Regional analysis: The why and how of spatial econometrics

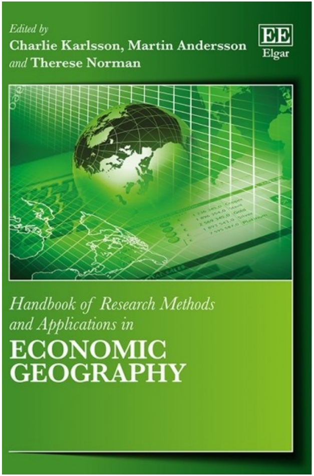
ERSA - OECD Winter School 2019

Diego Giuliani - diego.giuliani@unitn.it

21 January 2019 - Trento

Spatial Econometrics and Regional Science (1)

A very nice reference for an introduction to spatial econometrics for regional sciences is LeSage (2014a),



Edited by
Charlie Karlsson, Martin Andersson
and Therese Norman

EE
Elgar

Handbook of Research Methods and Applications in
ECONOMIC GEOGRAPHY

Spatial Econometrics and Regional Science (2)

Broadly speaking, spatial econometrics can be viewed as a sub-branch of econometrics that is specifically concerned with the empirical contexts where the sample data are characterized by some form of spatial dependence.

Applied regional science has to often deal with data-sets which statistical units are geo-referenced, at some level of spatial resolution (points, polygons, lines).

Geo-referenced data are subjected to Tobler's first law of geography: "*everything is related to everything else, but near things are more related than distant things*" (Tobler 1970).

Spatial Econometrics and Regional Science (3)

The first law of geography implies that geo-referenced observations are not likely to be independent and hence the fundamental assumption of independence for the classical linear regression model is likely to be violated. For example, the economic activity occurring in a region is likely to affect, in some way, that occurring in the neighbouring regions.

Spatial econometrics approaches allow to relax the independence assumption by incorporating a form of spatial dependence among statistical units within a regression modelling framework.

Moving from the assumption of independence to that of spatial dependence is of particular interest to regional scientists because it allows to assess **spatial spillovers**.

Spatial Econometrics and Regional Science (4)

According to LeSage and Pace (2009), in the context of regional data, a *spatial spillover* occurs when a change in the characteristics (or action) x of region i exerts a relevant influence on the outcome (or action) y of other regions j.

Therefore, under a regression modelling framework, a spatial spillover is detected if the cross-partial derivative of y_j with respect to x_i is non-zero, that is when $\frac{\partial y_j}{\partial x_i} \neq 0$.

One important reason why spatial econometric models are very useful for applied regional scientists is that they allow to quantify and test spatial spillovers (LeSage 2014a).

A stylized motivating example (LeSage 2014a) (1)

Suppose we are interested in assessing empirically the relationship between exam scores and time spent studying using a sample of six students that took the same exam,

```
  student   y  x
1    John  70  0
2   Devon  85  5
3   Steve  70 15
4  Denise  80 30
5   Billy  90 60
6    Mary 100 90
```

where y is the variable representing the exam score and x indicates the study time (in minutes).

A stylized motivating example (LeSage 2014a) (2)

The relationship between x and y can be properly studied by estimating a linear regression model,

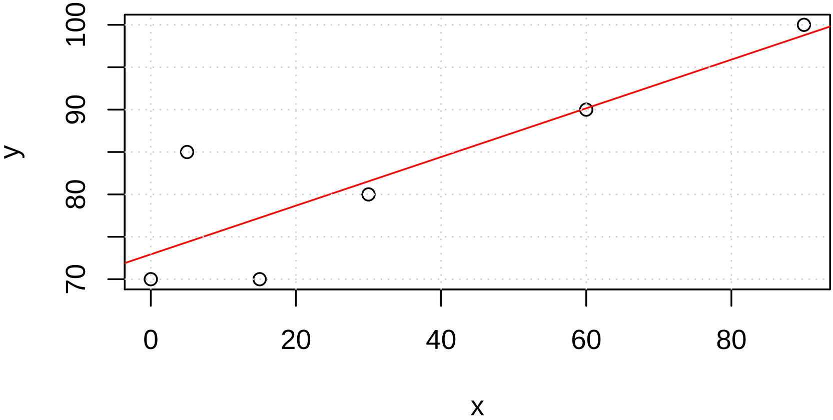
$$y_i = \alpha + \beta x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \dots, n$$

Using the OLS estimator we get the following results

$$\hat{\alpha} = 72.93, \qquad \hat{\beta} = 0.29,$$

thus we estimate that, on average, one more minute spent studying leads to additional 0.29 points on the exam score. Or, alternatively, 20 more minutes spent studying lead to a 5.8 points greater exam score.

A stylized motivating example (LeSage 2014a) (3)



A stylized motivating example (LeSage 2014a) (4)

The estimated regression parameters can be useful for

- studying the underlying phenomenon of interest, such as how study affects the outcome of an exam scores
- testing the validity of theoretical propositions
- perhaps use the model to predict the variable y, such as predicting the score of a student outside the sample who studied for 45 minutes.

The independence assumption of the classical linear regression analysis implies, in this case, that we consider that the study time of a single student affects only her/his exam score and not that of the other students.

A stylized motivating example (LeSage 2014a) (5)

This assumption, however, cannot be considered valid if we take into account the spatial characteristics of sample data, that is if we consider where the students were seated during the exam. Suppose that the seats occupied by the students followed a single row of seats.

Seats occupied	John	Steve	Mary	Devon	Billy	Denise
Study time	0	15	90	5	60	30
Exam score	70	70	100	85	90	80

If we examine the data in combination with the spatial location of the sample units, we can clearly see a spatial pattern: students seated closer to students with a high exam score tend to have a higher score. This kind of pattern is quite typical in regional data.

A stylized motivating example (LeSage 2014a) (6)

Spatial econometric models accounts for spatial dependence by specifying a **spatial weight matrix**, W, which is a $n \times n$ matrix containing information on the spatial connectivity between all pairs of n sample units.

The i, j^{th} element of W, denoted w_{ij}, specifies the level of spatial closeness between the i^{th} and j^{th} sample units, where $w_{ij} = 0$ for $i = j$.

Typically, W is also row-standardized so that each row's elements sum to one. This kind of normalization is useful for computational and statistical reasons.

A stylized motivating example (LeSage 2014a) (7)

A proper W matrix for the sample of students can be

$$W = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

which states that John and Denise have one neighbour while all the other students have two neghbours. The powers of W identify the the higher order neighbours. For example, W^2 identifies the neighbours to the neighbours, that is the second-order neighbours.

A stylized motivating example (LeSage 2014a) (8)

The spatial weight matrix can be used to obtain the so-called **spatial lag** of the dependent variable, Wy, that is

$$Wy = \begin{pmatrix} y_2 \\ 0.5y_1 + 0.5y_3 \\ 0.5y_2 + 0.5y_4 \\ 0.5y_3 + 0.5y_5 \\ 0.5y_4 + 0.5y_6 \\ y_5 \end{pmatrix},$$

which is a $n \times 1$ vector providing, for each sample unit, the average value of y of its neighbours. For example, for John, whose only neighbour is Steve, the spatial lag is given by Steve's exam score; for Mary, whose neighbours are Steve and Devon, the spatial lag is given by the average exam scores of Steve and Devon.

A stylized motivating example (LeSage 2014a) (9)

If we use Wy as a further independent variable in the linear model for the exam scores, we obtain the common **spatial autoregressive model (SAR)**,

$$y = \rho W y + X\beta + \varepsilon, \qquad \varepsilon \sim N(0, I\sigma^2)$$

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon$$

which specifies that the exam scores of students can depend on the scores of neighboring students (those seated to the left and right). The nature and strength of this dependence is assessed by the spatial autocorrelation parameter $\rho < |1|$.

If ρ is significantly different from zero, then the spatial location of students is relevant for the exam scores because of spatial dependence.

A stylized motivating example (LeSage 2014a) (10)

According to the SAR model, spatial dependence is such that a change in study time of one student can affect the exam scores of neighbouring students, as well as neighbours to those neighbouring students, and so on.

In other words, because of copying behaviours of students, the effects of study-time on exam score can spillover from students to neigbouring students and then to neighbours to neigbouring students and so on.

These global spillovers can be quantified by the partial cross-derivatives of y with respect to X (LeSage and Pace 2009).

A stylized motivating example (LeSage 2014a) (11)

It can be easily shown that, for a any regressor x,

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & & \\ \vdots & \vdots & \ddots & \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} = (I - \rho W)^{-1}\beta,$$

which is a $n \times n$ matrix providing how a change in the study time of each student affects both her/his own exam score (the main diagonal) and also the other students' exam score (the off-diagonal elements).

A stylized motivating example (LeSage 2014a) (12)

(LeSage and Pace 2009) proposed to summarize the information contained in the $\partial y/\partial x$ matrix by computing the

- *Direct impact*, as the average of the diagonal elements
- *Indirect impact*, as the average of either the row sums or the column sums of the non-diagonal elements
- *Total impact*, as the sum of the direct and indirect impacts.

The indirect impact represents a proper measure of (global) spatial spillover. Indeed, it expresses, on average, how a change in x reverberates on y all over the sample.

A stylized motivating example (LeSage 2014a) (13)

In point of fact, the indirect impact cumulates spillovers falling on immediately neighbouring units, neighbours to these units, neighbors to the neighbors of these units, and so on. This becomes clear if we look at the infinite series decomposition of $(I - \rho W)^{-1}$,

$$(I - \rho W)^{-1} = \sum_{q=0}^{\infty} \rho^q W^q = I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \dots$$

Since $\rho < |1|$, the effect of spillovers falling on first-order neighbours would be greater than those falling on second-order neighbours which, in turn, would be greater than those falling on third-order neighbours, and so on.

A stylized motivating example (LeSage 2014a) (14)

In general, for the phenomena of interest for regional science, it is reasonable to assume that spillovers are stronger between neighbouring regions than those located far apart, so this aspect of spatial decay of the SAR model is appropriate.

With respect to the students example, this could mean that the students seated most closely would be able to copy answers most accurately, whereas copying by the next nearest student may be less accurate, diminishing the influence of increased study time by a single student on exam scores of more distant students.

Useful Econometric Models for Regional Science (1)

In order to choose the proper econometric model specification, the analyst should understand whether the phenomenon under study is likely to generate *local* or *global* spatial spillovers.

We have a **local spatial spillover** when $\partial y_j / \partial x_i$ implies an impact on neighbouring units that do not generate endogenous interaction and feedback effects (LeSage 2014b).

We have a **global spatial spillover** when $\partial y_j / \partial x_i$ implies an impact on neighbouring units plus neighbours to the neighbouring units, neighbours to the neighbours, and so on, thus generating endogenous interaction and feedback effects (LeSage 2014b).

Useful Econometric Models for Regional Science (2)

Theoretical and substantial considerations should help in concluding if a *global* or *local* spillover specification is the proper one.

Examples of local spillover modeling situations could be: region border crossing by medical patients, cross-border shopping by consumers to avoid higher local taxes and crossing school district boundaries by households.

Examples of global spillover modeling situations could be: changes in levels of public assistance or local taxation in a region may lead to a reaction by neighboring regions, resources shared by more regions such as a river or a highway.

Global spillover specifications (1)

The most important global spatial spillover specification is the **spatial Durbin model** (SDM),

$$y = \rho W y + X\beta + WX\theta + \varepsilon, \qquad \varepsilon \sim N(0, I\sigma^2)$$

Similarly to the case of the SAR model, direct and indirect impacts for the SDM for the k^{th} independent variable in the X matrix can be computed from the matrix of partial derivative of y with respect to X'_k,

$$\partial y/\partial X'_k = (I - \rho W)^{-1}(I\beta^k + W\theta^k).$$

As with SAR, direct impact can be obtained as the average of the diagonal elements of $\partial y/\partial X'_k$, while indirect impact as the average of sums of the non-diagonal elements (LeSage and Pace 2009).

Global spillover specifications (2)

If $\theta = 0$, the SDM reduces to the SAR model. An important advantage of the SDM over the SAR model lies in the fact that the former does not pose restrictions on the partial derivatives.

Elhorst (2010) shown that in the SAR model the ratio between the indirect and direct impacts is the same for every independent variable, and that its magnitude depends only on ρ and W. In many regional science applications, this is not very plausible

On the other hand, in the SDM, both the direct and indirect impact of a particular independent variable k will also depend on the estimate of θ_k.

Local spillover specifications

For the empirical situations where theoretical or substantive considerations indicate the potential occurrence of local spillovers, the proper specification is the **spatial Durbin error model** (SDEM),

$$y = X\beta + WX\theta + u$$

$$u = \lambda Wu + \varepsilon, \qquad \varepsilon \sim N(0, I\sigma^2)$$

Following the partial derivative perspective, $\partial y/\partial X_k'$, β represents properly the direct impacts, while θ expresses the indirect effects and hence the local spatial spillovers.

If $\lambda = 0$ SDEM reduces to the *spatial lag of X model* (SLX), while if $\theta = 0$ SDEM reduces to the *spatial error model* (SEM) (LeSage and Pace 2009).

Comparing the two model specification (1)

There are empirical situations where we do not know a priori whether the phenomenon under study is characterized by local or global spatial spillovers, and hence we do not know if we should use the SDM or the SDEM.

Testing which of the two specification better describes the true underlying data generating process is complicated because they are non-nested and they both collapse to the SLX model (LeSage and Pace 2009).

Elhorst (2010) proposed an interesting inferential strategy to make a valid comparison between the two models.

Comparing the two model specification (2)

According to Elhorst (2010), in order to find the best specification (closer to the true DGP) under the ML estimation approach, one should

1. estimate an OLS model and then use the robust LM-test by Anselin et al. (1996) to verify if the SAR model or SEM is more proper
2. if the OLS model is rejected in favour of the SAR, the SEM or in favour of both models, then the SDM should be estimated
3. test if SDM can be reduced to SAR;test if SDM can be reduced to SEM (LeSage and Pace (2009) shown that SDM reduces to SEM if $\theta + \rho\beta{=}0$)
4. if both restrictions are rejected, then SDM best describes the data
5. if only the SAR (or SEM) restriction cannot be rejected and the LM-test points to the SAR (or SEM) then the SAR (or SEM) best describes the data; otherwise the SDM is better

Comparing the two model specification (3)

6. if the SEM is favoured, estimate a SDEM and test if θ is significant
7. if the LM-test points to the OLS model, estimate a SLX model and test if θ is significant.

As an alternative model comparison strategy LeSage and Pace (2009) suggested a Bayesian procedure.

An illustration (1)

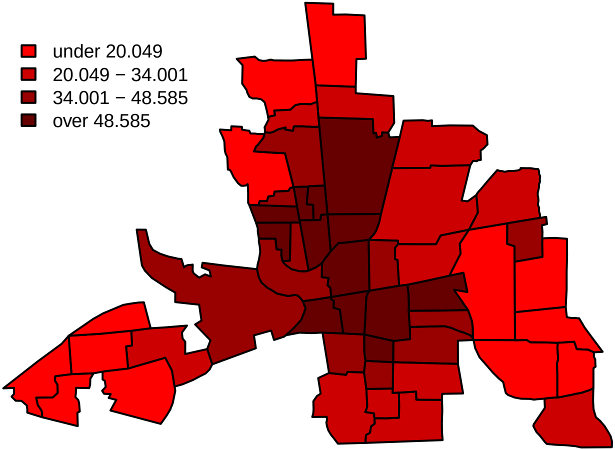
Following LeSage (2014a), as a way of illustration, let's consider a spatial regression analysis of the popular 1980 dataset "Columbus" from Anselin (1988) about 49 neighbourhoods in Columbus, OH, with variables

- HOVAL housing value (in $1,000)
- INC household income (in $1,000)
- CRIME residential burglaries and vehicle thefts per thousand households in the neighborhood.

The aim of the analysis is study CRIME as a function of INC and HOVAL while considering that increases in household income levels (or in housing values) in a neighborhood may lead to a reduction in crime in the own neighborhood but also in nearby neighborhoods.

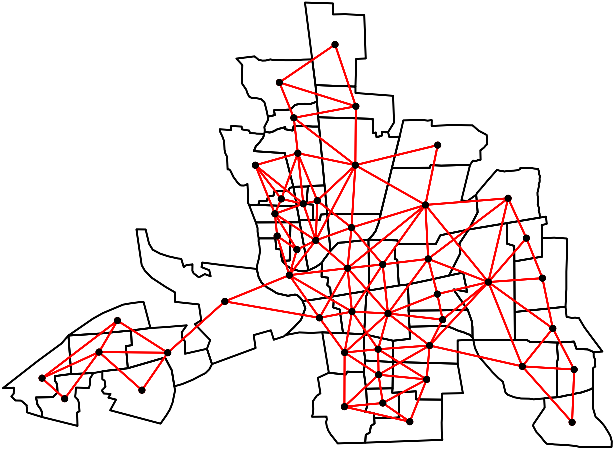
An illustration (2)

Spatial distribution of CRIME



An illustration (3)

Contiguity-based W matrix



An illustration (4)

The Moran's I test of spatial autocorrelation

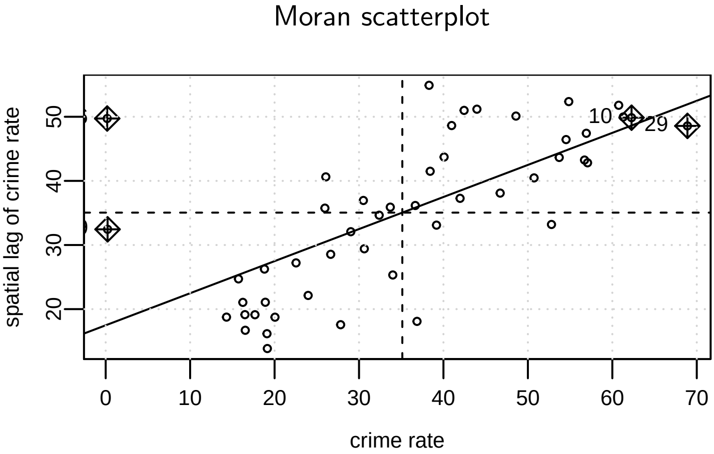
```
    Moran I test under randomisation

data:  columbus$CRIME
weights: listw

Moran I statistic standard deviate = 5.6, p-value =
1e-08
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
         0.500189         -0.020833          0.008689
```

Spatial autocorrelation of CRIME is statistically significant.

An illustration (5)



An illustration (6)

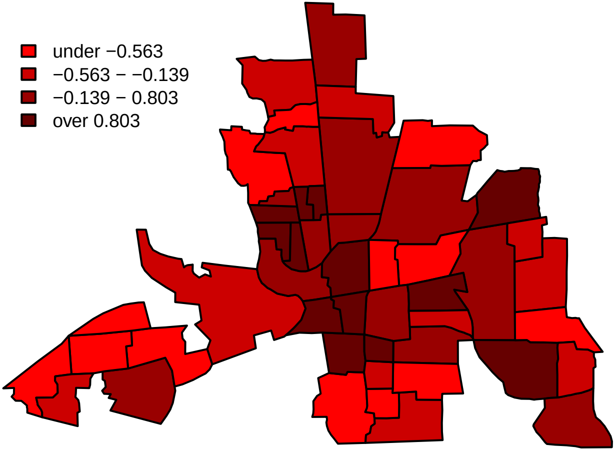
Let's apply the Elhorst (2010)'s strategy for best model selection. We start with the OLS model,

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  68.6190     4.7355  14.490 9.211e-19
INC          -1.5973     0.3341  -4.780 1.829e-05
HOVAL        -0.2739     0.1032  -2.654 1.087e-02
```

Then we perform the classic LM test and the robust LM test to see whether the SAR or the SEM are better description of the data.

An illustration (7)

Spatial distribution of OLS residuals



An illustration (8)

The Moran's I test of spatial autocorrelation in OLS residuals

```
    Global Moran I for regression residuals

data:
model: lm(formula = CRIME ~ INC + HOVAL, data =
columbus)
weights: listw

Moran I statistic standard deviate = 2.4, p-value =
0.009
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
        0.178521        -0.033418         0.008099
```

An illustration (9)

Anselin et al. (1996)'s LM test

```
...
    Lagrange multiplier diagnostics for spatial
    dependence

       statistic parameter p.value
LMerr     5.2062         1  0.0225 *
LMlag     8.8980         1  0.0029 **
RLMerr    0.0439         1  0.8340
RLMlag    3.7357         1  0.0533 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
...
```

An illustration (10)

Since we have quite evidence in favour of the SAR model, we estimate the SDM,

```
...
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 44.320005  13.045474  3.3973 0.0006804
INC         -0.919906   0.334742 -2.7481 0.0059941
HOVAL       -0.297129   0.090416 -3.2863 0.0010153
lag.INC     -0.583913   0.574225 -1.0169 0.3092139
lag.HOVAL    0.257684   0.187235  1.3763 0.1687404

Rho: 0.4035, LR test value: 4.663, p-value: 0.030825
...
```

An illustration (11)

Now, with the likelihood ratio test, we can see if the SDM can simplified to the SAR, on one hand; and to the SEM, on the other hand:

```
    Model df AIC logLik Test L.Ratio p-value
SDM     1  7 377   -182    1
SAR     2  5 375   -183    2    2.07   0.355
```

```
    Model df AIC logLik Test L.Ratio p-value
SDM     1  7 377   -182    1
SEM     2  5 377   -184    2    4.22   0.121
```

Both restrictions cannot be rejected, however the (robust) LM test points to the SAR model and the hence this is the specification that should be adopted.

An illustration (12)

Finally, we can estimate and test the impacts under the SAR model specification,

```
...
Impact measures (lag, exact):
       Direct Indirect   Total
INC   -1.1009  -0.7177 -1.8186
HOVAL -0.2796  -0.1823 -0.4618

Simulated p-values:
      Direct  Indirect Total
INC   0.00055 0.044    0.0011
HOVAL 0.00365 0.117    0.0166
...
```

An illustration (13)

Results show, in particular, that

- negative (and significant) direct impacts of changes in both neighborhood household income and housing values, implying that higher levels of these variables in a neighborhood lead to reductions in crime rates in that same neighborhood.
- the significant indirect impact associated to household income income indicates that, on average, because of spatial spillovers a change in household income in a neighborhood would also lead to a cumulative decrease of 0.7177 criminal events per 1000 households.

References

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers (Dordrecht).

Anselin, L., A. K. Bera, R. Florax, and M. J. Yoon. 1996. “Simple Diagnostic Tests for Spatial Dependence.” *Regional Science and Urban Economics* 26 (1): 77–104.

Elhorst, J. P. 2010. “Applied Spatial Econometrics: Raising the Bar.” *Spatial Economic Analysis* 5 (1). Routledge: 9–28.

LeSage, J. 2014a. *Spatial Econometrics*. Edited by Edward Elgar Publishing Limited. *Handbook of Research Methods and Applications in Economic Geography*.

———. 2014b. “What Regional Scientists Need to Know About Spatial Econometrics.” *Review of Regional Studies* 44 (1): 13–32.

LeSage, J., and R.K. Pace. 2009. *Introduction to Spatial Econometrics*. Statistics: A Series of Textbooks and Monographs. CRC Press.

Tobler, W. R. 1970. “A Computer Movie Simulating Urban Growth in the Detroit Region.” *Economic Geography* 46. Clark University, Wiley: 234–40.