Street Value: Assessing Non-Euclidean Metrics in Spatial Analysis Applied to Property Values

Chris Brunsdon
Department of Geography and Planning
School of Environmental Sciences
University of Liverpool
Christopher.Brunsdon@liverpool.ac.uk

Sunday 11th November, 2012

1 Introduction

Many spatial analysis techniques depend on notions of distance, or spatial proximity. However, representing these attributes quantitatively is not straightforward. Many models and associated analysis techniques exploit the ideas underlined by *Tobler's First Law* -

> "Everything is related to everything else, but near things are more related than distant things" Tobler (1970)

however, it is not always clear what 'near' means in this context. The simplest definition, if the data being analysed consists of spatial point data, is to work with Euclidean (or straight line) distances - however this is not always a reliable indicator of proximity. A number of alternative measures of proximity are possible - for example travel times or travel distances based on routes on roads connecting locations. In this study a number of spatial analytical techniques for predicting property prices on the basis of spatial trend are compared. For this initial study, variability in prices due to the size and type of property is accounted for by considering only two-bedroom flats. Future investigations will consider a fuller study

of properties and make use of regression models whose explanatory variables are characteristics of the property (hedonic models).

As well as being a useful way of evaluating the methods used to represent distance, the application area of property price analysis is important in its own right - for example measuring the effect of location on property value is an important tool in a number of tasks associated with regional economic analysis - for example the effect on property values of being located near an area of woodland can be assessed by identifying spatial trends (Powe et al., 1997) .

In this extended abstract a brief overview of the techniques being considered will be presented, together with a comparison of their performance when applied to a data set of property values in Liverpool. Approaches based on local modelling are found to perform better than those based on modelling spatial autocorrelation.

2 Overview of Techniques

2.1 Spatial Statistical Approaches

In many spatial statistical approaches, Tobler's law is incorporated by allowing random error terms to exhibit spatial autocorrelation, where instead of the classical assumption that the error terms are independent, it is assumed that they are correlated, and that the error term is related to proximity. For example a dependent variable y_i (here, property value) for observation i is predicted using a regression model

$$y_i = \alpha + \sum_{k=1,j} \beta_k x_{ik} + \varepsilon_i \ . \tag{1}$$

with the x_{ij}'s being predictor variables - for example characteristics of each propertyy. However, rather than assuming independence between ε_i and ε_j as in the standard OLS model, for observations are assumed to be correlated, with correlation being related to the nearness of these observations. However, one characteristic of these models is that although there is an unobserved parameter (or set of parameters) relating proximity to correlation which are calibrated when the model is fitted, the proximities themselves are taken as given. For example Dubin (1992) models the correlation between the error terms as

$$\rho_{ij} = \exp(-d_{ij}/\lambda) \tag{2}$$

where d_{ij} is the distance between observations i and j. The corresponding covariance for each i and j is then $\Sigma = \sigma^2 [\rho_{ij}]$, where σ^2 is the variance of the error terms. The error terms are also assumed to have zero means. In the case study here, a particular subset of properties having similar characteristics are chosen (that is, they are all two-bedroom flats) and therefore the analysis is simplified so that

$$y_i = \alpha + \varepsilon_i \tag{3}$$

is the model used - again with the error terms having a covariance structure as used by Dubin.

However, although the parameter λ is estimated from observed data - here it is done using a method set out by Breusch (1980) - the values of the d_{ij}'s are prescribed in this model specification. Here the effect of varying the definition of the d_{ij}'s will be considered. The paper by Dubin also demonstrates how a Kriging-based approach may be used to predict property values at locations where no y-value is supplied.

However, one technical issue must also be considered when doing this, and this is outlined in the following subsection.

2.2 Positive Definite Correlation Functions

Not all matrices Σ may be covariance matrices. A necessary and sufficient condition is that Σ is *positive definite* - this means that for any real-valued vector $\boldsymbol{a}$ with n elements we have

$$\boldsymbol{a'}\Sigma\boldsymbol{a} \geq 0$$

with equality only occurring if $\boldsymbol{a} = 0$. Essentially this implies that any linear combination of the error terms has a well-defined distribution, since if the condition did not hold the prospective variance of some linear combinations of error terms would be negative. A function relating distances d_{ij} to Σ is called a *positive definite function* if it always yields a positive definite Σ for any distance values and any parameters λ.

One issue with the above approach is that the correlation function defined in eq. 2 is always well-defined for any value of λ when d_{ij} is derived from Euclidean distances, but not necessarily so for other distance metrics (Armstrong and Diamond, 1984). An implication of this is that it is important to identify ranges of λ for which Σ *is* positive definite if working with non-Euclidean distance metrics. In practice this can be done by plotting the lowest eigenvalue of Σ against λ and identifying ranges for which this exceeds zero. This is because a positive definite matrix has the property that all of its eigenvectors are greater than zero.

2.3 k-Nearest Neighbour Modelling

This technique owes more to the discipline of machine learning than inferential statistical approaches based on a *data model* such as that in eq. 1, since *from the outset* it specifies an algorithm for predicting property values given other values in the area, rather than deriving such an algorithm via a principle such as maximum likelihood applied to a data model. The algorithm itself is straightforward. For each y_i, select the k observations closest to observation i and predict y_i to be the arithmetic mean y-value of these observations. In plain terms, just take the average of the k nearest property prices.

Two issues arise here. Firstly, this approach does not take into account explanatory variables. In this case this is not an issue due to the fact that the dataset is intentionally restricted to properties of a single type. However the question of how this approach can be generalised is still a pertinent one. One approach is to sub-divide the data set into subsets according to a classification of property characteristics - and apply the algorithm to each subset in turn. The classification used is likely to depend on the level of detail in the data - for example one could just subset the data by property type, but if more detailed attribute data about properties were available, another option might be to use cluster analysis, and subset the data in terms of the clusters.

Secondly, some consideration must be given to the choice of k. The method suggested here is by *leave-one-out* cross-validation. In this approach, each observation y_i in turn is removed from the data set, and on the basis of the remaining $n-1$ observations $\{y_1, y_2, ..., y_{i-1}, y_{i+1}, ..., y_n\}$ obtain a prediction (with k neighbours) of y_i, say $\hat{y}_{-i}(k)$. The overall accuracy of these predictions is measured by

$$\text{mse}(k) = \sqrt{\frac{\sum \left(y_i - \hat{y}_{-i}(k)\right)^2}{n}} \tag{4}$$

and the value of k minimising this expression is chosen. The reason that y_i is removed from the data is that otherwise the expression in eq. 4 would always be maximised by $k = 1$.

Finally, as in section 2.1 two definitions of distance will be used to determine the k nearest neighbours. These will be the Euclidean and road-based distances.

2.4 Moving Window Smoothing

A similar, but alternative method to k-nearest neighbours is that of *moving window smoothing*. In this, the each y_i is predicted as a *weighted* mean of its neighbours, where the weight is a decreasing function of distance. In this case the value of a property is still predicted by those nearby, but in this case the distance itself determines the influence that a nearby property price has, rather than just using a cut-off of the k nearest neighbours, and beyond this, no further weighting. For the study here, the weight, when predicting y_i will be

$$w_{ij} = \begin{cases} \left(1 - \frac{x^2}{h^2}\right)^2 & \text{if } x < h \\ \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where h is the *bandwidth*. The value of this determines how nearby other properties must be, to influence the prediction. The predicted value for y_i will be

$$\hat{y}_i = \frac{\sum_j w_{ij} y_j}{\sum_j w_{ij}} \tag{6}$$

As in section 2.3 the influence of attributes of properties can be accounted for by sub-setting. Also h can be chosen by leave-one-out cross validation. Again, as with the k-nearest neighbours approach, both Euclidean and road-distance metrics will be used.

3 A Practical Example

3.1 Housing Data

The house price data used for this example were obtained from the *Nestoria* web site[1]. It is possible to interrogate this web site via an *Application Programming Interface* (API) - enabling asking prices for properties, together with their location as latitude and longitude to be downloaded. The data are shown in figure 1. Here, each semi-transparent red disc represents the location of a property, and its radius reflects the asking price. There are clearly some spatial trends - although there are also some cluster with notable heterogeneity - for example south of Liverpool near to Sefton Park, and also in Birkenhead. Fig. 2 shows the distribution of asking prices as a boxplot.

[1] `http://www.nestoria.co.uk/`

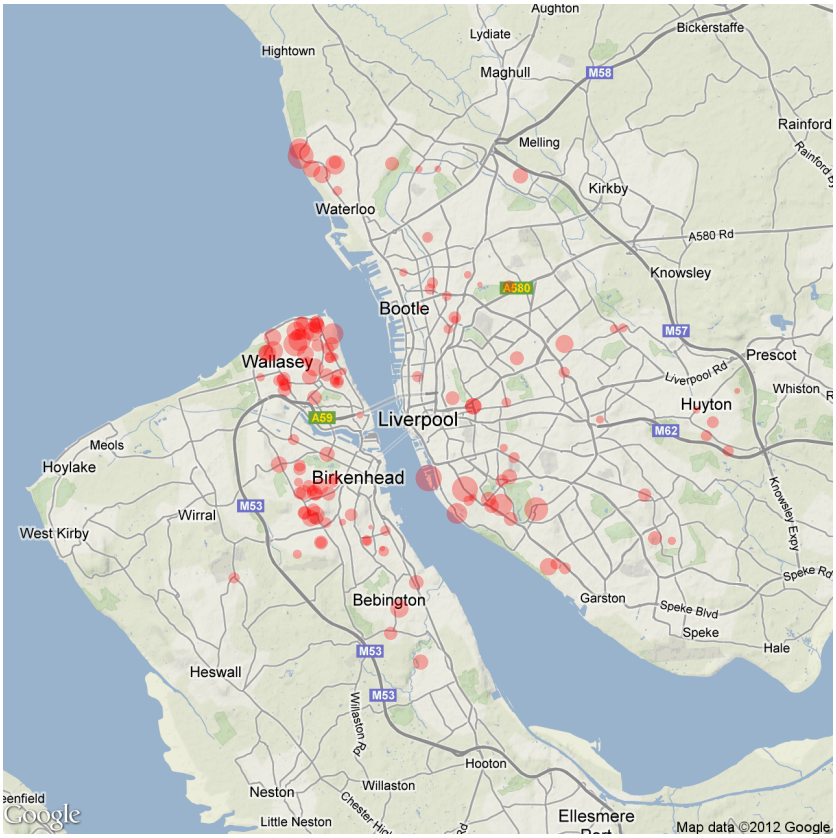

Figure 1: Locations of Properties

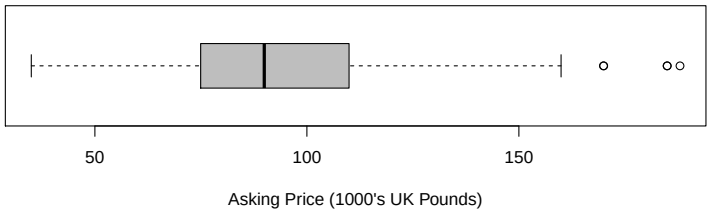

Figure 2: Box Plot of Property Asking Prices

3.2 Distance Estimation

As stated above, the techniques considered in the previous sections will be evaluated by being applied to data relating to property prices. Also as stated, the techniques will be applied using two different ways of representing distance - Euclidean distance and road distance. Although the house locations are supplied as latitude and longitude they may be converted to OS National Grid coordinates via the `proj4` program[2]. From this, Euclidean distances were computed.

For road-based distances (essentially network distances), the information was obtained via the Routino program[3]. Essentially his is a routing program making use of OpenStreetMap data[4]. Although at the time of writing it is not possible to make calls to Routino via a subroutine library, it is possible to write scripts to identify routes between each property location, and from these to obtain the road distance between each property pair.

3.3 Valid Road-Based Distance Models

Recall that when working with non-Euclidean metrics and using models such as that specified in eq. 2 care must be taken to ensure that the coefficient λ yields a valid autocorrelation matrix for the random error terms $\{\varepsilon_i\}$, and that this may be achieved by identifying values of λ for which the smallest eigenvalue of the correlation matrix is positive. This is shown in fig. 3 - and indicates that λ yields a valid model only for values below around 0.5km.

3.4 Visualisation of Road Distances

A useful method for assessing the difference between road distance and Euclidean distance for the data is to use multidimensional scaling (Shepard, 1966). Here, given a set of non-Euclidean distances between entities, a set of points whose Euclidean distance matrix matches the given matrix as closely as poossible is found. Applying this algorithm to the road distance matrix for the Liverpool property price data yeilds the spatial arrangement seen in fig. 4. As before, the discs represent properties and thier radius reflects asking price. From this, perhaps the most notable effect is that the groups of properties on each side of the River Mersey are moved further apart than they are in reality. This reflects the fact that, as is often

[2] `http://trac.osgeo.org/proj/`
[3] `http://www.routino.org`
[4] `http://www.openstreetmap.org`

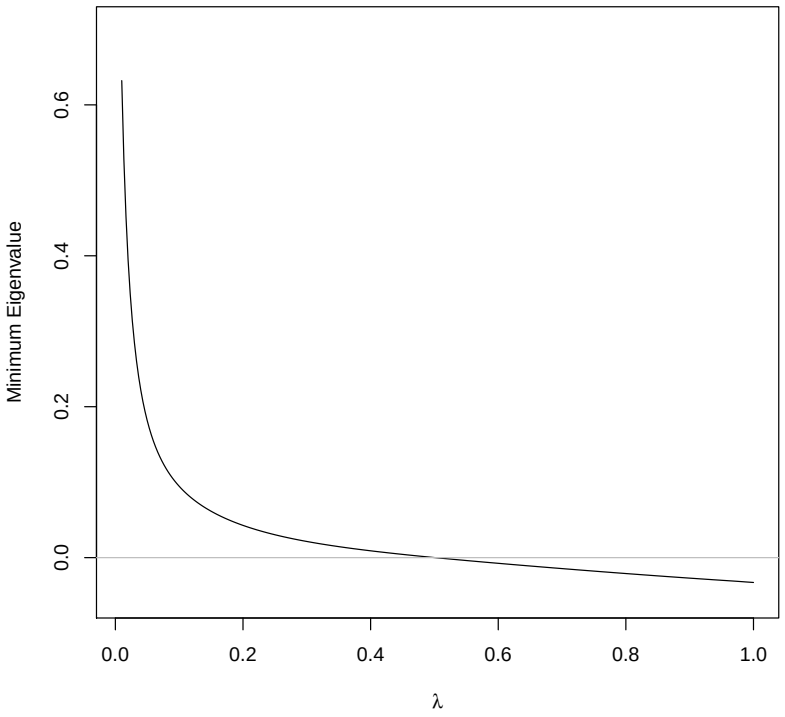

Figure 3: Minimum Eigenvalue vs. λ for Road Distance

the case with rivers, there are a relatively small number of crossing points, and if a pair of properties are on different sides of the Mersey, the road distance is notably greater than the Euclidean distance as one often has to make a journey to a bridge to cross the river.

4 Results of Analysis

The three approaches to modelling were fitted to the property value data, with both Euclidean and road-distance metrics. The results (in terms of the scoring method specified in eq. 4) are listed in table 1. The scores here are for the best-performing k (k-nearest neighbours) and h (moving window smoothing). For the modelled autocorrelation approach, the scores are obtained by predicting each y_i by cali-

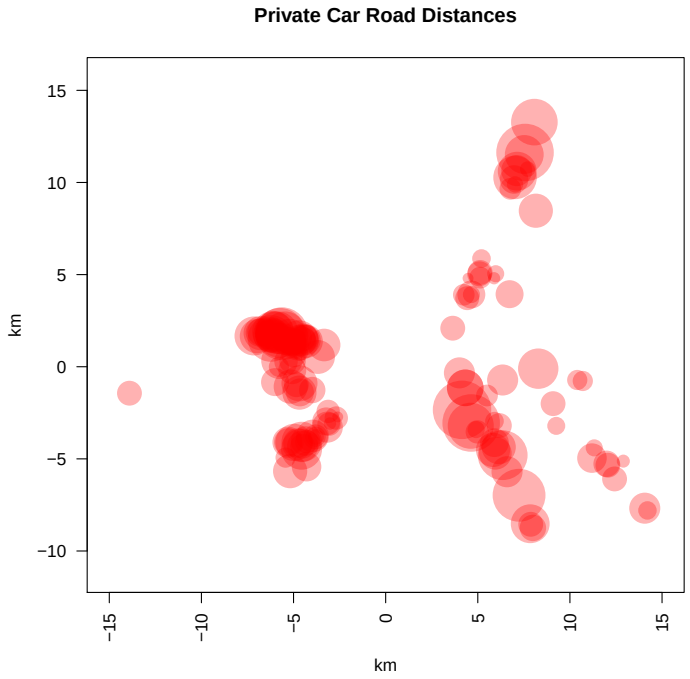

Figure 4: Multidimensional Scaling via Travel Times

brating the the correlation model on the data without observation i. From this it may be seen that the best performing approach is the moving window approach, when Euclidean distances are used. The autocorrelation approaches performed relatively poorly. For each distance metric, the values of λ were very low (0.09 for the Euclidean and 0.12 for road-distyance), suggesting virtually no spatial autocorrelation. In both cases, the predicted values for the y_i's were very close to the mean for the entire sample of prices.

5 Conclusions

The two notable findings here were that the Euclidean distances outperformed the road distances, and the two machine-learning based approaches performed better than the autocorrelation modelling approach. THe abest performance of Euclidean distance may be attributable to the fact that these are property *asking* prices - which are generally set on the advice of estate agents - and the perceptoons of estate

Table 1: Root mean square scores for property value predictions - † denotes best performance.

Metric \ Method	k Nearest Neighbour	Moving Window	Autocorrelation
Euclidean	27.37	† 26.56	31.03
Road Distance	27.50	27.39	31.03

agents may be better represented using Euclidean space. The poor performance of the autocorrelation modelling approach is perhaps explained by the fact that the autocorrelation relationship is modelled as being *universal* - so that residuals with the same degree of separation will always have the same correlation regardless of location. Examining the map suggests that there are some places where this correlation is strong, but others where it is not. This 'one-size-fits-all' approach may lead to poor performance. The other techniques rely only on relatively local data, and arguably it is this characteristic that causes this problem to be avoided.

The implications of this are most pertinent to econometric studies, where frequently autocorrelation approaches are often adopted uncritically. Although this is a limited study and there is clearly a need to consider a broader data set and further analytical approaches, the findings here suggest that there are situations in which they do not encapsulate the property pricing process particularly well.

References

Armstrong, M. and Diamond, P. (1984). Testing variograms for positive-definiteness. *Mathematical Geology*, 16(4):407–421.

Breusch, T. S. (1980). Useful invariance results for generalised regression models. *Journal opf Econometrics*, 13:327–340.

Dubin, R. A. (1992). Spatial autocorelation and neighbourhood quality. *Regional Science and Urban Economics*, 22:433–452.

Powe, N. A., Garrod, G. D., Brunsdon, C. F., and Willis, K. G. (1997). Using a ge-

ographic information system to estimate an hedonic price model of the benefits of woodland access. *Forestry*, 70(2):139–149.

Shepard, R. N. (1966). Metric structure in ordinal data. *Journal of Mathematical Psychology*, 3(287–315).

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240.