Forecasting Spatially Dependent Origin and Destination Commodity Flows

James P. LeSage
Fields Endowed Chair of Urban and Regional Economics
McCoy College of Business Administration
Department of Finance and Economics
Texas State University - San Marcos
San Marcos, TX 78666
jlesage@spatial-econometrics.com

Carlos Llano
Departamento de Análisis Económico: Teoría Económica e Historia Económica
Facultad de Ciencias Económicas y Empresariales
Universidad Autónoma de Madrid
28049 Madrid, Spain
carlos.llano@uam.es

April 17, 2013

Abstract

We explore origin-destination forecasting of commodity flows between 15 Spanish regions, using data covering the period from 1995 to 2004. The one-year-ahead forecasts are based on a recently introduced spatial autoregressive variant of the traditional gravity model. Gravity (or spatial interaction models) attempt to explain variation in $N = n^2$ flows between n origin and destination regions that reflect a vector arising from an n by n flow matrix. The spatial autoregressive variant of the gravity model used here takes into account spatial dependence between flows from regions neighboring both the origin and destinations during estimation and forecasting. One-year-ahead forecast accuracy of non-spatial and spatial models are compared.

KEYWORDS: gravity models, Bayesian spatial autoregressive regression model, spatial connectivity of origin-destination flows.

1 Introduction

The term 'spatial interaction models' has been used by Sen and Smith (1995) to label models that focus on flows between origin and destination locations. Others refer to these as 'gravity models', since they seek to explain variation in the level of flows across a sample of $N = n^2$ origin-destination (O-D) flows by relying on a function of distance between the origin and destination locations as well as explanatory variables reflecting size of the regions. As in the case of gravity, the push and pull forces associated with origin and destination locations further apart are thought to be weaker, and the attraction between larger regions is greater.

Flows sometimes take the form of counts, for example a count of persons migrating from location A to B over time, or the number of workers commuting from home to work locations. The estimation and forecasting methods set forth here do not apply to flows taking the form of counts, since our methods require flows that exhibit a normal distribution. Extending our methods to the case of Poisson or Negative Binomial distributions is beyond the scope of this article. The methods we outline pertain to commodity or trade flows between locations measured in dollars, that can be log transformed to exhibit an approximately normal distribution. One can also encounter cases where flows between a large number of the regions are absent. This might be true for flows between small isolated regions measured over a short time horizon, for example monthly migration of persons from the state of Rhode Island to South Dakota might be zero. For a discussion of issues that plague regression and maximum likelihood models of flows see LeSage and Fischer (2010).

Ordinary least-squares estimation of spatial interaction models relies on two matrices of explanatory variables that we label X_d, X_o constructed to reflect characteristics associated with origin and destination locations for each O-D pair. Distance between each O-D pair is also included as an explanatory variable. Use of least-squares to estimate these models assumes that the O-D flows are spatially independent as well as following an approximately normal distribution. The validity of the independence assumption has long been questioned. For example, Griffith and Jones (1980) conjecture that flows from an origin are "enhanced

or diminished in accordance with the propensity of emissiveness of its neighboring origin locations". They also state: that flows associated with a destination are "enhanced or diminished in accordance with the propensity of attractiveness of its neighboring destination locations". In other words, spatial dependence is more likely than spatial independence when considering O-D flows. Spatial dependence in our flow setting refers to the fact that flows from nearby locations (either origins or destinations) will be similar in magnitude. The interested reader is referred to LeSage and Pace (2008) for a more complete development of these ideas.

For the case of spatial dependence in situations involving flows taking the form of counts, a number of modeling and estimation procedures have appeared in the literature. Grimpe and Patuelli (2011) model regional patent counts (as the dependent variable) using a Negative Binomial model after filtering the dependent variable for spatial dependence, while Scherngell and Lata (2011) use a Poisson model that filters the dependent variable patent counts for spatial dependence. Pre-filtering for spatial dependence allows application of standard Negative Binomial or Poisson regression methods. Fischer and Griffith (2008) assume that spatial dependence exists only in the model disturbances, and model this using a spatial autoregressive process applied to the disturbance terms. These works pertain only to estimation and inference, but not to the topic of forecasting flows reflecting count magnitudes, which is an area for future research.

In this study we compare forecasting performance of the classical regression gravity model with a spatial regression model that incorporates a dependence structure for the flows. The forecast accuracy of robust and non-robust Bayesian variants of the spatial regression model for commodity flows is compared to the more conventional least-squares model as well as robust variants of these. Spatial regression models are widely used in applied spatial econometric work, but a comparison of forecasting accuracy for these models with more conventional regression models in a flow context has not appeared in the literature.

Our sample data consists of $n = 15$ Spanish regions where the commodity flows have been organized as an n by n 'origin-destination (O-D) flow matrix' that we label Y. Without

loss of generality, the row elements of the matrix $Y_{ij}, j = 1, \ldots, n$ reflect flows from origin regions j to each destination region $i = 1, \ldots, n$. In our applied example, we use (logged) dollar values of commodity flows, and treat the columns as 'origins' of the commodity flows and the rows are 'destinations' of the flows. We explore the space-time dimension of these commodity flows using the C-intereg database. This allows us to estimate the Spanish commodity flow model annually for the period 1995-2004. One-year-ahead forecasts for the set of $N = n^2 = 225$ commodity flows are used in our real-time forecasting experiments.

Apart from the spatial autoregressive extension of the classical gravity model, we explore a method proposed by LeSage and Pace (2008) to address a problem that often arises in flow modeling. Intra-regional flows (those within the regions) tend to be very large relative to inter-regional flows (those between regions). Researchers interested in interregional aspects of flows often set the intraregional flows that reside on the main diagonal of the flow matrix to zero values in an effort to prevent these observations from entering the model (see Tiefelsdorf, 2003 and Fischer et al., 2006). The phenomenon of large within region or country flows versus smaller between region flows has been extensively analyzed in the international trade literature under the rubric of 'border effects' (see: McCallum,1995; Wolf, 2000; Anderson and van Wincoop, 2003). According to this literature, interregional borders result in larger trade within two neighboring regions than between these regions even if the two regions exhibit similar production and consumption levels.

In our spatial regression model we avoid setting the intraregional flows to zero since this will have an adverse impact on local averages of the dependent variable used as an explanatory variable in these models (see LeSage and Pace, 2008). The approach proposed by LeSage and Pace (2008) involves creating a separate model for the within region flows (those on the main diagonal of the flow matrix). Use of a distinct set of explanatory variables (and associated parameters) to model variation in the main diagonal (intraregional) flows prevents parameter estimates associated with the origin and destination explanatory variables from being influenced by large within region flow magnitudes. Our forecasting experiments compare the forecast accuracy of models constructed using this approach to

more conventional approaches.

Section 2 of the paper describes the conventional spatial interaction model along with the spatial regression extension from LeSage and Pace (2008). Section 3.1 describes the data and methodology used in our forecasting experiments based on the Spanish commodity flows. Section 3.2 is devoted to a procedure that can be used to forecast dependent data, which requires additional effort relative to the case of independent data. Analysis of forecasting accuracy comparisons for the various models is the subject of section 3.3.

2 Empirical modeling of commodity flows

In section 2.1 we review the traditional gravity or spatial interaction model that assumes the O-D flows contained in the dependent variable vector $y = vec(Y)$ are independent, consistent with the Gauss-Markov assumptions for least-squares. Section 2.2 describes a procedure from LeSage and Pace (2008) for dealing with the scale differences that arise due to the presence of large intraregional flows versus smaller interregional flows. This procedure can be applied to the traditional as well as spatial versions of the flow model. Section 2.3 describes the spatial regression extension of the gravity or spatial interaction model.

2.1 Conventional gravity models

The conventional gravity (or spatial interaction) model attempts to explain variation in the set of $N = n^2$ O-D flows expressed as a vectorized version of the flow matrix: $y = \text{vec}(Y)$, where we assume that the flow matrix Y is organized so that the columns represent 'origins' of the commodity flows and rows are 'destinations'. A conventional n by k explanatory variables matrix X is strategically repeated using the Kronecker product ($\otimes$) to form: $X_d = \iota_n \otimes X$, where ι_n is an n by 1 vector of ones. This produces an N by k matrix of explanatory variables that reflect characteristics of the destination region of the flows. Similarly, we can use the Kronecker product to construct an N by k matrix $X_o = X \otimes \iota_n$, representing characteristics of the origin regions. The other important explanatory variable use to explain variation in commodity flows is distance, constructed by vectorizing a distance

matrix D that measures distances between all origins and destinations, e.g., $d = \text{vec}(D)$.

These variables are used to form a log-linear regression relationship shown in (1). If one starts with the standard gravity model and applies a log-transformation, the resulting structural model takes the form of (1) (c.f., equation (6.4) in Sen and Smith, 1995). In (1), β_d and β_o represent k by 1 parameter vectors associated with the N by k matrices $X_{d,t-1}$ and $X_{o,t-1}$. The scalar parameter γ reflects the effect of the distance vector d, and α denotes the constant term parameter. The N by 1 vector ε_t represent disturbances, and we assume $\varepsilon \sim N[0, \sigma_t^2 I_N]$, where we use $N[\mu, \Sigma]$ to represent a normal distribution with mean μ and variance-covariance Σ. In (1), past period characteristics of the destination and origin regions $(X_{d,t-1}, X_{o,t-1})$ are posited to explain flows during the next time period (t). Conventional cross-sectional models typically rely on contemporaneous characteristics $(X_{d,t}, X_{o,t})$ to explain current (period t) flows, which could introduce simultaneity bias in least-squares estimates.

$$y_t = \alpha \iota_N + X_{d,t-1}\beta_d + X_{o,t-1}\beta_o + \gamma d + \varepsilon_t \tag{1}$$

In (1), we establish a relationship between flows in the current year, y_t and the N by k explanatory variable matrices, $X_{d,t-1}$, $X_{o,t-1}$ containing destination and origin characteristics from the previous year. Of course, the distance vector d does not change over time. This allows us to produce one-year-ahead forecasts using the estimated k by 1 parameter vectors β_d and β_o along with the scalar parameter estimates α and γ. Forecasts for period $t+1$ are calculated using parameter estimates based on period t information, as would occur in real-time forecasting of the model.

Interpretation of the individual coefficient estimates from the vector β_d in this model are that: a positive coefficient for a particular variable vector from the matrix $X_{d,t-1}$ results in larger flows (in the next year) to destinations having larger values of the variable. Similarly, for the case of positive β_o estimates, we would expect to see larger outflows (during the next year) from regions having larger values of the variable. For example, both origin and destination regions that have larger populations might be expected to have larger in-

and out-flows, since more population reflects an increase in the 'size' of the regions. Origin variables having positive parameters are sometimes referred to as 'push factors', since larger values for these variables lead to more outflows, whereas destination variables with positive parameters are considered 'pull factors', producing more inflows.

Estimation of this variant of the model can be accomplished using ordinary least-squares. One problem that arises in using the model in (1) is that the vector y_t contains intraregional flows that represent the main diagonal of the flow matrix Y. These are typically very large relative to interregional flows from off-diagonal elements of the flow matrix. This results in parameter estimates for β_d, β_o that reflect a compromise between explaining the large intra- and small interregional flows. We turn attention to this issue in the next section.

2.2 Treatment of intra- and interregional flows

Recent empirical research on interregional and international trade has produced a wide range of results regarding the impact of borders on trade. For example, studies have found that pairs of regions within a country will trade between 4 to 20 times as much as otherwise identical pairs of regions separated by borders.[1] Other work has focused on the magnitude of domestic market fragmentation when intraregional trade flows are available. Again intraregional trade flows appears to be substantially larger than interregional, with a ratio that varies between 2 and 20.[2]

Based on these findings, we would expect a differential response of intraregional flows to explanatory variables $X_{d,t-1}, X_{o,t-1}$ than interregional flows, based on scale differences alone. In a least-squares setting, these scale differences will bias the parameter estimates β_d, β_o in favor of explanatory variable characteristics that do a good job of explaining the large intraregional flows. We note that of the $n \times n$ flows which we are modeling, there are only n intraregional flows compared to $n \times n - n$ interregional flows. Tiefelsdorf (2003) and Fischer et al. (2006) proposed setting the intraregional flows (the main diagonal

[1] The external border or frontier effect has been studied by McCallum (1995), Helliwell (1996, 1998), Anderson and Smith (1999), Anderson and van Wincoop (2003), Okubo (2004), Gil et al. (2005), Nitsch (2000), Evans (2003) and Chen (2004).

[2] Helliwell (1996), Wolf (2000) and Combes et al. (2005) among others

elements of the flow matrix Y to zero in an effort to focus on interregional flow responses to the explanatory variables. Since our objective is to produce forecasts for both intra- and interregional commodity flows, this approach is ruled out.

Our approach was to follow LeSage and Pace (2008) and create a separate model for intraregional flows, which can be accomplished by setting all elements of the matrices X_d, X_o and the intercept vector ι_N corresponding to the intraregional flows to zero. This prevents the large magnitudes associated with these observations from entering the interregional flow model by forcing the non-zero observations in the explanatory variable matrices $X_{d,t-1}, X_{o,t-1}$ to explain variation in the interregional flows.

This is accomplished using $\iota_i = \text{vec}(I_n)$ as a new intercept for the main diagonal elements of the flow matrix, and adjusting the intercept $\iota_N - \iota_i$ to have zeros for these observations. A separate intraregional model is constructed using an additional matrix of explanatory variables that we label $X_{i,t-1}$ to explain intraregional flows (perhaps regional size measures such as population and area which were used in our application). This is created using $X_{i,t-1} = \iota_i \odot (\iota_n \otimes [\text{pop}_t \quad \text{area}_t])$, where pop_t and area_t are $n \times 1$ vectors, and $\odot$ is the Hadamard or element-by-element matrix product. Only observations associated with the intraregional flows from the vector y contain non-zero values in the matrix $X_{i,t-1}$ and ι_i. The matrices X_d, X_o are adjusted to have zero elements for these observations, which can be formally expressed as: $X_d - \iota_i X_d, X_o - \iota_i X_o$.

Use of the separate model for intraregional commodity flows should downweight the impact of the large values on the main diagonal of the flow matrix, preventing them from exerting undue impact on the resulting estimates for $\beta_{d,t-1}\beta_{o,t-1}$. We interpret these parameters as reflecting the relationship of origin and destination characteristics on variation in interregional flows, and estimates for $\beta_{i,t-1}$ and the separate intercept that we denote α_i, represent the relationship between regional characteristics and intraregional flows.

One point to note is that there are only n non-zero observations in the matrix $X_{i,t-1}$, which might limit the number of explanatory variables that can be used for samples involving a small number of regions n. Plausible variables that should explain the magnitude of

intraregional flows would be: area of the region, population and gross regional product of the region. We would expect larger regions with more population and production to have higher levels of intraregional (within region) flows than smaller regions with less population and production.

The modified version of our model from (1) is presented in (2), where we have added the matrix $X_{i,t-1}$ and associated parameter vector β_i along with another intercept ι_i and associated parameter α_i. It should also be noted that the matrices of explanatory variables $X_{o,t-1}$, and $X_{d,t-1}$, and the intercept vector ι_N have been transformed so the values associated with observations representing intraregional flows are set to zero.

$$y_t = \alpha\iota_N + \alpha_i\iota_i + X_{d,t-1}\beta_d + X_{o,t-1}\beta_o + X_{i,t-1}\beta_i + \gamma d + \varepsilon_t \tag{2}$$

2.3 The Bayesian spatial autoregressive gravity model

LeSage and Pace (2008) point to the implausible nature of the assumption that O-D flows contained in the dependent variable vector y exhibit no spatial dependence. They note that the gravity model makes an attempt at modeling spatial relations between observations using the distance vector alone. If regions exert an influence on their neighbors as suggested by Griffith and Jones (1980), this might be inadequate. For example, neighboring origins and destinations may exhibit estimation errors of similar magnitude if underlying latent spatial forces are at work.

LeSage and Pace (2009, pp. 27-28) show that the presence of spatially dependent omitted variables that are uncorrelated with included variables produces a spatial error model specification, while (spatially dependent) omitted variables that are correlated with included variables leads to a spatial lag model specification. In addition, we note that forecasts for a spatial error model take the same form as those from the non-spatial least-squares model, and do not involve exploiting dependence between flows from nearby regions. Therefore, improved forecasts resulting from a spatial lag model specification would point to the lag model as the specification most consistent with the sample data.

There are several motivations for spatial dependence in variables omitted from spatial interaction models. For example, agents located at origins nearby in space may experience similar transport costs and profit opportunities when evaluating alternative destinations. Other motivations for spatial dependence in observed commodity flows might be common factor endowments or complemetary/competitive sectoral structures. For example, natural resource endowments are commonly thought to explain patterns of trade specialization. Demand and supply shocks should then have a similar impact on regions with similar resource endowments. Since many resource endowments are conditioned by space (e.g. arable land, natural resources, climate, common transport infrastructures, etc.), it seems plausible that nearby regions might exhibit similar patterns of sectoral specialization in production. As a concrete example for our sample of Spanish regions, trade flows of agricultural products originating in the Andalusia region of Spain are not likely to be independent of agricultural outflows from the Murcia region, since these two neighboring regions have many natural features in common as well as similar agricultural production specializations. Demand and supply factors that influence the agricultural flows from one of these two regions are likely to impact the flows from the other region in a similar fashion.

In addition to the example of positive spatial dependence above, there may also be negative spatial dependence arising from situations where two nearby regions compete by producing similar products for the national market. This would lead to a situation where increases in outflows of a specific product from one of the regions might be associated with decreased flows from the competing region. LeSage and Llano (2007) provide a hierarchical Bayesian model that contains spatially structured individual effects parameters that can be used to identify positive and negative origin/destination spatial effects. However, individual effects models are not well-suited for the task of forecasting.

We use a spatial autoregressive model to accommodate possible spatial dependence in commodity flows. In a typical cross-sectional model with n regions where each region represents an observation, spatial regression models rely on an n by n non-negative weight matrix that describes the connectivity structure between the n regions. For example, $W_{ij} >$

0 if region i is contiguous to region j. Besides contiguity, various measures of proximity such as cardinal or ordinal distance have been used to specify non-zero elements of the matrix W. By convention, $W_{ii} = 0$ to prevent an observation from being defined as a neighbor to itself, and the matrix W is typically standardized to have row sums of unity.

In the case of O-D flows, where we are working with $N = n^2$ observations, a key issue is how to construct a meaningful spatial weight matrix that describes connectivity between regions treated as origins and destinations. LeSage and Pace (2008) provide a solution by noting that $W_d = I_n \otimes W$, represents an N by N row-standardized spatial weight matrix that captures connectivity between regions viewed as destinations, and $W_o = W \otimes I_n$ produces another N by N row-standardized spatial weight matrix that captures connectivity between origin regions.[3] For our model of commodity flows we rely on a row-standardized matrix $W_c = W_d + W_o$, to form a *spatial lag* of the dependent variable, $W_c y$. This additional explanatory variable vector formed by multiplying the N by N matrix W_c and the N by 1 vector y is added to the model from (2), as shown in (3).

$$y_t = \alpha \iota_N + \alpha_i \iota_i + \rho W_c y_t + X_{d,t-1}\beta_d + X_{o,t-1}\beta_o + X_{i,t-1}\beta_i + \gamma d + \varepsilon_t \tag{3}$$

The additional explanatory variable captures both 'destination' and 'origin' based spatial dependence relations using an average of flows from neighbors to each origin and destination region. Intuitively, forces leading to flows from any origin to a particular destination region may create similar flows from neighbors to this origin to the same destination, a situation labeled origin-based dependence by LeSage and Pace (2008). This formally captures the point of (Griffith and Jones, 1980) that flows from an origin are "enhanced or diminished in accordance with the propensity of emissiveness of its neighboring origin locations". The spatial lag vector $W_c y$ also captures destination-based dependence reflecting the intuition that forces leading to commodity flows from a particular origin region to a destination region may create similar flows to nearby or neighboring destinations. This is the notion of Griffith and Jones (1980) that flows associated with a destination are "enhanced or diminished in

[3] We use the symbol $\otimes$ to denote a kronecker product.

accordance with the propensity of attractiveness of its neighboring destination locations".

We note that the model in (3) subsumes the non-spatial model as a special case when the scalar parameter ρ takes on a value of zero. This allows us to test for the presence of significant spatial dependence in the O-D flows. If there is spatial dependence, it has been shown that least-squares estimates are biased and inconsistent, which suggests that ignoring spatial dependence should result in less accurate forecasts.

A final problem that arises in modeling and forecasting commodity flows is the presence of outliers. These might appear as O-D flow observations that are aberrantly large of small, given values of the explanatory variables. These can exert a great deal of impact on the parameter estimates, which will in turn influence forecast accuracy. In order to downweight these aberrant observations, we rely on a robust variant of the spatial regression model set forth in LeSage (1997).

A formal statement of the *Bayesian heteroscedastic SAR model* is shown in (4), where we omitted the time subscript for simplicity. The model adds an *independent* normal and inverse-gamma prior for δ and σ^2, and a uniform prior for ρ. In addition, a chi-squared prior is used for a set of $N = n^2$ variance scalars.

$$\begin{aligned}
y &= \rho W_c y + Z\delta + \varepsilon \\
Z &= \begin{pmatrix} \iota_N & \iota_i & X_d & X_o & X_i & d \end{pmatrix} \\
\delta &= \begin{pmatrix} \alpha & \alpha_i & \beta_d & \beta_o & \beta_i & \gamma \end{pmatrix}' \\
\varepsilon &\sim N(0, \sigma^2 V) \\
V_{mm} &= v_m, m = 1, \ldots, N,\ V_{ml} = 0,\ m \neq l \\
\pi(\delta) &\sim N(c, T) \\
\pi(r/v_m) &\sim iid\ \chi^2(r), m = 1, \ldots, N \\
\pi(\sigma^2) &\sim IG(a, b) \\
\pi(\rho) &\sim U(1/\lambda_{\min}, 1/\lambda_{\max})
\end{aligned} \tag{4}$$

The variance scalars follow an approach introduced by Geweke (1993). These represent a set of latent variance parameters for each O-D flow dyad, allowing us to replace $\varepsilon \sim \mathrm{N}[0, \sigma^2 I_N]$, with:

$$\begin{aligned} \varepsilon &\sim \mathrm{N}[0, \sigma^2 V] \\ V &= \begin{pmatrix} v_{11} & 0 & \dots & 0 \\ 0 & v_{22} & & \\ \vdots & & \ddots & \vdots \\ 0 & & & v_{NN} \end{pmatrix} \end{aligned} \tag{5}$$

Estimates for the N variance scalars in (5), are produced using an *iid* $\chi^2(r)$ prior on each of the variance scalars v_{mm}, with a mean of unity and a mode and variance that depend on the hyperparameter r of the prior. Small values of r between 5 and 8 result in a prior that allows for the individual v_{mm} estimates to be centered on their prior mean of unity, but deviate greatly from the prior value of unity in cases where the model residuals are large. We used a prior setting of $r = 5$ for our implementation described in the next section. Watanabe (2001) provides a method for producing posterior inferences regarding this parameter, but there is evidence in other work that this approach seldom makes a difference in the estimates and inferences (Keith and LeSage 2004). Large residuals are indicative of outliers or origin-destination combinations that are atypical or aberrant relative to the majority of the sample of origin-destination flows. Geweke (1993) points to the equivalence of this modeling approach and the assumption of disturbances that follow a Student $t-$distribution. We rely on log-determinant calculations set forth in LeSage and Pace (2009) to implement a Bayesian model similar to that set forth in LeSage (1997).

Markov Chain Monte Carlo (MCMC) estimation is based on the idea that rather than work with the posterior density of our parameters, the same goal can be achieved using a large random sample from the posterior distribution. Let $p(\theta|D)$ represent the posterior, where θ denote the parameters and D the sample data. If the sample from $p(\theta|D)$ were

large enough, one could approximate the form of the probability density using kernel density estimators or histograms, eliminating the need to find the precise analytical form of the density. To implement this approach we sample sequentially from the complete set of conditional distributions for the parameters of the model.

We need the conditional posterior distributions for the parameters β, σ, and ρ as well as the variance scalars $v_{mm}, m = 1, \ldots, N$ in this model to implement an MCMC sampling scheme. The conditional distribution for β takes the form of a multivariate normal shown in (6).

$$\begin{aligned} p(\delta|\rho,\sigma,V) &\propto N(c^*,T^*) \\ c^* &= (Z'V^{-1}Z+\sigma^2T^{-1})^{-1}(Z'V^{-1}(I_N-\rho W_c)y+\sigma^2T^{-1}c) \\ T^* &= \sigma^2(Z'V^{-1}Z+\sigma^2T^{-1})^{-1} \end{aligned} \tag{6}$$

The expression needed to produce a draw from the conditional posterior distribution of σ^2 takes the form in (7).

$$\begin{aligned} p(\sigma^2|\delta,\rho,V) &\propto IG(a^*,b^*) \\ a^* &= a+N/2 \\ b^* &= (2b+e'V^{-1}e)/2 \\ e &= Ay-Z\delta \\ A &= I_N-\rho W_c \end{aligned} \tag{7}$$

The expression needed to produce draws for the parameter ρ takes the unknown distributional form shown in (8). LeSage and Pace (2009) provide details on two approaches that can be used to sample from this distribution, one that relies on numerical integration followed by a draw via inversion, and the second being a Metropolis-Hastings method.

$$p(\rho|\delta,\sigma^2,V) \propto \ |A| \ \exp\left(-\frac{1}{2\sigma^2}e'V^{-1}e\right) \tag{8}$$

Geweke (1993) shows that the conditional distribution of V given the other parameters is proportional to a chi-square density with $r+1$ degrees of freedom. Specifically, we can express the conditional posterior of each v_m as in (9), where $v_{-m} = (v_1, \ldots, v_{m-1}, v_{m+1}, \ldots, v_N)$ for each m. That is, we sample each variance scalar conditional on all others. The term e_m represents the mth element of the vector $e = Ay - Z\delta$.

$$p(\frac{e_m^2 + r}{v_m}|\delta,\rho,\sigma^2,v_{-m}) \propto \chi^2(r+1) \tag{9}$$

We produce robust least-squares estimates using this model by setting the spatial dependence parameter $\rho = 0$.

3 An application to commodity flows between Spanish regions

To illustrate our method, we analyze the C-intereg database that contains interregional commodity flows for 18 Spanish NUTS2 regions measured in Euros. A description of the main features of the dataset is provided in section 3.1. A procedure for producing forecasts with dependent sample data is proposed in section 3.2. Section 3.3 describes five alternative models used in our forecasting experiments and forecast accuracy results from comparing the alternative models.

3.1 The Data

The data used in this study for the intra-national flows corresponds to the estimates of Spanish intra and interregional trade, resulting from the C-intereg project. Currently, the C-intereg database contains estimates for the interregional trade of goods (30 categories) between 18 Spanish regions (NUTS2, Autonomous Communities), 52 NUTS-3 provinces,

and four transport modes for the period 1995-2010. For the sake of homogeneity, in this study we focus on the "aggregate flows" (with no sectoral breakdown) observed during period 1995-2004 and the 15 NUTS-2 regions located on the Iberia peninsula, without considering Balearic Islands, Canary Islands, and the autonomous cities of Ceuta and Melilla located in Africa. A detailed description of the methodology used for estimating this intra-regional and inter-regional flows is described with full detail in flows (Llano et al., 2010). Such estimates are based on the most accurate data on Spanish transport flows of goods by transport modes (road, rail, ship and plane) with additional information used to estimate specific export price vectors, one per each region of origin, transport mode and type of product. Finally, a process of harmonization is applied to produce flow magnitudes in tons and euros that are consistent with total output magnitudes from the Spanish Industrial Survey and the National Accounts.This novel database has been used in several papers modeling intra-national and inter-national flows with the aim of estimating the internal (home bias) and the external border effect in Spain at different spatial levels (Requena and Llano, 2010; Ghemawat et al, 2010; Llano-Verduras et al, 2011; Garmendia et al, 2012). However it has not been used yet with forecasting purposes.

Apart from the information on the Spanish interregional trade, an additional database of variables reflecting regional characteristics was constructed for every region and year using information from the Spanish National Institute of Statistics (INE) and the REGIO Database from Eurostat.

Following classical applications of gravity models to bilateral trade flows, explanatory variables meant to reflect size of the origin and destination regions were used in $X_{o,t}, X_{d,t}$. Specifically, population (**pop**) and gross domestic product (**gdp**) were included in the set of destination size explanatory variables to reflect "pull factors", but (**pop**) and gross value added (**gva**) were used in the set of origin size explanatory variables to reflect "push factors".

Variables selected for the matrix $X_{i,t}$ that capture intraregional flows were population (**pop**) and the square meter area of the region (**area**). These along with the additional intercept term were adjusted to reflect non-zero values for only observations corresponding

to the main diagonal of the flow matrix as described earlier. A similar statement applies to population variable used in the set of origin and destination explanatory variables. Finally, the (**distance**) between regions constructed as a vectorized version of the O-D distance matrix. The sources and definitions of these variables are described in Table 1.

Table 2 shows summary statistics for the (logged) flows and explanatory variables over the period from 1997 to 2007. A single zero flow was present for the years 1996, 1997, and 2006, with no zero flows for the other years 1998-2005 and 2007. This lack of zero flows at the regional (NUTS2) level reflect that the intensity of trade between regions within a country appears much higher than between countries likely due to border effects. The summary indicates that minimum values for logged flows shows consideration variation over time whereas the maximum (logged) flows ranged between 24.1 and 24.5 over the years. Logged flows as well as population and GVA exhibit the expected right-skew, and population and GVA exhibit upward trending growth over time.

3.2 A forecasting procedure for dependent data

For each year over the period 1996-2003, the model from (3) was estimated using y_t, $X_{o,t-1}$, $X_{d,t-1}$ and $X_{i,t-1}$ with all variables in log form. For simplicity variables used in the explanatory variables matrices were keep the same for all years.

Using the one-period lag relationship between (logged) O-D flows and the explanatory variables one-period-ahead, a least-squares forecast can be computed using:

$$\hat{y}_{t+1} = \hat{\alpha}\iota_N + \hat{\alpha}_i\iota_i + X_{d,t}\hat{\beta}_d + X_{o,t}\hat{\beta}_o + X_{i,t}\hat{\beta}_i + d\hat{\gamma} \tag{10}$$

The gold standard for predicting a dependent variable using variable values at all observations and conditioning on known values of the dependent variable for other dependent observations is Best Linear Unbiased Prediction (BLUP) (Goldberger, 1962). Intuitively, conditioning on sample data values for the dependent variable and the covariance relationship implied by the spatial dependence structure of observations on those from other locations should aid prediction accuracy. Past work has focused on prediction of dependent

data, but not forecasting. For example, LeSage and Pace (2004) consider predicting missing house values in a real estate application, and Pace and LeSage (2008) provide computationally efficient approaches to producing predications for spatial regression models. We note that this work points out that use of the reduced form expression for (3) will not produce best linear unbiased predictions.

Here, we develop a forecasting scheme that takes advantage of the covariance structure between spatially dependent observations as well as the time lag imposed during estimation. For simplicity of exposition, we rewrite the spatial lag (SAR) model as in (11), where we note that use of $A = (I_N - \rho_d W_d - \rho_o W_o)$ would produce forecasts from the more general model which does not combine the spatial weight matrices. Since the focus of this study is whether: 1) the adjustment for large flows on the diagonal of the flow matrix (intraregional flows), and 2) spatial dependence in flows, can be used to improve forecast accuracy relative to conventional least-squares forecasts, we did not pursue the more elaborate model involving two weight matrices. An assessment of gains from use of the more flexible spatial lag structure involving two weight matrices is a topic for future research.

$$\begin{aligned} y_{t+1} &= A^{-1} Z_t \delta + \varepsilon_t \\ A &= (I_N - \rho W_c) \\ Z_t &= \begin{pmatrix} \iota_N & \iota_i & X_{d,t} & X_{o,t} & X_{i,t} & d \end{pmatrix} \\ \delta &= \begin{pmatrix} \alpha & \alpha_i & \beta_d & \beta_o & \beta_i & \gamma \end{pmatrix}' \end{aligned} \tag{11}$$

We partition the model in (11) into two parts, one involving flow i from the vector of flows y_{t+1}, and the other involving all other flows $j \neq i$ in the vector. This requires partitioning the matrix A into $A_{ii}, A_{ij}, A_{ji}, A_{jj}$, as shown in (12).

$$A^{-1} = \begin{pmatrix} A^{ii} & A^{ij} \\ A^{ji} & A^{jj} \end{pmatrix} \tag{12}$$

where A^{ii} is a scalar, A^{ij} is a $1 \times N-1$ vector consisting of elements $j \neq i$ from the ith row of the matrix A^{-1}, A^{ji} is an $N-1 \times 1$ vector containing elements $j \neq i$ from the ith column of A^{-1}, and A^{jj} is an $N-1 \times N-1$ matrix. The non-singular matrix A need not be symmetric, but $\Omega = \sigma_\varepsilon^2(A'A)^{-1}$ is symmetric and positive definite.[4]

Given the partitioned model format, we forecast observation i conditional on all other observations $j \neq i$ using the expression in (13), where $A_{ii}^{-1}Z_{it}\hat{\delta}$ denotes the conditional mean of the forecast for observation i, and $(y_{j,t} - A_{jj}^{-1}Z_{jt-1}\hat{\delta})$ are the residuals from the previous period for observations $j \neq i$.

$$\hat{y}_{i,t+1} \quad = \quad A_{ii}^{-1}Z_{it}\hat{\delta} + \Omega_{ij}(\Omega_{jj})^{-1}(y_{j,t} - A_{jj}^{-1}Z_{jt-1}\hat{\delta}) \tag{13}$$

Our procedure for producing a forecast for observation i at time $t+1$ involves a number of operations. First, it takes the prediction errors for observations $j \neq i$ made by the model during the previous time period, $(y_{j,t} - A_{jj}^{-1}Z_{jt-1}\hat{\delta})$, and removes the covariance between these observations and observation i that we wish to forecast. This is accomplished through multiplication by the part of the inverse variance-covariance matrix that pertains to observations $j \neq i$, Ω_{jj}^{-1}. Second, the procedure reimposes the pattern of dependence appropriate for the forecast observation i using the multiplication involving the partition from the variance-covariance matrix that relates observation i and observations $j \neq i$, Ω_{ij}. The forecasted value $\hat{y}_{i,t+1}$, then uses these new dependent prediction errors from the previous time period to improve on the simple conditional mean prediction $A_{ii}^{-1}Z_{it}\hat{\delta}$. In other words, our forecast uses the previous period prediction errors from observations j (all observations other than i) by translating these through use of the spatial dependence structure to reflect dependence with observation i that we wish to forecast.

It should be clear that for cases where we have no spatial dependence so that $\rho = 0$, which implies $A = A^{-1} = I_N$, and $\Omega = \sigma_\varepsilon^2 I_N$, our forecast would collapse to that from the

[4]For the robust model $A = (I_N - \rho W_c)V^{1/2}$ and $\Omega = \sigma_\varepsilon^2(A'V^{-1}A)^{-1}$, where V is a diagonal matrix containing the variance scalar estimates.

least-squares model. (Of course, the adjustments described to produce a separate model for the intraregional flows would need be implemented.) This suggests that the greatest gains in forecast accuracy would occur for situations involving high levels of spatial dependence, since our forecast then rely more heavily on flows from nearby regions.

The models are estimated using sample data for the year 1995 and 1996 to produce coefficient estimates based on the one-year lagged relationship between flows and regional characteristics which are used to produce a one-year-ahead forecast for 1997. The models are then re-estimated using sample data for 1996 and 1997 to produce a forecast for 1998, and so on, with the last forecast for 2004. This procedure represents a real-time forecasting experiment that reflects how the models would be used in actual forecasting practice.

To provide an assessment of the introduction of variables $X_{i,t-1}$ and intercept ι_i to separate out within-region versus between-region flows during estimation, we considered the least-squares model that excludes the explanatory variable $X_{i,t-1}$ and intercept ι_i from the model, and does not set observations in the explanatory variable matrices X_o, X_d to zero for the within-region flows. This is the conventional specification and estimation approach for gravity models.

3.3 Forecasting Accuracy Results

In this section we analyze forecasting accuracy results from the five different specifications and estimation schemes for the model (SAR robust; SAR non-robust; OLS-robust; OLS non-robust; OLS conventional model).

Forecasting accuracy of the five models was compared using Mean Absolute Percentage Error (MAPE) and the Percentage Mean Squared Error (PMSE), calculated using conventional formulas shown in (14) and (15). We use $y_{i,ft+1}$ to denote the ith forecasted flow and $y_{i,at+1}$ to represent the ith actual flow magnitude in (14) and (15).

Mean squared error penalizes larger forecast errors quadratically whereas mean absolute error treats these in a linear fashion. Which measure one uses to analyze forecast accuracy would depend on the forecaster's loss function.

$$MAPE = 100 \times \sum_{i=1}^{N} \left[\frac{|y_{i,ft+1} - y_{i,at+1}|}{y_{i,at+1}} \right] /N \tag{14}$$

$$PMSE = 100 \times \sum_{i=1}^{N} \left[\frac{y_{i,ft+1} - y_{i,at+1}}{y_{i,at+1}} \right]^2 /N \tag{15}$$

The forecast accuracy comparisons allow an examination of three issues. First, there is the question of the role played by the adjustment proposed by LeSage and Pace (2008) to produce a separate model to explain large variation in the main diagonal elements of the flow matrix. The goal of this adjustment is to produce estimates that more adequately model interregional flow variation, which could improve forecast accuracy. Second, there is the question of whether downweighting aberrant observations during estimation improves forecast accuracy. The third issue is whether spatial dependence in flows (which result in a multivariate distribution for flows), can be exploited to improve forecast accuracy. The multivariate distribution arising from spatial dependence allows us to exploit a decomposition provided by the multivariate normal distribution when producing forecasts. Forecasts for each observation are conditional on the mean prediction/forecast plus any covariance with flows from nearby observations.

The difference in forecast accuracy between the conventional OLS model and the OLS non-robust provides information regarding the usefulness of the strategy for modeling inter- and intra-regional flows separately. This is because the conventional OLS model from (1) does not make adjustments to the intercept term or explanatory variables for the purpose of producing a separate model for intraregional flows, whereas the model we label OLS non-robust does this.

Differences in accuracy between the OLS non-robust and OLS robust models as well as between the SAR non-robust and SAR robust provide information regarding the role played by outlier observations that impact the model parameter estimates. The robust parameter estimates of the same model specification should differ only because of aberrant observations that are downweighted by the robust estimation procedures.

Finally, differences between the SAR robust and non-robust and the OLS robust and non-robust point to the importance of taking spatial dependence into account. This of course depends on the magnitude of spatial dependence in the flow data, with highly dependent data leading to a greater advantage for the spatial models and very low levels of dependence resulting in no big differences between the OLS and SAR models. A theoretical upper bound on ρ is unity, so values closer to one reflect higher levels of spatial dependence while values closer to zero a lack of dependence.

The estimated parameters ρ reflecting the strength of spatial dependence for each time period along with the PMSE and MAPE errors and standard deviations are reported in Table 3. Accuracy comparisons were made based on the one-year ahead forecast of the 225 flows carried out for the period 1997 to 2004. In cases where the mean errors are the same, a forecaster with a symmetric loss function would pick the forecasting method with the smallest standard deviation in the forecast errors. An asterisk symbol $\star$ was used to denote the best of the five models for each year and each accuracy criterion.

In our application, the levels of spatial dependence measured by estimates for the parameter ρ were low, ranging from 0.19 to 0.33 for the robust model and from 0.19 to 0.34 for the non-robust model. Nonetheless, all coefficient estimates for ρ were statistically significantly different from zero based on the reported standard deviations.

The first point to note is that the conventional model that did not adjust for intraregional flows produced the largest forecast errors relative to the four models that adjusted for intraregional flows during all years. This was true using either the PMSE or MAPE accuracy criterion.

Asterisk symbols were used to indicate the best model for each year in the table. This makes it clear that one of the two spatial models produced the best forecast in all years, using either the PMSE or MAPE accuracy measures. For the PMSE accuracy criterion, the robust SAR model produced the most accurate forecasts in five of the eight years, with the non-robust SAR model winning in the other three years. Using the MAPE measure of forecast accuracy, the robust SAR model performed better than the non-robust in six of

the eight years. From this we conclude that modeling spatial dependence and using this in the forecasting procedure improves forecast accuracy. The last two rows of the table show the PMSE and MAPE cumulated over the eight years. These point to an average improvement in PMSE accuracy over the conventional least-squares model equalling 1.4 percent per year, $((70.5 - 59)/8 = 1.43)$, and around 0.2 percent per year for the MAPE measure, $((16.85 - 15.23)/8 = 0.2)$.

Another focus of the forecasting experiments was the value of using robust estimation methods that downweight outliers on forecast accuracy. If we consider the cumulative measures of accuracy for all eight years, the robust spatial model produced a small but inconsequential improvement in both PMSE and MAPE accuracy over the non-robust spatial model. Similarly, the least-squares models produced results indicating no large differences between models based on robust and non-robust estimates.

The last issue is that of the value of using a model that separates out intra- and interregional flows, by providing separate explanatory variables and parameters for these two types of flows. The two columns showing OLS results not labeled *conventional OLS* both use the separate model for intraregional flows. Using the PMSE and MAPE measures cumulated over the eight years, we see an improvement in PMSE forecast accuracy from around 70 for the conventional model to 65.31 and 66.48 for the non-robust and robust OLS models. For the MAPE criterion there is also an improvement from 16.85 to 16.15.

4 Conclusions

Gravity interaction models have traditionally relied on least-squares estimation methods, ignoring the issue of spatial dependence between interregional flows. We propose a forecasting methodology for a spatial regression model extension of the classical gravity model proposed by LeSage and Pace (2008). Their spatial regression approach also includes a procedure for dealing with the presence of large diagonal elements in the flow matrix that relate to within or intraregional flows.

We tested the forecasting accuracy of five alternative model specifications that allow us

to analyze the marginal contribution of: 1) taking into account spatial dependence between origin-destination flows; 2) the LeSage and Pace (2008) procedure for dealing with the presence of large diagonal elements in the flow matrix using a separate model for these; and 3) the importance of using a robust estimation method that downweights outlying or aberrant observations.

Using both percent mean-square error (PMSE) and mean absolute percent error (MAPE) accuracy measures, the procedure for dealing with the presence of large diagonal elements in the flow matrix produced a consistent improvement in forecast accuracy over the conventional least-squares model.

We find no overwhelming evidence that use of robust models produce more accurate forecasts than non-robust. Evidence in favor of the robust models is a bit stronger for the case of spatial versus non-spatial models, but the difference in accuracy appears inconsequential.

To our knowledge, this is the first study of forecasting accuracy for spatial interaction models that rely on a spatial autoregressive specification that models spatial dependence in flow magnitudes. Griffith (2007) makes the point that alternative approaches to estimation of these recently proposed models have produced a revival of interest in spatial variants of gravity models.

One important area for future exploration would be a forecasting procedure for cases where the flows represent count magnitudes, requiring Poisson or Negative Binomial estimation procedures. Lambert, Brown and Florax (2010) set forth a two-step maximum likelihood estimation procedure for a spatial autoregressive Poisson model, which would need to be extended to the case of flows involving $N = n^2$ observations.

Given our results which pointed to low levels of spatial dependence in the flow magnitudes, it would be of interest to explore situations where the sample data exhibited higher levels of dependence and more variation in dependence as well as a larger number of forecasting experiments. This would allow a more extensive exploration of the relationship between forecast accuracy and levels of spatial dependence.

References

Anderson, J. E. and E. van Wincoop (2003) "Gravity with Gravitas: A Solution to the Border Puzzle", *American Economic Review*, 93:1, 17092.

Chen, N. (2004), "Intra-national versus International Trade in the European Union: Why Do National Borders Matter?", *Journal of International Economics*, 63:1, 93118.

Combes, P. P., Lafourcade, M., Mayer, T. (2005) "The trade-creating effects of business and social networks: evidence from France." *Journal of International Economics*, Volume 66:1, 1-29.

Evans, C. L. (2003) "The Economic Significance of National Border Effects", *American Economic Review*, 93:4, 1291312.

Fischer, M.M. and D.A. Griffith (2008) "Modeling spatial autocorrelation in spatial interaction data: An application to patent citation data in the European Union," *Journal of Regional Science* 48, 969-989.

Fischer, M.M., T. Scherngell, and E. Jansenberger (2006) "The Geography of Knowledge Spillovers between High-Technology Firms in Europe Evidence from a Spatial Interaction Modelling Perspective," *Geographical Analysis*, 38:3, 288-309.

Garmendia, A., Llano-Verduras, C. and Requena-Silventre, F. (2012) "Network and the disappearance of the intranational home bias" *Economic Letters*, 116, 178-182.

Geweke, J. (1993) "Bayesian Treatment of the Independent Student t Linear Model." *Journal of Applied Econometrics*, 8, 19-40.

Ghemawat, P., Llano-Verduras. C., Requena-Silventre, F., (2010). "Competitiveness and interregional as well as and international trade: The case of Catalonia", *International Journal of Industrial Organization*, 28, 415-422

Gil, S., Llorca, R., Martnez, J.A. y Oliver, J. (2005) "The Border Effect in Spain", *The World Economy*, 28, 1617-1631

Goldberger, A. (1962) "Best Linear Unbiased Prediction in the Linear Model," *Journal of the American Statistical Association*, 57, 369-375.

Griffith, D.A. (2007) "Spatial Structure and Spatial Interaction: 25 Years Later," *The Review of Regional Studies* 37:1, 28-38.

Griffith, D.A., Jones K. (1980) "Explorations into the Relationship Between Spatial Structure and Spatial Interaction," *Environment and Planning A*, 12:2, 187201.

Grimpe C. and R. Patuelli (2011) "Regional knowledge production in nanomaterials: A spatial filtering approach, *Annals of Regional Science* 46, 519-541.

Helliwell, J. F. (1996) "Do National Borders Matter for Quebecs Trade?", *Canadian Journal of Economics*, 29:3, 50722.

Helliwell, J. F. (1998) *How Much Do National Borders Matter?* (Washington, DC: Brookings Institution Press).

Keith, K. and J. P. LeSage (2004) "Robust decomposition analysis of wage differentials," *Journal of Economic and Social Measurement* 29:4, 487-505.

Lambert, D. M., J. P. Brown and R. J. G. M. Florax (2010) "A two-step estimator for a spatial lag model of counts: Theory, small sample performance and an application," *Regional Science and Urban Economics*, 40:4, 241-252.

LeSage, J.P. (1997) "Bayesian Estimation of Spatial Autoregressive Models," *International Regional Science Review*, 20:1/2, 113-129.

LeSage, J. P. and M. M. Fischer (2010) "Spatial Econometric Modeling of Origin-Destination Flows," in *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, M. M. Fischer and A. Getis (Eds.) Springer, pp. 409-433.

LeSage, J.P., C. Llano (2007) "A Spatial Interaction Model With Spatial Structured Origin and Destination Effects". SSRN working paper, available at SSRN: http://ssrn.com/abstract=924603

LeSage, J.P. and R.K. Pace (2004) "Conditioning upon All the Data: Improved Prediction via Imputation," *Journal of Real Estate Finance and Economics*, 29, 233-254.

LeSage, J.P. and R.K. Pace (2008) "Spatial econometric modeling of origin-destination flows," *Journal of Regional Science*, 48:5, 941-967.

LeSage, J.P. and R.K. Pace (2009) *Introduction to Spatial Econometrics*, New York: Taylor and Francis/CRC Press.

Llano C., Esteban, A, Pérez, J., Pulido, A. (2010) "Opening the interregional trade black box: The C-Intereg Database For The Spanish Economy (1995-2005)". *International Regional Science Review*, 33:3, 302-337.

Llano-Verduras C.; Minondo A., Requena-Silvente F. (2011). "Is the Border Effect an Artefact of Geographical Aggregation?" *The World Economy*, 34, 10, 1771-1787.

McCallum, J. (1995) "National Borders Matter: Canadian-U.S. Regional Trade Patterns", *American Economic Review*, 85:3, 61523.

Nitsch, V. (2000) "National Borders and International Trade: Evidence from the European Union", *Canadian Journal of Economics*, 33:4, 10911105.

Okubo, T. (2004) "The Border Effect in the Japanese Market: A Gravity Model Analysis", *The Japanese and International Economies*, 18, 111.

Pace, R. K. and J. P. LeSage,(2008) "Spatial Econometric Models, Prediction," in *Encyclopedia of Geographical Information Science*, Shashi Shekhar and Hui Xiong (eds.), Springer-Verlag.

Requena, F. and Llano, C. (2010) "The border effects in Spain: an industry level analysis", Empirica, 37, pp. 455-476.

Scherngell, R. and R. Lata (2011) "Towards an integrated European Research Area? Findings from Eigenvector spatially filtered spatial interaction models using European Framework Programme data," (forthcoming in) *Papers in Regional Science*.

Sen, A. and T. E. Smith (1995) *Gravity Models of Spatial Interaction Behavior*, Heidelberg: Springer-Verlag.

Tiefelsdorf, M., (2003) "Misspecifications in interaction model distance decay relations: A spatial structure effect," *Journal of Geographical Systems*, 5, 25-50.

Watanabe, T. (2001) "On sampling the degree-of-freedom of Student's-t disturbances," *Statistics & Probability Letters*, 52, 177-181.

Wei, S. (1996) "Intra-national versus International Trade: How Stubborn are Nations in Global Integration?" NBER Working Paper 5531.

Wolf, H. (2000) "Intra-national Home Bias in Trade", *The Review of Economics and Statistics* 82:4, 555-563.

Table 1: Description and source of the explanatory variables used

Variable	Description	Source
area	Surface by regions NUTS2 in squared Km.	INE
pop	Population by regions NUTS2.	INE
gdp	Gross domestic product (GDP).	INE
gva	Gross value added (GVA).	INE
distance	Actual distance in Km traveled by heavy trucks (freight flows).	M.Fomento

Table 2: Summary statistics for the dependent and explanatory variables used

Variables	Mean	Median	Std	min	max
1996 log(flows)	19.5	19.8	2.0	0.0	24.1
1996 gva	15142.5	9212.0	13815.5	2365.4	53299.9
1996 pop	2478273.9	1712529.0	2159928.7	264941.0	7234873.0
area	32901.7	23260.0	31173.2	5045.0	94225.0
1997 log(flows)	19.6	19.8	2.0	0.0	24.1
1997 gva	16330.7	10063.1	14889.0	2602.7	57441.0
1997 pop	2486261.9	1716152.0	2172542.2	263644.0	7236459.0
1998 log(flows)	19.8	20.0	1.6	14.6	24.2
1998 gva	16861.5	10497.1	15411.2	2722.3	59484.3
1998 pop	2505134.5	1726199.0	2195214.3	265178.0	7305117.0
1999 log(flows)	19.8	19.9	1.8	6.4	24.2
1999 gva	17584.9	11051.7	16011.0	2916.6	61685.2
1999 pop	2519758.7	1734261.0	2214176.2	264178.0	7340052.0
2000 log(flows)	19.9	20.1	1.6	13.7	24.3
2000 gva	18622.8	11853.9	16963.8	3086.3	65207.7
2000 pop	2554157.7	1755053.0	2250462.4	270400.0	7403968.0
2001 log(flows)	19.9	20.1	1.6	12.1	24.4
2001 gva	20062.7	13026.4	18033.8	3350.4	69324.1
2001 pop	2595455.7	1782038.0	2295043.5	281614.0	7478432.0
2002 log(flows)	20.0	20.1	1.6	15.6	24.4
2002 gva	21413.5	13810.1	19406.9	3557.9	74544.4
2002 pop	2648762.7	1815781.0	2359386.8	287390.0	7606848.0
2003 log(flows)	20.0	20.1	1.5	16.3	24.4
2003 gva	22348.8	14540.5	19943.9	3627.0	76462.5
2003 pop	2678961.9	1848881.0	2392957.3	293553.0	7687518.0
2004 log(flows)	20.1	20.2	1.6	14.6	24.4
2004 gva	23604.5	15544.1	20909.7	3864.1	79614.1
2004 pop	2734423.7	1894667.0	2456955.6	301084.0	7849799.0
2005 log(flows)	20.1	20.3	1.6	14.5	24.5
2005 gva	24973.0	16434.7	22051.6	4042.9	83620.6
2005 pop	2771289.1	1932261.0	2499410.0	306377.0	7975672.0
2006 log(flows)	20.1	20.4	2.0	0.0	24.5
2006 gva	26852.4	17748.1	23568.5	4244.0	88394.9
2006 pop	2734423.7	1894667.0	2456955.6	301084.0	7849799.0
2007 log(flows)	20.3	20.5	1.5	15.3	24.5
2007 gva	28483.9	17991.9	25077.7	4548.2	94018.2
2007 pop	2771289.1	1932261.0	2499410.0	306377.0	7975672.0

Table 3: A comparison of one-year-ahead forecast errors

	SAR Robust	SAR non-robust	OLS robust	OLS non-robust	OLS conventional
1997					
PMSE (std)	8.59* (35.03)	8.71 (34.14)	9.82 (38.56)	9.66 (35.57)	10.31 (36.54)
MAPE (std)	1.99* (2.15)	2.01 (2.16)	2.12 (2.31)	2.14 (2.25)	2.21 (2.33)
$\hat{\rho}$ (std)	0.24 (0.065)	0.24 (0.071)			
1998					
PMSE (std)	5.10 (9.84)	5.08* (9.52)	6.50 (12.93)	6.24 (11.78)	6.92 (12.41)
MAPE (std)	1.67* (1.51)	1.69 (1.48)	1.94 (1.66)	1.90 (1.61)	2.01 (1.69)
$\hat{\rho}$ (std)	0.29 (0.065)	0.30 (0.071)			
1999					
PMSE (std)	7.80 (25.71)	7.54* (25.67)	8.71 (31.32)	8.63 (30.32)	9.37 (31.58
MAPE (std)	1.91 (2.03)	1.85* (2.02)	1.99 (2.18)	2.00 (2.15)	2.08 (2.24)
$\hat{\rho}$ (std)	0.33 (0.058)	0.34 (0.062)			
2000					
PMSE (std)	8.90* (37.47)	8.93 (36.25)	9.73 (41.88)	9.62 (39.60)	10.22 (40.69)
MAPE (std)	2.01* (2.20)	2.03 (2.19)	2.09 (2.31)	2.11 (2.27)	2.20 (2.32)
$\hat{\rho}$ (std)	0.25 (0.063)	0.25 (0.071)			
2001					
PMSE (std)	8.45* (32.49)	8.76 (39.26)	9.14 (41.06)	9.04 (39.87)	9.73 (41.35)
MAPE (std)	1.93 (2.17)	1.90* (2.26)	1.97 (2.29)	1.98 (2.26)	2.05 (2.35)
$\hat{\rho}$ (std)	0.34 (0.069)	0.30 (0.070)			
2002					
PMSE (std)	6.13* (10.56)	7.02 (12.50)	6.96 (12.66)	6.83 (11.69)	7.46 (12.31)
MAPE (std)	1.89* (1.60)	1.99 (1.75)	1.97 (1.75)	1.98 (1.71)	2.04 (1.81)
$\hat{\rho}$ (std)	0.30 (0.056)	0.26 (0.066)			
2003					
PMSE (std)	6.76 (17.02)	6.66* (16.34)	7.69 (17.41)	7.57 (17.10)	8.14 (17.31)
MAPE (std)	1.91* (1.76)	1.92 (1.72)	2.02 (1.90)	2.01 (1.87)	2.12 (1.91)
$\hat{\rho}$ (std)	0.19 (0.070)	0.19 (0.070)			
2004					
PMSE (std)	7.21* (19.18)	7.22 (19.37)	7.92 (19.71)	7.73 (19.05)	8.32 (19.50)
MAPE (std)	1.92* (1.87)	1.93 (1.87)	2.05 (1.93)	2.03 (1.89)	2.14 (1.93)
$\hat{\rho}$ (std)	0.27 (0.071)	0.26 (0.073)			
$\sum_{i=1}^{8}$ PMSE	58.95*	59.93	66.48	65.31	70.47
$\sum_{i=1}^{8}$ MAPE	15.23*	15.32	16.15	16.15	16.85