call Y_1 the new vector of flows. We are going to test the predictive ability of the above models by leave-one-out cross validation. For a given eliminated flow corresponding to origin o and destination d, after fitting the model with the non eliminated units (Y_1 data and the corresponding X_1 variables), you will predict Y_{od} by the following formula for the LAG model

$$\hat{Y}_{od} = X_{od}\hat{\beta} + \hat{\rho}W^{od}Y_1,$$

where the matrix W^{od} is obtained by extracting from the original $W_O + W_D$ matrix row od and columns corresponding to non eliminated units (Y_1).

1. Explain the proposed formula for the LAG model
2. Compute a prediction with the traditional gravity model
3. Compute the predictions for all eliminated flows (for the LAG model) and the compute the quadratic mean error of prediction as well as the relative quadratic mean error of prediction. Compare with the accuracy of the traditional predictions and comment the results.