Big Data: Big Models for Urban Planning: GWR and Urban Analytics

A. Stewart Fotheringham

Professor of Computational Spatial Science
School of Geographical Sciences and Urban Planning
Arizona State University
stewart.fotheringham@asu.edu
Setting the Scene I

• It is important to differentiate spatial data from spatial processes

• **Spatial data** are the observations we make about our environment. Spatial data often exhibit **spatial dependency**

• Spatial regression models are designed to account for this affect
How Britain voted in the referendum

Northern Ireland shares a completely porous border with Ireland, which is in the European Union. Trade issues could arise between the two.

The majority of Wales voted strongly to leave, except for the largest city Cardiff, which voted to remain by 60 per cent.

The Scottish first minister has said that a leave vote could trigger a referendum vote in Scotland to leave Britain. Scots rejected independence in a referendum in September 2014 by 55 per cent to 45 per cent.

London, along with Scotland, led the vote to remain in the European Union, though the east side of the city voted to leave.

Sources: Preliminary results data from the BBC/ NYT
• **Spatial processes** are the unobserved processes that generate the spatial data we observe. Spatial processes may exhibit **spatial heterogeneity**

• Geographically Weighted Regression (GWR) is designed to capture this affect
To clarify…

Spatial Data

Spatial Processes

Spatial Associations

Measure

Infer
Setting the Scene II

• Deeper understanding comes from asking the ‘why’ of spatial processes than the ‘what’ of spatial data.

• One common and useful way of uncovering information on spatial processes is to formulate and calibrate models, such as regression models, and to obtain estimates of the model parameters.

• The more parameters we can reliably estimate, the more information on urban processes we can generate.
<table>
<thead>
<tr>
<th>Urban Issue</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crime rates</td>
<td>house prices; unemployment rates</td>
</tr>
<tr>
<td>House prices</td>
<td>property variables; neighbourhood variables</td>
</tr>
<tr>
<td>Electricity usage</td>
<td>house size; number of people per dwelling; income</td>
</tr>
<tr>
<td>Travel demand</td>
<td>income; car ownership rates; age</td>
</tr>
<tr>
<td>Pollution levels</td>
<td>car ownership; elevation; road density</td>
</tr>
<tr>
<td>Shopping trips</td>
<td>store variables; car ownership; income</td>
</tr>
</tbody>
</table>
Regression

In a typical linear regression model applied to spatial data we assume a stationary process:

- the same stimulus provokes the same response in all parts of the study region process:

\[ y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \varepsilon_i \]
The assumption of stationarity in regression

\[ y_i = \alpha + \beta x_i \]

Assumption is that the values of \( \beta \) are the same everywhere.
What if relationships are not the same everywhere? For example…

Suppose we regressed the price of a house on a set of house attributes, including presence/absence of a garage.

Would we really expect this relationship to be the same for places in the City center and the Suburbs?
To overcome this problem, we can apply **GEOGRAPHICALLY WEIGHTED REGRESSION (GWR)**

GWR is a local statistical technique to analyse spatial variations in *relationships*

We are not satisfied with global averages of spatial data e.g. climate data / health data

Why then should we be satisfied with global averages of statistics that measure relationships within spatial data e.g. regression?
To allow the relationships we are measuring to vary over space is quite simple and *is the essence of GWR*

\[ y_i = \beta_0(i) + \beta_1(i)x_{1i} + \beta_2(i)x_{2i} + \ldots + \beta_n(i)x_{ni} + \varepsilon_i \]
Geographically Weighted Regression
Geographically Weighted Regression
Main output from GWR is a set of location-specific parameter estimates which can be mapped and analysed to provide information on spatial non-stationarity in relationships plus an optimal bandwidth which is a measure of the spatial scale at which the processes operate
An Example using Educational Attainment Data in Georgia
In GWR, we can also ...

- estimate local standard errors
- derive local t statistics
- calculate local goodness-of-fit measures
- perform tests to assess the significance of the spatial variation in the local parameter estimates

GWR is now widely used in a great variety of application areas. Implemented within ArcGIS, R, Stata and several stand-alone packages available including GWR 4
Customized GWR Software (GWR4)

- GUI
- Control file
- Summary file
- Local estimates file
The concept of spatial weighting can be applied to any type of model – does not have to be linear regression

- GW PCA
- GW Discriminant Analysis
- GW Spatial Interaction Models
- etc
Example: Spatially Weighted Interaction Models (SWIM)

Spatial Interaction is movement or transmission over space resulting from a decision-making process.

Movement involves physical transfer (shopping trips; commuting; migration etc).

Transmission involves non-physical transfer (phone calls; social network messages; information etc).
A Typical Spatial Interaction Matrix

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>M_{11}</td>
<td>M_{12}</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>M_{1n}</td>
</tr>
<tr>
<td>2</td>
<td>M_{21}</td>
<td>M_{22}</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>M_{2n}</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>m</td>
<td>M_{m1}</td>
<td>M_{m2}</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>M_{mn}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Inflows</th>
<th>D1</th>
<th>D2</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>Dn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Outflows</td>
<td>O1</td>
<td>O2</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>Om</td>
</tr>
</tbody>
</table>

Overall Total M
Spatial Interaction Models

\[ M_{ij} = k \, v_i^\alpha \, w_j^\theta \, d_{ij}^\beta \]

Objective: Obtain accurate parameter estimates for each relationship in the model to inform on the processes affecting flows.

Why?: We then have a better understanding about what causes different volumes of interaction and can modify the patterns of flows (e.g. Migration) and/or predict flows more accurately under changing scenarios (e.g. Retailing).
Typically a matrix of flows is used to calibrate a SIM which generates one estimate of each parameter in the model.
## Global Poisson Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>origin active pop</td>
<td>0.85</td>
<td>316</td>
</tr>
<tr>
<td>dest employment</td>
<td>1.01</td>
<td>450</td>
</tr>
<tr>
<td>road distance</td>
<td>-0.67</td>
<td>-131</td>
</tr>
</tbody>
</table>

Pseudo $R^2$ 0.945

This doesn’t give us any information on the spatial variation of processes operating within this agglomeration
Are we able to disaggregate this system in some way so that we can estimate more parameters?

Yes, through using GW concepts to create Spatially Weighted Interaction Models
Spatially Weighted Individual Flow Models

*The Ultimate in SI Modelling?*

**Question:** What’s happening around the flow $1 \rightarrow 1$

**Solution:** Weight the flows $1 \rightarrow 2$, $1 \rightarrow 3$, $2 \rightarrow 1$, $2 \rightarrow 2$, $2 \rightarrow 3$, $3 \rightarrow 1$, $3 \rightarrow 2$, $3 \rightarrow 3$ according to their proximity to $1 \rightarrow 1$.

This generates a **Fully localised SIM**
$$\text{Dis} \ (O'D',OD) = \left[ (x'_1-x_1)^2 + (x'_2-x_2)^2 + (x'_3-x_3)^2 + (x'_4-x_4)^2 \right]^{\frac{1}{2}}$$
Spatially Weighted Individual Flow models

\[ M_{ij} = k_{{\{i\},\{j\}}} v_{i}^{\alpha_{{\{i\},\{j\}}}} w_{j}^{\theta_{{\{i\},\{j\}}}} d_{ij}^{\beta_{{\{i\},\{j\}}}} \]

where \{\} denotes a parameter estimate relating to a region around i or j

We can now calibrate a SIM for individual flows…
An Example of Commuting Flows in Lausanne

70 communes of the agglomeration of Lausanne

Lake Geneva

Source: Swisstopo, VECTOR200, 2013
Global Poisson Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>origin active pop</td>
<td>0.85</td>
<td>316</td>
</tr>
<tr>
<td>dest emp. opps</td>
<td>1.01</td>
<td>450</td>
</tr>
<tr>
<td>road distance</td>
<td>-0.67</td>
<td>-131</td>
</tr>
</tbody>
</table>

Pseudo $R^2$ 0.945

This doesn’t give us much information on the spatial choice processes operating within this agglomeration.
Spatially Weighted Individual Flow Model

\[ M_{ij} = k_{\{i, j\}} v_i^{\alpha_{\{i, j\}}} w_j^{\theta_{\{i, j\}}} d_{ij}^{\beta_{\{i, j\}}} \]

where \( \{ \} \) denotes a parameter estimate relating to a region around i or j
Summary

• The availability of big data sets relating to urban areas is increasingly rapidly e.g. using GPS tracking for movement; the use of transportation cards (Oyster; octopus etc)

• There is an increasing demand to employ these large data sets to better understand urban processes
The traditional BIG data approach

Big, highly multivariate, lots of confounding effects

Want this to be reliable, robust, and useful to decision-makers
The BIG models approach is the opposite of the BIG data approach.
Combining BIG Models with BIG data

DATA

INFO
For example, in case of commuting flows in Lausanne, Switzerland, we have moved from a model with only 3 behavioral parameters (Global) to one with 14,700 parameters (Individual SWIM)!
Thank you!