Extended abtract

EXTENDED ABSTRACT

Title: **S**patial Grouped-Data Probit models for the analysis of the inter-regional population flows in Spain

Authors and e-mail of all:

Luc Anselin
anselin@uchicago.edu
Department: The Center for Spatial Data Science
University: University of Chicago

Patricio Aroca
patricio.aroca@uai.cl
Department: Economics
University: Adolfo Ibáñez University

Coro Chasco
coro.chasco@uam.es
Department: Applied Economics
University: Autonomous University of Madrid

Subject area: *S04 – Spatial flows: towards an empty Spain and a full Spain*

Abstract: *(minimum1500 words)*

Human migration is a social process with widespread consequences for both individuals and the society. Particularly, internal migration refers to human migration within one geopolitical entity, usually a nation. A general trend of movement from rural to urban areas is a form of internal migration, which has led to rapid urbanization in many countries. In general, motives for internal migration prominently include travel for education and for economic reasons. The increasing role of this topic in the socioeconomic and demographic development of countries has promoted the interest in the formulation of mathematical models of migration.

As a socio-economic process, migration modeling has been applied to both, micro and macro-levels (Aleshkovski and Iontsev 2006). The micro approach focuses on the migration behavior of individuals or households and intends to explain the decision-making process by potential migrants to remain in a current residence or to migrate to another one. This type of models is based on disaggregated data usually delivered by surveys. In many areas of economics, choices made by individuals are costly to collect or inaccessible. However, analysts may have access to the choice data aggregated across groups of individuals in the form of counts or shares. This is the macro-approach of

migration modeling, which studies the patterns of migration of certain social groups within a given territory.

One of the most well-known modeling methods at the aggregate level, where observations are referred to discrete administrative regions, are the spatial interaction models, which are also referred as gravity models (Sen and Smith 1995, LeSage and Fischer 2008). In a seminal paper, LeSage and Pace (2008), hereafter LSP, provide specifics regarding how spatial regression methods can be applied to spatial interaction models trying to overcome different kind of caveats. First, they explain carefully the notation and conventions to describe origin-destination (OD) flows considering that in contrast to typical spatial regression models, where the sample involves n regions or observations, spatial interaction models involve $n^2=N$ OD pairs, with each OD pair being an observation. Second, they use the algebra of the Kronecker products to form moment matrices without dealing with usually big N by N matrices, which simplifies and speeds the computation of the model estimation. Third, since the OD flows are expressed as count data which follow a Poisson distribution, the normality assumption for the ordinary least squares (OLS) and maximum-likelihood (ML) methods, is seriously threatened in spatial interaction models, where additionally a large number of zero flows usually exist. Hence, they propose to transform the dependent variable of counts in logarithms –among other methods– in order to approximate normality. Forth, they created a separate model, with different explanatory variables, for intra-regional flows since they are different and larger than inter-regional ones. Fifth, they recognize the possible existence of endogeneity problem in certain explanatory variables like population and employment. Sixth, they augment the conventional gravity model with richer forms of spatial dependence in order to consider neighboring locations' influence over both origin and destination regions.

Additionally to these limitations of spatial interaction regression models, it must be added the fact that the use of count data (absolute frequencies) for OD flows has little meaning alone, unless the size of the population of the regions from which they are derived is similar, or relatively so, which is often not the case in this context. Hence, we propose to use rates, in which the numerator is the number of people who migrate during a certain period of time and the denominator is the total population "at risk" of migrating for the same period of time. In fact, these rates are shares or '*observed relative frequencies of interactions which yield to meaningful estimates of interaction probabilities between origin-destination pairs*' (Sen and Smith 1995).

For all these reasons, the aim of this paper is to propose other strand of the literature based on a choice-theoretic perspective, since the advantages of the application of this kind of models to OD flows is definitive. Choice models have been specified to explain not only individual behavior but also grouped data when observations no longer consist of single individuals but sets of several persons who share similar characteristics (e.g. living in the same region). In the grouped-data choice models, the dependent variable

consists of a number of observed proportions or relative frequencies, which can be considered as an estimation of the theoretical proportions. Hence, we can treat this problem as a simple one of sampling from a Bernoulli population so as the observed probability that an individual moves from an origin to a destination is equal to the population proportion, which in turn is a function of a set of explanatory variables and their corresponding impact parameters, plus an error term (Aroca and Hewings 2002 and Borjas 2006.).

Grouped-data logit and probit models can be linearized in the impact parameters so as the dependent variable is the inverse form of the observed proportions (Gourrieroux 2000, section 4.2). Specifically, we propose employing the groped-data probit (GProbit) model for OD flows instead of the spatial interaction or gravity model, due to the clear benefits of this application. First and foremost, the GProbit specification for a “heard” behavior as a clustering process of individual decision behavior is more consistent with the neoclassical consumer theory, for which consumers scan the expected utilities obtainable at each alternative location open to them and select the one with the greatest expected benefit (Mueller 1982). Second, the normality assumption is more likely in grouped-data choice models than in gravitational models, since the dependent variable is not specified as counts but an inverse function of the cumulative standard normal distribution. Third, due to the probability nature of the grouped-data flow proportions, intra-regional flow rates can be easily derived as differences of interregional flow rates from the unity, with the creation of a separate model for the estimation of intra-regional flows no longer needed. Forth, the linear form of the probit model for grouped-data OD flows also allows for on the one hand, dealing with possible endogeneity problems and on the other hand, extending spatial dependence and/or heterogeneity models as it is usual in spatial econometric models.

We illustrate the performance of a spatial grouped data probit model to estimate internal regional migration flows in Spain, in order to test for spatial dependence in the regression residuals, and we compare the results with the obtained with a spatial autoregressive interaction model. Internal migration has been a key component in Spain’s inter-regional population dynamics, particularly during the 1960s and 1970s, when the rural exodus was at its peak (Coll and Stillwell 2000). This prominent process of labor migration declined during the period of industrial restructuring, but it recovered pulse in the expansion years reaching a stabilization level during the Great Recession (2008-2012), which is the time period we analyze in this paper (Izquierdo et al. 2015).

The outline of the paper is as follows. In the next section, we review the problems that confront the empirical implementation of the conventional spatial interaction model. In section 3, we present the GProbit model for OD flows as a more effective alternative than the spatial interaction models to overcome these problems. In Section 4, we estimate both type of models for internal migration flows across Spanish regions during the period 2008-2012. Finally, Section 5 concludes.

References

Aleshkovski I, Iontsev V (2006) Mathematical models of migration. In: Livchits VN, Tokarev VV (eds) *Systems analysis and modeling of integrated world systems. Encyclopedia of Life*, Ontario (Canada), pp 185-213

Aroca P, Hewings GJD (2002) Migration and regional labor market adjustment: Chile 1977–1982 and 1987–1992. *Ann Reg Sci* 36:197–218

Borjas GJ (2006) Native internal migration and the labor market impact of immigration. *The Journal of Human Resources* 41(2):221-258

Coll AG, Stillwell J (2000) *Internal migration and regional population dynamics in Europe: Spain case study*. Working paper, School of Geography, University of Leeds (UK)

Gourrieroux C (2000) *Econometrics of qualitative dependent variables*. Cambridge University Press, Cambridge (UK)

Izquierdo M, Jimeno JF, Lacuesta A (2015) *Spain: from immigration to emigration?* Working Paper, Banco de España nº 1503, Madrid (Spain).

LeSage JP, Fischer MM (2008) *Spatial econometric methods for modeling origin destination flows*. WU Vienna University of Economics and Business, Vienna. Available at: http://epub.wu.ac.at/3957

LeSage JP, Pace RK (2008) Spatial econometric modeling of origin-destination flows, *Journal of Regional Science* 48(5), 941-967

Mueller CF (1982) *The economics of labor migration: A behavioral analysis*. Academic Press, Inc., New York (US)

Sen A, Smith TE (1995) *Gravity models of spatial interaction behavior*. Springer, Berlin, Heidelberg and New York

Keywords: *spatial probit, grouped data, migration flows, Spanish regions*
JEL codes: C21, C25, R23