A comparison of singly-constrained and spatially structued random effects gravity model specifications with and without eigenfunction spatial filters: 2000 journey-to-work flows in Pennsylvania

Daniel A. Griffith
Ashbel Smith Professor
School of Economic, Political and Policy Sciences, University of Texas at Dallas, 800 W. Campbell Road, Richardson, Texas 75080-3021, USA, dagriffith@utdallas.edu

Abstract. A resurgence of interest is occurring concerning the role spatial dependency effects play in spatial interaction flows models. Gravity model specifications studied to date include: a spatial lag formulation with a log-normal probability model approximation with no constraints, and an eigenvector spatial filter formulation with a Poisson probability model, both with and without being doubly-constrained. A third possible specification includes a random effects term for the origin or for the destination areal units, and links directly to the two possible singly-constrained gravity model formulations. This paper summarizes findings from an analysis of these latter specifications. Results are illustrated with the 2000 journey-to-work dataset for the state of Pennsylvania, USA.

KEY WORDS: balancing factor, eigenfunction spatial filter, gravity model, spatial autocorrelation, spatial interaction

1. Introduction

The simple gravity model (e.g., Sen and Smith, 1995) motivated its singly-constrained versions containing either origin or destination balancing factors (e.g., Wilson, 1967), for which Poisson regression can be used to estimate its parameters (e.g., Flowerdew and Aitkin, 1982). This Poisson probability model specification led to the use of either origin or destination indicator variables (a separate indicator variable is included for each origin and each destination—i.e., n 0-1 binary variables, each having a single 1 and n-1 0s), whose sets of coefficients are equivalent to the logarithm of the balancing factors when amalgamated (i.e., the respective sets of coefficients are concatenated into an n-by-1 vector). Today, attention is focusing on the role spatial autocorrelation plays in this model (e.g., Chun, 2007, 2008; Griffith, 2007, 2009a,b; Fischer and Griffith, 2008; LeSage and Pace, 2008). The purpose of this paper is twofold: (1) to introduce a singly-constrained spatial interaction model specification that accounts for the spatial autocorrelation in flows; and, (2) to compare this with a specification that includes a random effects term in place of the n fixed effects indicator variables.

The conceptualization described in this paper is illustrated with the 2000 journey-to-work dataset for the state of Pennsylvania, USA. Pennsylvania is partitioned into 67 counties, to which 6 additional counties from neighboring states surrounding Philadelphia are included (one from Delaware and five from New Jersey; see Klios, 2004), resulting in 5,329 flow dyads (of which 2,359 had a 0 flow in 2000). The total journey-to-work flows for 2000 by county total 6,287,022 (of which 5,381,466 originate in Pennsylvania, 671,371 originate in New Jersey, and 234,185

originate in Delaware). Inter-county distance is calculated with the geometric centroids of these counties.

2. The singly-constrained gravity model: a Poisson specification that accounts for spatial autocorrelation

An expanded singly-constrained gravity model may be written as

$$\text{origin-constrained: } F_{ij} = O_i \exp(\alpha + \sum_{h=1}^{n-1} I_{i,h,o}\beta_{h,o} - \gamma d_{ij} + \rho_{ij}) \text{ , or} \quad (1)$$

$$\text{destination constrained: } F_{ij} = D_j \exp(\alpha + \sum_{k=1}^{n-1} I_{i,k,d}\beta_{k,d} - \gamma d_{ij} + \rho_{ij}) \text{ ,} \quad (2)$$

where F_{ij} is the journey-to-work flow between origin (county) i and destination (county) j,

O_i and D_j are the total number, respectively, of workers in origin i and of jobs in destination j that travel within the system[1], and enters the Poisson regression model specification as the offset variable $LN(O_i)$ or $LN(D_j)$,

$\exp(\beta_{h,o})$ and $\exp(\beta_{k,d})$ respectively are the balancing factors for origin h (ensuring that the predicted number exactly equals the observed number of workers leaving each origin h) and for destination k (ensuring that the predicted number exactly equals the observed number of workers arriving in each destination k), where the n^{th} areal unit is assigned $\beta_{n,o}$ and $\beta_{n,d}$ values of 0 (to avoid perfect multicollinearity in the specification)— this value is absorbed into the intercept term,

d_{ij} is the distance separating origin (county) i and destination (county) j (measured here between county centroids),

ρ_{ij} is the spatial autocorrelation term representing dependencies between flows from nearby counties of i to nearby counties of j (Figure 1), and

$\exp(\alpha)$ and γ respectively are the constant of proportionality and the distance decay parameters.

The balancing factors $\exp(\beta_{i,o}) = A_i = \dfrac{1}{\sum_{j=1}^{n} D_j \exp(\gamma d_{ij})}$ and $\exp(\beta_{j,d}) = B_j = \dfrac{1}{\sum_{i=1}^{n} O_i \exp(\gamma d_{ij})}$

respectively ensure that $\sum_{i=1}^{n} \hat{F}_{ij} = O_i$ and that $\sum_{j=1}^{n} \hat{F}_{ij} = D_j$, where $\hat{F}_{ij}$ denotes the predicted flow between counties i and j. Meanwhile, Curry (1972) argues that inclusion of the spatial autocorrelation term ρ_{ij} results in a substantial change in the estimate of γ.

[1] Of all Pennsylvania journey-to-work flows, 0.69% is external to the state. This is not the case for the included Delaware and New Jersey counties, where only those parts of these states in the catchment area of Philadelphia are included in this analysis.

3. Spatial filtering: a brief overview

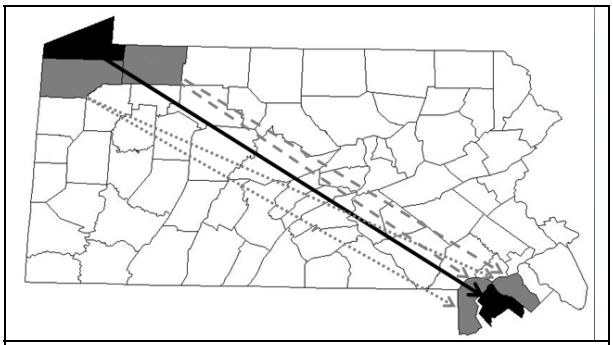

Figure 1. An example set of correlated spatial interaction flows. The directional flow of interest is denoted by the solid black arrow. Those directional flows correlated with it are denoted by the dotted gray arrows.

Spatial filtering, which allows positive spatial autocorrelation to be accommodated in a Poisson model specification, furnishes an approach that can account for spatial autocorrelation (see Griffith, 2002) in flows data described by a Poisson probability model specification. This result is achieved by constructing a linear combination of a subset of the eigenvectors of a modified geographic weights matrix—a spatial filter; this subset can be determined with, for example, backward selection stepwise Poisson regression based upon statistical significance. An individual eigenvector represents a global, regional, or local map pattern of geographic dependency (Figure 2).

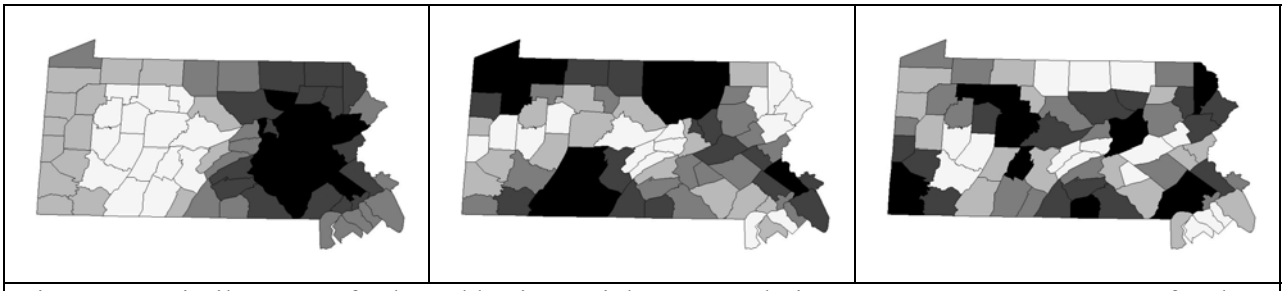

Figure 2. Quintile maps of selected basic spatial autocorrelation component map patterns for the Pennsylvania journey-to-work county map; values directly proportional to darkness of grayscale. Left (a): global (E_1: $MC/MC_{max} = 1.00$, $MC_{max} = 1.06$). Middle (b): regional (E_{11}: $MC/MC_{max} = 0.52$). Right (c): local (E_{18}: $MC/MC_{max} = 0.27$).

The modified geographic weights matrix for a given surface partitioning, say $(\mathbf{I} - \mathbf{11}'/n)\mathbf{C}(\mathbf{I} - \mathbf{11}'/n)$—where $\mathbf{I}$ is an n-by-n identity matrix, $\mathbf{1}$ is an n-by-1 vector of 1s, and $'$ denotes the matrix transpose operation—from which eigenfunctions are extracted appears in the numerator of the Moran Coefficient (MC) spatial autocorrelation test statistic. The eigenvectors of this matrix exhibit the following property when mapped: the first eigenvector, say $\mathbf{E}_1$, is the set of real numbers that has the largest MC value achievable by any set of real numbers for the spatial arrangement defined by the geographic connectivity matrix $\mathbf{C}$; the second eigenvector is the set of real numbers that has the largest achievable MC value by any set that is uncorrelated with $\mathbf{E}_1$; the third eigenvector is the third such set of values; and so on through $\mathbf{E}_n$, the set of real numbers that has the largest negative MC value achievable by any set that is uncorrelated with the preceding (n-1) eigenvectors. As such, these eigenvectors furnish distinct map pattern descriptions of latent spatial autocorrelation in georeferenced variables, because they are both orthogonal and uncorrelated. Their corresponding eigenvalues index the nature and degree of spatial autocorrelation portrayed by each eigenvector (Tiefelsdorf and Boots, 1995), which can be standardized by dividing by the largest MC value (MC_{max}, which often is greater than 1).

The spatial autocorrelation of interest here is contained in the network of flows that constitutes spatial interaction—the flows themselves are spatially dependent. These

origin/destination paired variables can be linked to the individual flows data dyads through Kronecker products, which results in an n^2-by-n^2 connectivity matrix for flows that is asymptotically correct: $\mathbf{C}_{n^2} = \mathbf{C}_n \otimes \mathbf{C}_n$, where $\mathbf{C}_{n^2}$ is the n^2-by-n^2 binary 0-1 connectivity matrix for the n^2 spatial interaction flows, and $\mathbf{C}_n$ is the n-by-n connectivity matrix for the tessellation of n origins/destinations. As with most geographic weights matrices, $\mathbf{C}_{n^2}$ is very sparse. It also is quite sizeable. For example, the 73-by-73 $\mathbf{C}_n$ matrix for Pennsylvania becomes a 5,329-by-5,329 $\mathbf{C}_{n^2}$ matrix. This Kronecker product specification yields the origin and destination variates $\mathbf{E}_i \otimes \mathbf{E}_j$ and $\mathbf{E}_j \otimes \mathbf{E}_i$ for constructing spatial filters. Because a minimum level of spatial autocorrelation should be represented by a given eigenvector, say 0.25 (i.e., roughly 5% of the variance of a given eigenvector's values is attributable to redundant information resulting from the presence of spatial autocorrelation), and the eigenvalues of $\mathbf{C}_n \otimes \mathbf{C}_n$ are the pairwise products of the individual eigenvalues of $\mathbf{C}_n$, this threshold value should be increased to 0.5 (i.e., $0.5^2 = 0.25$) for each of the eigenvectors in a Kronecker product. The net result of this specification is a dramatic reduction in computational intensity. For the Pennsylvania case study, it results in 11 candidate eigenvectors from matrix $\mathbf{C}_n$, and hence 121 candidate eigenvectors from matrix $\mathbf{C}_{n^2}$.

Because the eigenvectors used to construct spatial filters are Kronecker products of vectors with zero means (i.e., $\mathbf{1}^T\mathbf{E} = 0$), eigenvector medians can be used to visualize the spatial filters.

4. Random effects: a brief overview

Random effects models are increasing in popularity (see, for example, Demidenko, 2004), partially because they have become estimable. One common specification is for the intercept term to be cast as a random effects, resulting in it representing variability about the conventional single-value, constant mean. The role of a random effects in this context may be twofold: (1) supporting inferences beyond the specific fixed values of covariates employed in an analysis; and, (2) accounting for correlation in a non-random sample of data being analyzed. Random effects are used if the values of independent variables—which were not deliberately selected by an experimenter—are thought to be a small subset of all possible values to which inferences are to be made, to account for heterogeneity/overdispersion/inter-observation correlation, or to handle observations that are not obtained by simple random sampling but come from a cluster or multi-level sampling design (assumption: the latent correlation structure is exchangeable within a cluster). Random effects yield larger standard errors than are obtained with fixed effects. During estimation, a random effects variable is integrated out (with numerical methods) of its likelihood function.

A random effects term is factored from the residuals for a model, yielding the specification

$$\mathbf{Y} = f(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \boldsymbol{\varepsilon}), \tag{3}$$

where ξ is a random observation effect (differences among individual observational units), and ε is a residual error. The composite error term is the sum of the two. In order for this factoring to occur, data must contain repeated measures. The simplest geographic form specifies a linear model, and these repeated measures are costituted by the spatial autocorrelation latent in georeferenced data. For example, a geostatistical model captures the random effects term, separating it from

residuals. The variogram describes spatial structuring in these random effects, whereas the nugget describes any aspatial random effects. This same conceptualization applies to a Bayesian model specification, where a spatially structured random effects term almost always is specified with a conditional autoregressive (CAR) model, in both its proper and improper (ICAR) forms, for hierarchical modeling purposes; this is the version available in WinBUGS. A more complex form specifies a linear or non-linear model with multiple years of data (i.e., a space-time dataset), or multiple variables (i.e., a general linear model with Q response variables), or some other repeated measurement situation (e.g., for flows data, an origin with n destinations).

5. Estimation results

Numerical experiments were conducted to explore relationships between the balancing factors and random effects.

5.1. Specifications ignoring network autocorrelation

Estimating the balancing factors for singly constrained specifications accounts for spatial autocorrelation in origin and destination geographic distributions, but not network autocorrelation. Because only one set of indicator variables is involved, the intercept term can be added to each factor, forcing α to 0 in equations (1) and (2), and allowing a more meaningful significance test for each balancing factor (the sum, with 2 dfs, and a variance that is the sum of the individual squared standard errors). Estimating random effects also overlooks network autocorrelaiton, and treats the n destination flow recipients as repeated measures for each origin, or the n origin flow sources as repeated measures for each destination. All of these specifications posit a unique value for each origin/destination for the n^2 flows data.

Estimation summary results appear in Table 1. The generalized linear mixed model (GLMM) random effects estimates are nearly identical to their balancing factor fixed effects counterparts (Figure 3):

$$\hat{a}_i = -0.00211 + 1.00242\,\hat{\xi}_O + \hat{\varepsilon}_O,\ R^2 = 1.0000, \text{ and}$$
$$\hat{b}_i = -0.00204 + 1.00271\,\hat{\xi}_D + \hat{\varepsilon}_D,\ R^2 = 1.0000,$$

where $\hat{\xi}_j$ and $\hat{\varepsilon}_j$ (j = O, D; O denoting origin and D denoting destination) are the resepctive random effects and residual terms appearing in equation (3). Furthermore, the random effect terms essentially are identical:

$$\hat{\xi}_O = -0.00000 + 1.00311\,\hat{\xi}_D + \hat{\varepsilon},\ R^2 = 1.0000.$$

In other words, no consequential differences exist between the estimated fixed and random effects.

Spatial filter descriptions of these variates are nearly identical, too (Table 2), and comprise 11 of the 18 candidate eigenvectors. These spatially structured random effects account for roughly two-thirds of the total random effects. This spatial structuring represents moderate-to-strong positive spatial autocorrelation, and is the principal reason the individual terms deviate

from a normal frequency distribution. These linear spatial filters account for virtually all of the spatial autocorrelaiton latent in these variates. Because the fixed and random effects are identical, the spatial filter residuals constitute spatially unstructured random effects. Figure 4 portrays these two components; because the four geographic distributions essentially are identical, Figure 4 represents all four of them.

Table 1. Summary statistics for estimations involving equations (1)-(3) with $\rho_{ij} \equiv 0$

statistics	Model specification			
	Origin-constrained	GLMM with origin random effects	Destination-constrained	GLMM with destination random effects
Deviance	476.14	476.14	296.43	296.43
Distance decay (γ)	0.6951	0.6951	0.6967	0.6967
# significant coefficients:				
< 0.01	21	3	29	3
0.01-0.05	9	24	10	24
0.05-0.10	7	21	5	21
Mean	-0.00208	0.00002	-0.00202	0.00003
Variance	0.01831	0.01822	0.01821	0.01811
P(Shapiro-Wilk)	0.0002	0.0002	0.0002	< 0.0001
Pseudo-R^2	0.9565	0.9565	0.9805	0.9805

NOTE: Shapiro-Wilk statistic furnishes a diagnostic test for normality
NOTE: GLMM estimation utilized a Newton-Raphson optimization technique

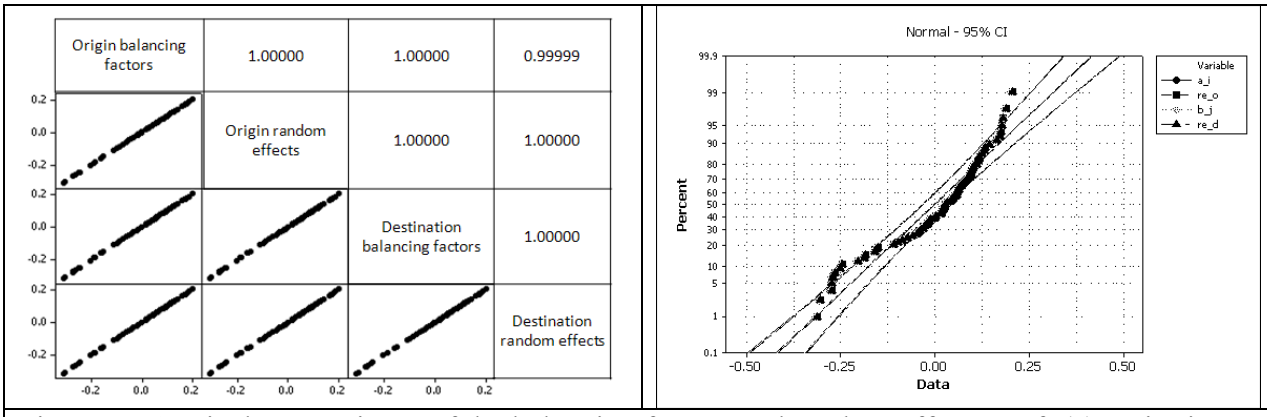

Figure 3. Grapical comparisons of the balancing factors and random effects. Left (a): pairwise scatterplots. Right (b): superimposed normal quantile plots.

Table 2. Summary results for normal probability model-based linear spatial filters

Estimated term	eigenvectors	R^2	P(Shapiro-Wilk)	SF_{MC}	Residual z_{MC}
a_i	E_1, E_3, E_4, E_6-E_{11}, E_{13}, E_{17}	0.7770	0.8346	0.74537	1.67
ξ_O		0.7767	0.8303	0.74556	1.68
b_j		0.7769	0.8360	0.74532	1.67
ξ_D		0.7767	0.8379	0.74557	1.68

NOTE: SF_{MC} denotes the Moran Coefficient for a given spatial filter

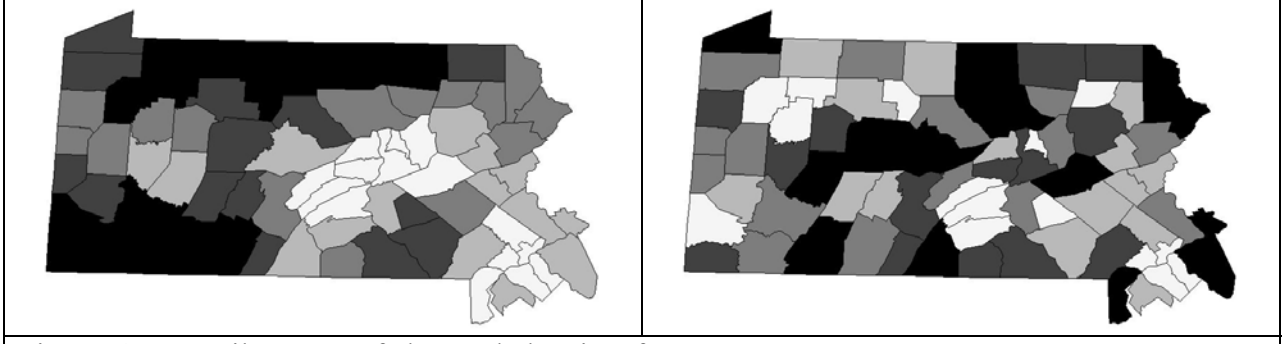

Figure 4. Quantile maps of the A_i balancing factors; values directly proportional to darkness of grayscale. Left (a): spatial filter. Right (b): residual.

5.2. Specifications accounting for network autocorrelation

Because the balancing factors are not independent of network autocorrelation components, in part because the two groups of variates share common eigenvectors, the balancing factors and spatial filters must be simultaneously rather than sequentially estimated. These flows-based spatial filters: represent moderate positive spatial autocorrelation, decrease overdispersion (i.e., extra-Poisson variation) by a third or more, only slightly increase pseudo-R^2 values, markedly decrease distance decay parameter value estimates (e.g., their confidence intervals do not overlap; Figure 5), and comprise about half of the 121 candidate eigenvectors (Table 3).

Table 3. Origin- and destination-constrained model specifications accounting for network spatial autocorrelation

statistic	Origin-constrained	Destination-constrained
Deviance	268.51	200.55
Distance decay (γ)	0.6343	0.6535
# of eigenvectors	73	64
Origin SF_{MC}	0.48991	0.58786
Destination SF_{MC}	0.53375	0.69940
Pseuco-R^2	0.9811	0.9897

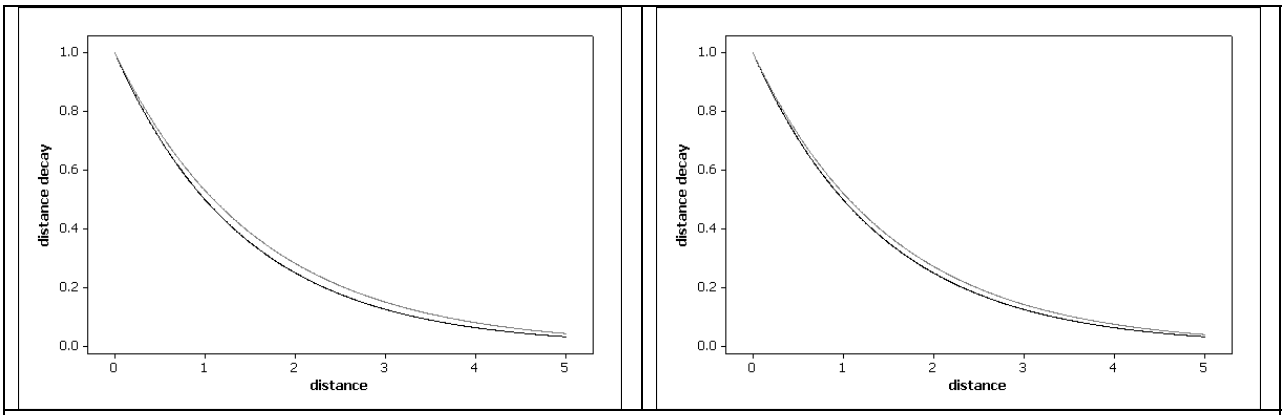

Figure 5. Distance decay effect; black line denotes unadjusted, and gray line denotes adjusted, for network spatial autocorrelaiton. Left (a): origin-constrained model. Right (b): destination constrained model.

The scatterplots of the observed versus predicted flows (Figure 6) display a close alignment of these sets of values. Imposing flow matrix row or column total constrains coupled with inclusion of the spatial autocorrelation components shrink especially the larger predicted flow values toward the perfect prediction line; the scatterplot also no longer displays a typical Poisson scatter of increasing variance with increasing amount of flow.

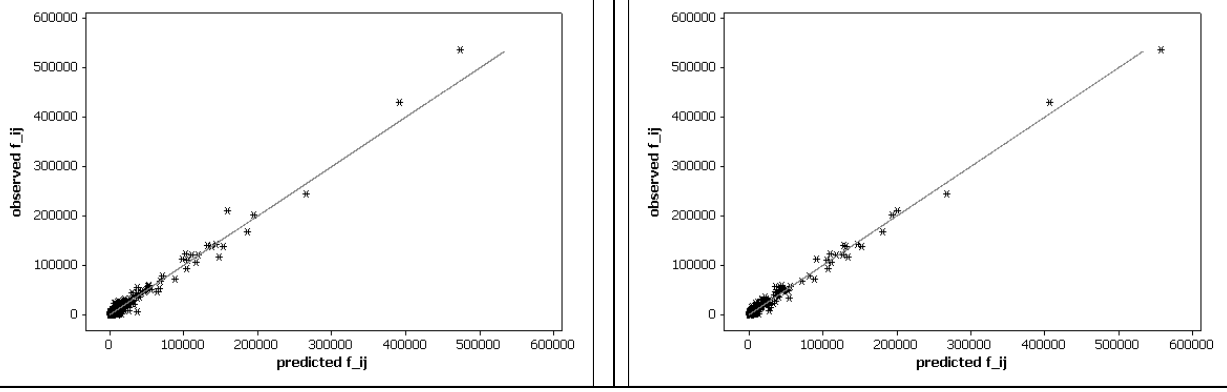

Figure 6. Scatterplots of observed versus predicted flows; gray line denotes the line of perfect prediction. Left (a): origin-constrained model. Right (b): destination-constrained model.

The flows-based spatial filters indicate mostly regional map patterns of dependency (Figure 7), which is sensible given that the flows are journeys to work. For the origin-constrained model specification, the origin spatial filter highlights the Philadelphia region as well as a northwest-central region, whereas the destination spatial filter highlights Philadelphia, too, and a north-central region. In contrast, for the destination-constrained model specification, both origin and destination spatial filters highlight a northwest region.

6. Conclusions and implications

Findings of this research suggest a number of intersting conclusions and implictions. Foremost, and quite counter-intuitive, for all practical purposes (i.e., except for slight rounding/algorithm-convergence error), fixed and random effects are identical. This finding is the outcome of an equivalency between assigning a single fixed effects indicator variable to each origin/destination, on the one hand, and estimating a single random effects value for an origin/destination while treating the corresponding n destinations/origins as repeated measures, on the other hand. This finding also indicates that the number of degrees of freedom associated with the random effects term in this context is n-1. Substituting the estimated random effects term into a model specification yields an overall quasi-likelihood standard error estimate of 0.0589 for it in the origin-constrained, and of 0.0482 for it in the destination-constrained, model specification. In contrast, the fixed effects specification furnishes such a standard error estimate individually for each value. Furthermore, latent spatial autocorrelation masks the bell-shaped curve characterizing the statistical distributions for these variates.

An expected finding is that the distance decay parameter estimates adjusted for network spatial autocorrelation are less than their unadjusted counterparts. These values differ substantially from their unadjusted counterparts; the pairs of values do not have overlapping

confidence intervals [origin-constrained: (0.6232, 0.6455) versus (0.6867, 0.7036); destination-constrained: (0.6440, 0.6629) versus (0.6900, 0.7034)]. Figure 5 reveals that the adjusted distance decay estimate is shallower than its unadjusted counterpart.

Figure 7. Quantile maps of the network autocorrelation spatial filters based upon medians; values directly proportional to darkness of grayscale. Top left (a): origin-constrained origin spatial filter. Top right (b): origin-constrained destination spatial filter.

An interesting finding is that both network and geographic distribution correlation components entail moderate and postive spatial autocorrelation. The map patterns are dominated by regional dependencies (Figures 4 and 7).

Finally, as with both the unconstrained and doubly-constrained spatial interaction models, adjusting for spatial autocorrelation in flows data improves model performance (e.g., higher R^2, lower deviance statistic). And, the cost in degrees of freedom (dfs) is minor (on average, 36-39 dfs are available for each parameter estimated here).

7. References

Chun, Y. (2007) *Behavioral specifications of network autocorrelation in migration modeling: an analysis of migration flows by spatial filtering*, unpublished doctoral dissertation, Department of Geography, The Ohio State University.

Chun, Y. (2008) Modeling network autocorrelation within migration flows by eigenvector spatial filtering, *J. of Geographical Systems*, 10(4): 317-344.

Curry L. (1972) Spatial analysis of gravity flows, *Regional Studies* 6, 131–147.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. NY: Wiley.

Fischer, M., and D. Griffith. (2008) Modeling Spatial Autocorrelation in Spatial Interaction Data: A Comparison of Spatial Econometric and Spatial Filtering Specifications, *Journal of Regional Science*, 48: 969-989.

Flowerdew R., and M. Aitkin. (1982) A method of fitting the gravity model based on the Poisson distribution, *Journal of Regional Science*, 22: 191-202.

Griffith, D. (2002) A spatial filtering specification for the auto-Poisson model, *Statistics & Probability Letters*, 58: 245-251.

Griffith, D. (2007) Spatial structure and spatial interaction: 25 years later, *The Review of Regional Studies*, 37, #1: 28-38.

Griffith, D. (2009a) Modeling spatial autocorrelation in spatial interaction data: empirical evidence from 2002 Germany journey-to-work flows, *J. of Geographical Systems*, 11: 117-140.

Griffith, D. (2009b) Complexity-to-simplicity features of spatial autocorrelation in spatial interaction: empirical descriptions of journey-to-work in a space-economy.

Klios, Inc. (2004) Economic analysis of interstate commute patterns in the greater Philadelphia region, www.klios.net/GPCC_JTW_Economic_Analysis.pdf (accessed on 1/27/2009).

LeSage, J., and R. Pace. (2008) Spatial econometric modelling of origin-destination flows, *Journal of Regional Science*, 48: 941-968.

Sen, A., and T. Smith. (1995) *Gravity Models of Spatial Interaction Behavior*. Berlin: Springer.

Tiefelsdorf M, Boots B (1995) The exact distribution of Moran's I, *Environment and Planning A*, 27: 985-999.

Wilson, A. (1967) A statistical theory of spatial distribution models, *Transportation Research*, 1: 253-269.