Modelling spatial flows with R

Thibault Laurent - Paula Margaretic - Christine Thomas-Agnan

July 9, 2019

What are spatial flows ?

Origin-destination (OD) flow data are data doubly indexed by two geographical locations. They represent movements of people, money (or other) between these two locations. Typical examples are

- home-to-work commuting data
- air-passenger flows between two airports
- quantity of money spent in a given store by a customer living in a given area (geomarketing)
- amount of trade between two countries
- number of migrants moving from one country to another

Models for spatial flows

In econometrics, people use **gravity models** for modelling spatial flows. They are linear regression models explaining the logarithm of the flow as a function of

- characteristics of origin
- characteristics of destination
- characteristics of the couple origin + destination

In spatial econometrics, to take into account possible dependence between "neighboring flows", one can adapt spatial autoregressive models to the case of flows: **spatial interaction models**.
In this project we concentrate on the **Spatial Durbin model for flows**.

Estimation methods and their implementation

For fitting the spatial Durbin model, we consider three estimation methods for the parameters

- Maximum likelihood (ML)
- A Bayesian approach
- A two-stage least squares (S2SLS) approach

Existing R-code

- ML: possible to use R-code for non-flow data (spdep package) but some preformatting required and restrictions
- Bayesian estimators: only Matlab code (James LeSage), not public, restricted to particular cases
- 2SLS: possible to use R-code for non-flow data (spdep) but some preformatting required

Contribution

We distinguish between:

- Symmetric case: List of origins $=$ list of destinations
- Asymmetric case: List of origins $\neq$ list of destinations

- We provide preformatting functions
- We extend existing implementations in three directions
  1. We allow for a different list of locations for origins and destinations
  2. We allow for different characteristics at origin and destination, even in the symmetric case
  3. We allow for multiple spatial weight matrices

Project Overview

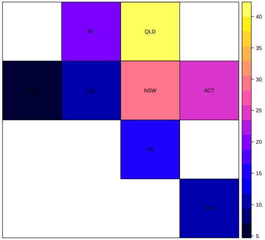
In black: existing, in red: our current contribution

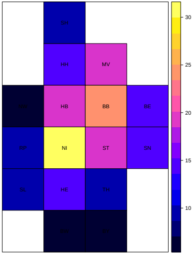
In green: forthcoming

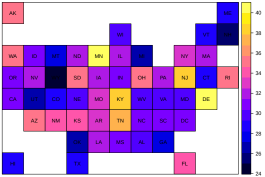
	Max Lik	Bayesian	2SLS
List orig. = List dest.	In vectorized format and with single W matrix possible to use non flow-specific code	no program freely available 1-we translate into R LeSage Matlab code for matrix format several W possible 2-we write a vectorized version	we construct a function specific to flows in vectorized format
List orig. $\neq$ List dest.	vectorized format works need write matrix implementation	vectorized code works several W possible need write matrix implementation	vectorized code works several W possible

Toy data for illustrations

Simplified maps of Australia, Germany and USA







from Many-to-Many Geographically-Embedded Flow Visualisation: An Evaluation (2016) IEEE transactions on visualization and computer graphics

Y.Yang, T.Dwyer, S. Goodwin and K. Marriott

Using Kronecker products

kronecker(A,B)

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} & a_{12} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \\ a_{21} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} & a_{22} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \end{pmatrix}$$

Matrix and vector formats

Flows F_{od}, with $o = 1, \quad n_o$ and $d = 1, \quad n_d$ (as well as explanatory variables which are origin-destination characteristics) can be presented in two different formats

- **matrix format**
- **vectorized format**

$$\mathbf{F} = \begin{pmatrix} o_1 \ ! \ d_1 & o_1 \ ! \ d_2 & \ldots & o_1 \ ! \ d_{n_d} \\ o_2 \ ! \ d_1 & o_2 \ ! \ d_2 & \ldots & \ldots \\ & & & o_{n_o \ 1} \ ! \ d_{n_d} \\ & & & o_{n_o} \ ! \ d_{n_d} \end{pmatrix}$$

Matrix and vector formats

The n_o n_d flow matrix $\mathbf{F}$ can be converted into a $n_o n_d$ 1 vector $\mathbf{F}$ in two different ways ($N = n_o n_d$)

- by stacking its rows (origin-centric ordering)
- by stacking its columns (destination-centric ordering)

With the **destination centric ordering**,

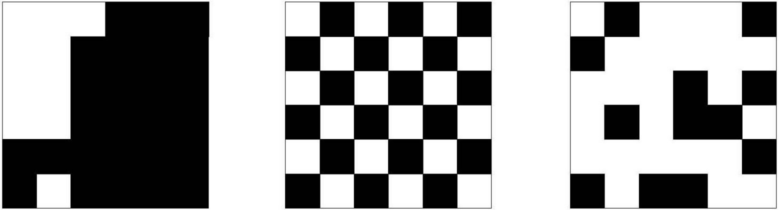
- an origin characteristic (vector OX of size n_o 1) will enter in the model as $\mathbf{X}_o = OX \bigotimes i_{n_d}$ (an N 1 vector)
- a destination characteristic (vector DX of size n_d 1) will enter in the model as $\mathbf{X_d} = i_{n_o} \bigotimes DX$ (an N 1 vector)

where i_n is a vector of ones of size n

Spatial Econometrics models

- Spatial data: indexed by a geographical location
- Spatial econometric data: the location is a zone
- Other approaches: continuous location (geostatistics) and random location (Spatial Point Process)

Objective of spatial econometrics models: take into account spatial heterogeneity and spatial autocorrelation



Spatial Weight matrices

The weight matrix is the spatial version of the lag operator in times series.

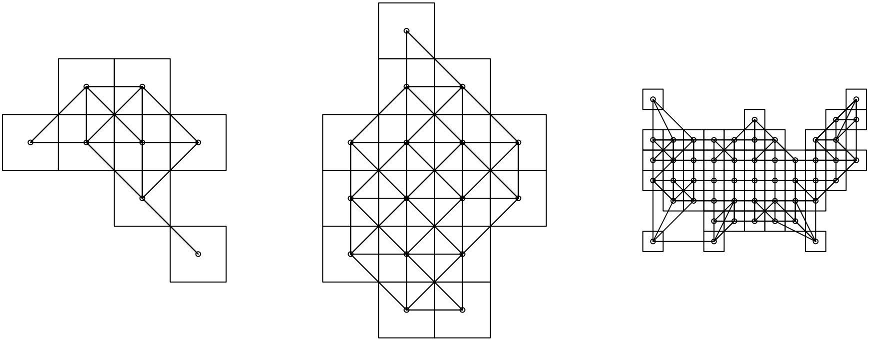
For n geographical sites, a weight matrix $\mathbf{W}$ is an $n \times n$ matrix (not necessarily symmetric)
its element w_{ij} is an indicator of the intensity of proximity between location i and location j (specifies the topology of the domain)

By convention $w_{ii} = 0$.

It is often row-normalized $\sum_{j=1}^{n} w_{ij} = 1$.

Lagged variable. if Z is a variable, WZ is the corresponding lagged version: if W is row-normalized, the term i of WX is the mean (weighted by proximity) of the values of X for neighbors of location i

Neighborhood structure for toy data



Neighborhood structure for flows

Given

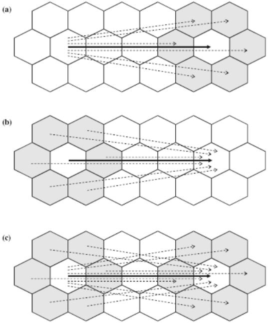
- OW of dimension $n_o \quad n_o$ for characterizing the proximity in the set of origins
- DW of dimension $n_d \quad n_d$ for characterizing the proximity in the set of destinations

we can then obtain the three types of neighborhood structures as follows

- origin based spatial neighborhood matrix: $\mathbf{W_o} = OW \otimes I_{n_d}$ two flows are neighbors if their origins are neighbors according to OW
- destination based spatial neighborhood matrix: $\mathbf{W_d} = I_{n_o} \otimes DW$ two flows are neighbors if their destinations are neighbors according to DW
- origin-to-destination based spatial neighborhood matrix: $\mathbf{W_w} = OW \otimes DW$ two flows are neighbors if their origins and their destinations are neighbors according respectively to OW and DW

Neighborhood structure for flows

Illustration from Chun (2008)



Gaussian log-linear specification of Durbin SIM model

- $XL_o = OLX \otimes i_{n_d}$, lagged characteristics of the spatial units acting as origins characteristics
- $XL_d = i_{n_o} \otimes DLX$, lagged characteristics of the spatial units acting as destinations characteristics.
- X_i intra-regional characteristics
- G matrix of variables characterizing both origin and destination

Model equation in vectorized form ($\mathbf{y} = \log(F)$)

$$A(W)\mathbf{y} = X_o b_o + X_d b_d + X_i b_i + XL_o d_o + XL_d d_d + Gg + e, \quad (1)$$

with the spatial filter matrix $A(W) = (I_N \;\; {}_N \quad r_o W_o \quad r_d W_d + r_w W_w)$

Some interesting submodels of the general gaussian log-linear spatial model

- **Specification 1**: Assumption $r_o = r_d = 0$ yields the gravity model with independent observations
- **Specification 2**: Assumption $r_d = 0$ yields a spatial dependence model using a single weight matrix $\mathbf{W_o}$ reflecting origin-based spatial dependence
- **Specification 3**: Assumption $r_o = 0$ yields a spatial dependence model using a single weight matrix $\mathbf{W}_d$ reflecting destination-based spatial dependence
- **Specification 4**: Assumption $r_o = r_d$ yields a spatial dependence model using a single weight matrix $\mathbf{W}_g \quad \frac{1}{2}(\mathbf{W_o} + \mathbf{W}_d)$ reflecting a cumulative, non separable origin and destination spatial dependence effect

MLE in ordinary spatial Durbin model

Why MLE? Least squares is biased in Durbin model. The computation of the MLE in the Durbin model proceeds in two steps. Stack X and WZ in a variable X_1 and stack b and d in a parameter g

1. Optimization wrt b for fixed r is in closed form

$$\hat{s}^2(r) = \frac{1}{n}(y \quad A(r) \ ^{1}(X_1\hat{g}(r))^{0}A(r)^{0}A(r)(y \quad A(r) \ ^{1}X_1\hat{g}(r))$$

   and

$$\hat{g}(r) = (X_1^{0}X_1) \ ^{1}X_1^{0}A(r)Y.$$

   with $A(r) = (I \quad rW)$

2. Plug in values from step 1 in the Log-Lik to obtain the so-called concentrated log-lik and optimize it numerically.

The concentrated LL contains a log(det) term, which is demanding, needs to be approximated for large data.

Specific to flow data: if several weight matrices, step 2 is more difficult.

Bayesian implementation

As in LeSage (2009)

- parameters associated to covariates are assigned uninformative priors
- s^2 is assigned an inverse gamma prior
- variance scalar parameters are assigned a c^2 prior
- r parameters are assigned uniform priors on $[\quad 1, 1]$ (plus stability restrictions)

About LeSage implementations of MLE and Bayes

For Bayesian and MLE, LeSage use computational tricks based on properties of Kronecker products **in the symmetric case** and in matrix format) number operations for recomputing concentrated Log Lik independent from number of sites and number of explanatory variables.

Possible extension to symmetric case: the tricks go through under the restriction that all origins have the same list of destinations (cartesian product). Not implemented yet.

	Bayesian	Log-Likelihood	S2SLS
Mean	77.46	0.3112	0.00592
Std. Dev.	0.947	0.00512	0.000787

Table: Comparison execution time of 3 methods (vectorized format) in seconds for Germany

Matrix versus vector format

After vectorization, any code for non-flow data can be used, however we run into a big data problem ... for example for Bayesian method in the symmetric case

	Matrix	Vector
Mean	6.714	77.46
Std. Dev.	0.201	0.947

Table: Mean execution time in seconds for Germany

Why spatial two stage least squares is appealing ?

S2SLS: Kelejian and Prucha (1998)
Based on two stage LS hence computationally simple

- regression of the lagged endogenous variable on H consisting in a selection of independent among the explanatory variables and their lagged versions with W and W^2.
- regression of the endogenous variable on the explanatory variables and the fitted value of the lagged endogenous variable obtained at step 1.

Flow-specific difficulty: products W_d times X_o is exactly equal to X_o. Hence products such as $W_d^s X_o$ and $W_o^s X_d$ should be removed from the list of variables in H.

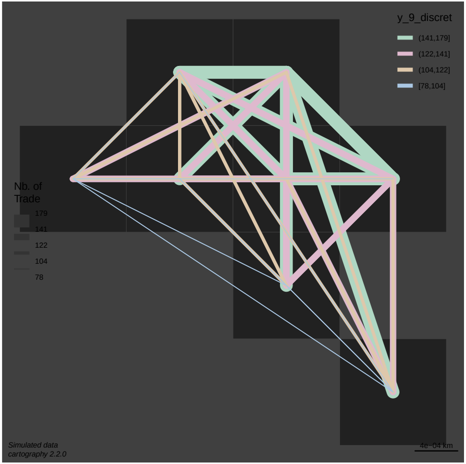
Australia toy data

We use the Australian simulated data

	origin	dest	x_o	x_d	g	W_dx_d	W_ox_o	y
1	NT	NT	20	20	0.0000000	21.75000	21.75	76.41378
2	NT	QLD	20	40	0.6931472	21.25000	21.75	96.04838
3	NT	WA	20	7	0.8813736	15.00000	21.75	58.23100
4	NT	SA	20	10	0.6931472	22.40000	21.75	66.07393
5	NT	NSW	20	30	0.8813736	22.00000	21.75	86.26428
6	NT	ACT	20	25	1.1743590	28.33333	21.75	87.27157

and model specification 3

Australia toy data



Comparison of the 3 methods on a single replication -Australia toy data

	Bayes	ML	S2SLS	True
intercept	2.65	3.66	7.81	0
x_d	0.97	0.97	0.95	1
$W_d x_d$	0.48	0.45	0.31	0.5
$W_o x_o$	0.22	0.2	0.12	0.25
G	1.83	1.78	1.6	2
r_d	0.46	0.5	0.65	0.4

Another comparison between Bayesian, ML and 2SLS

Taken from Thomas-Agnan and LeSage (2014); $n_o = n_d = n = 8$

Variables	Bayes	ML	S2SLS	True
r_d	0.399***	0.409***	0.419***	0.4
Intercept	0.44	0.352	0.278	0
$X1_d$	0.48***	0.477***	0.473***	0.5
$X2_d$	0.676**	0.685**	0.686***	1
$X1_o$	1.502***	1.478***	1.454***	1.5
$X2_o$	2.166***	2.134***	2.100***	2
G	-0.48***	-0.474***	-0.467***	-0.5

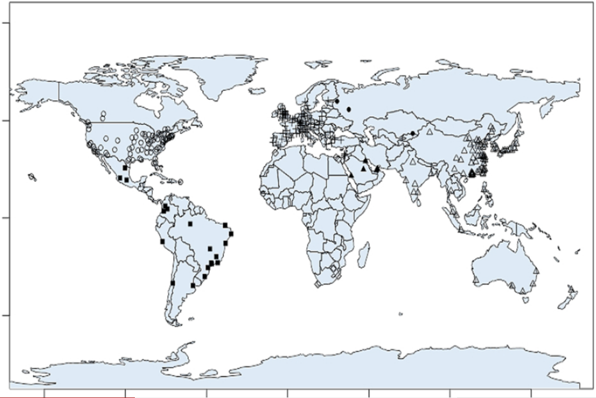
Comparison between Bayesian, ML and 2SLS in the asymmetric case

We use two grids with 30 origins and 12 destinations.

	Variables	Bayes	ML	S2SLS	True
1	rho_d	0.309	0.317	0.342	0.400
2	(intercept)	2.06	2.002	1.566	0.000
3	z_d	1.089	1.081	1.056	1.000
4	W_dz_d	0.578	0.568	0.542	0.500
5	x_o	0.468	0.462	0.449	0.500
6	W_ox_o	0.432	0.427	0.402	0.250
7	g	-2.26	-2.235	-2.161	-2.000

The air passenger data

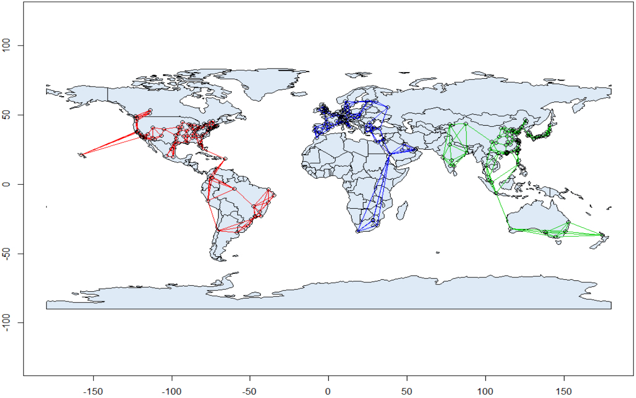
- OD, city to city, air passenger flows between 279 cities in 2012
- $n_o = n_d = n = 279$
- Covariates: GDP per capita, per city; distance ($\mathbf{g}$), air fares ($\mathbf{f}$), and two dummy variables for short and long haul



Weight matrix for air passenger flows

The n n weight matrix $\mathbf{W}$ is such that,

- $w_{ij} > 0$ if city i is one of the $k = 4$ nearest neighbours of j
- $\sum_j w_{ij} = 1$. By convention, $w_{ii} = 0$



The spatial auto-regressive model specification

We consider a specification including two weight matrices $\mathbf{W_o}$ and $\mathbf{W}_d$

$$\log(\mathbf{y}) = r_o\mathbf{W_o}\log(\mathbf{y}) + r_d\mathbf{W}_d\log(\mathbf{y}) + a i_N + \mathbf{X}_o b_o + \mathbf{X}_d b_d + \\ \mathbf{W_o X_o} d_o + \mathbf{W_d X_d} d_d + g\mathbf{g} + f\mathbf{f} + q_1\mathbf{d_1} + q_2\mathbf{d_1} + u$$

- $\mathbf{X_o}$ and $\mathbf{X_d}$ are GDP per capita of the cities acting as origins and destinations, respectively
- $\mathbf{g}$ and $\mathbf{f}$ denote distance and air fares respectively
- $\mathbf{d_1}$ and $\mathbf{d_2}$ are two dummy variables for short and long haul
- a, g, f, b_o, b_d, d_o, d_d, q_1 and q_2 are scalar parameters and $u \quad N(0, s^2 I_N)$

Bayesian model estimates with multiple neighborhood matrices

Table: SDM estimates with weight matrices W_o and W_d, $k = 4$ nearest neighbors

	Mean	Lower_{05}	Upper_{95}	T_{stat}
r_d	0.453	0.437	0.470	42.696
r_o	0.450	0.433	0.467	43.640
Intercept	0.988	0.826	1.159	9.275
W_o GDP capita_o	-0.381	-0.461	-0.302	-7.861
W_d GDP capita_d	-0.387	-0.468	-0.305	-7.826
Fares	-0.659	-0.711	-0.607	-20.899
GDP capita_o	0.448	0.395	0.501	13.924
GDP capita_d	0.459	0.405	0.513	14.046
Short Haul	0.231	0.068	0.397	2.282
Long Haul	0.643	0.067	1.222	1.849
Distance	0.175	0.108	0.239	4.360

Conclusions and Future Work

- We examine the problem of modelling OD flow data, using spatial autoregressive interaction models to account for spatial dependence
- Our contribution:
  1. We provide an R implementation of the ML, Bayesian and S2SLS methods for the spatial Durbin model
  2. We extend the implementations by allowing for possibly different origin and destination characteristics and for a possibly different list of locations for origins and destinations
- Forthcoming: code optimization, impacts computation, decomposition of total impacts in the asymmetric case, including more models (Spatial error model, Spatial Filtering), including prediction functions, etc.

Some references

- LeSage, J. P., and Pace, R. K. (2008). Spatial Econometric Modeling of Origin-Destination Flows. Journal of Regional Science, 48(5), 941-967.
- Fischer M, LeSage J (2010) Spatial econometric methods for modeling origin-destination flows. In: Fischer M, Getis A (eds) *Handbook of applied spatial analysis: Software tools, methods and applications*. Springer-Verlag Berlin Heidelberg.
- Modeling network autocorrelation within migration flows by eigenvector spatial filtering, Y. Chun (2008), J. Geogr. Syst. 10: 317-344.
- An open access modeled passenger flow matrix for the global air network in 2010, Z. Huang et al., PLOS ONE (2013) 8.
- LeSage, J. P., & Satici, E. (2016). A Bayesian Spatial Interaction Model Variant of the Poisson Pseudo-Maximum Likelihood Estimator. In Spatial Econometric Interaction Modelling (pp. 121-143). Springer, Cham.
- LeSage, J. P., Fischer, M. M., & Scherngell, T. (2007). Knowledge spillovers across Europe: Evidence from a Poisson spatial interaction model with spatial effects. Papers in Regional Science, 86(3), 393-421.
- Chun Y, Griffith DA (2011) Modeling network autocorrelation in space-time migration flow data: an eigenvector spatial filtering approach. *Annals of the Association of American Geographers* 101 (3): 523-536.
- LeSage, J. P., and Fischer, M. M. (2016) Spatial regression-based model specifications for exogenous and endogenous spatial interaction. In Spatial econometric interaction modelling (pp. 15-36), Patuelli R, Arbia G (eds), Springer: 15:36, Cham.

Personal contributions

- Interpreting Spatial Econometrics Origin-Destination Flow Models with J. LeSage (in Journal of Regional Science, 2014).
- Spatial econometric OD-Flow models, in : Handbook of Regional Science, Fischer M.M. and Nijkamp P (eds), Springer, 2014, 1653-1673.
- Spatial dependence in (origin-destination) air passenger flows, with Paula Margaretic and Romain Doucet (in Papers in Regional Science, 2015)
- with A. Ruiz-Gazen, T. Laurent and J. LeSage, unpublished manuscript, 20.
- Work in progress (with T. Laurent and P. Margaretic): asymmetric case - alternative estimation methods - impacts decomposition - R implementation