THE APPLICATION Of SPATIALLY DERIVED LOCATION FACTORS Within A GIS ENVIRONMENT

Professor William J. McCluskey
University of Ulster/
University of Lincoln NZ
Mccluskw@lincoln.ac.nz

Mr William G. Deddis
University of Ulster
WG.Deddis@ulst.ac.uk

Ian Lamont
Causeway Data Communications Ltd.
Ian@cdc-ni.com

This paper is drawn from research (funded by the RICS Education Trust) and disseminates the results to date. The aim of the research was to measure the effect of location on residential house prices/value perceptions. The methodology attempts to develop models that integrate spatial and a-spatial data within a GIS environment. Ideally, the end product of the research will be a working model with the capacity to predict residential property prices/values.

Introduction

The inherent spatial immobility of land and property means that location is an intrinsic attribute of a dwelling, which is a direct determinant of housing quality and market value. Significant differences in value can occur over short distances and valuers generally agree that location is the most important factor affecting value. This has been reiterated on many occasions, Fraser in 1991 concluded that "*of dominant importance in understanding the demand for any property is its location both in a regional as well as a local sense*". Modelling location, for the general purpose of property valuation, has proved difficult. The wide range of spatially defined attributes, that may or may not affect perceptions of value at given points in time, are notoriously difficult to standardise. There are many contributory factors/variables, such as 'what is the condition of the urban infrastructure and road network?' among numerous others. It is common knowledge that ***'valuers'*** infer a substantial amount of information about a property from its location, which in turn is based on local knowledge and experience. Banton (1993) noted that the problem facing any valuer is the abundance of data, relative to the valuers' capacity to assimilate and effectively incorporate all the potential value influences within the valuation process. Furthermore the valuer, has to rationalise the relativity between the various value impacts both 'negative' and 'positive', before arriving at a final figure.

Indeed, understanding why a particular location exerts a particular influence on property value is quite different to attempting a measurement of that locational factor. (Ref) Our research has served to confirm the fact that in the real world 'fuzzy zones' of spatial trends occur as opposed to distinct areas of homogeneous property subsets. (Ref) The most common approach adopted for examining the effect of locational accessibility on house prices is that which includes a distance variable from the Central Business District, (CBD). This implicitly assumes that locational distribution is monocentric, but clearly, a multiple nuclei/multi-centric model incorporating a concentric pattern is much more appropriate given the existence of urban sub-centres.

Academic research has explored a number of techniques and avenues of approach to the question, how can one successfully factor location into valuation practice. The Scientific Geography Series (Thrall 1984 -1988) disseminated relevant work which sought to bridge what was an obvious gap between spatial geography, economic disciplines and property valuation processes. Fleming and Nellis (1984) considered the application of hedonic modelling to house prices and concluded that it provided a worthwhile means of exploring price determinants and movements. This however was out-with a GIS environment. Pitman *et al*, (1990). Rodriguez *et al* (1995) pointed out that given each location is influenced by other locations, econometric analysis must be concerned not only with the possibility of errors due to the problem of time auto-correlation but also examine the problems of spatial correlation. In the recent past Wyatt (1995/6/7) agreed that locational influences on property value are of the utmost significance, yet within the valuation process remained largely implicit. The deductive ability and intuitive capacity of a skilled Valuer has largely accounted for this ability to factor in location influences. Wyatt also considered the potential of several computer based techniques which could undertake valuations. In this context Wyatt explored the use of Expert Systems, Neural Networks (as did Borst 1992) and Multiple Regression Analysis (MRA) but found each lacking in some important respect. Expert Systems failed effectively to deal with or represent temporal or spatial knowledge. Neural networks was too complex and regarded as something of a 'Black Box' technique which did not have sufficient transparency in particular with regard to settling valuation disputes. Multiple Regression Analysis has also been utilised to explain and predict property prices (Adair *et al* 1987, Fraser *et al* 1989, Czernowski 1990). Each recognised the problem of seeking to incorporate location by analysing small homogeneous areas within which it was assumed that residential properties were considered to share similar locational attributes. The problems with such techniques essentially rested in the simplistic nature of these basic assumptions. Additionally, the difficulty of defining the market area or neighbourhood and the question of sample sizes are also problems. Wyatt (1997) concluded that the MRA technique was suspect because of multi-collinearity and the spatial auto-correlation that could occur in geographical analysis. McCluskey *et al* (2000) accepted that from a mass valuation perspective to devise a predictive model requires the sub-division of a significant geographical area into realistic sub-markets or neighbourhoods. Only by this sub-division can one enable the model to more accurately reflect the influence of location. MRA sits quite comfortably within a mass valuation model and is considered the primary technique.

The Mass appraisal approaches

Various methods of incorporating location as a factor within mass appraisal techniques have been investigated; all of which require an assessment of neighbourhoods or sub-markets.
The housing market is a set of distinct but interrelated sub-markets, encompassing dwellings differentiated by one or several alternative dimensions, which arise due to the joint nature of structural and locational attributes. Accordingly, attributes and sub-market locational features are essential ingredients in predicting prices. Whilst the existence of sub-market locational features is *a priori* accepted, consensus on whether the sub-markets should be defined in spatial terms according to property characteristics or based on the actual house price is less evident. (Ref)

Sub-markets can be defined in a number of ways by employing the principle of stratification: a process of creating a number of homogeneous segments from a larger heterogeneous database. In a spatial context it is possible to create localised regions formed through the aggregation of areal units, such as postal zones, enumeration districts or ward boundaries. The use of 'political' or other non-property based locational areas creates problems related to boundary positioning. A ward (political electoral unit) boundary may in fact divide a homogenous area, which creates a non-optimal submarket. Alternatively, a submarket or neighbourhood can be created on the basis of environmental or locational characteristics. Another approach is based upon the quantitative characteristics of the dwellings such as house type, size, age etc. In addition, analysis of house prices can be used as the determining factor in identifying sub-groups. From a mass appraisal modelling perspective it is essential that a large geographical area be divided into realistic submarkets or neighbourhoods to enable the model to more accurately reflect the influence of location. Neighbourhood quality is arguably an unobservable variable although subjective assessments may be made on a variety of perceived quality indicators, ranging from site sizes and housing quality through infrastructure and environmental conditions to social structure and employment levels. There is no consensus in the literature regarding which variable best is deemed the most appropriate measure of neighbourhood quality.

In terms of mass appraisal modelling the primary technique utilised is multiple regression analysis. This traditional econometric approach, of using regression analysis on housing value as a function of various structural, accessibility and neighbourhood attributes of dwellings can be extended. Incorporating a spatial element within what is typically an aspatial modelling technique requires the proper specification of the spatial regressor. Generally speaking it is possible to derive individual models for each discrete submarket or alternatively to employ an overall model encompassing several neighbourhoods, where each neighbourhood enters into the model as a dummy or dichotomous variable. The application of separate models for stratified homogeneous subsets induces a problem of sample size, which could result in statistically unsound and biased results. Alternatively, the approach of using dummy neighbourhood variables to reflect the influence of location does have an intuitive appeal, but presupposes that the affect of that location is uniform across all properties within a particular neighbourhood. This form of delineation can be construed as static, which results in the potential effects of spatial trends being ignored. Since location is usually the most important variable in real estate value, it is imperative to account for locational variation before attempting to derive definitive market coefficients for individual property attributes, given that such variables as age, plot size are highly correlated with location.

An alternative to modelling on specific neighbourhoods is to accept the premise that location is an amalgam of several environmental factors which when combined attribute specific value to a location. For example, locational influences may arise from a number of sources such as assessibility, environmental factors, and neighbourhood amenity.

Contemporary GIS (or spatial analysis) approaches

Within a GIS framework the use of surface response analysis techniques has been shown to provide a three dimensional visualisation of the value of location as it varies geographically. Whilst research in this area has been limited, the work carried out to date has contributed to a better understanding of the measurement of locational effects.

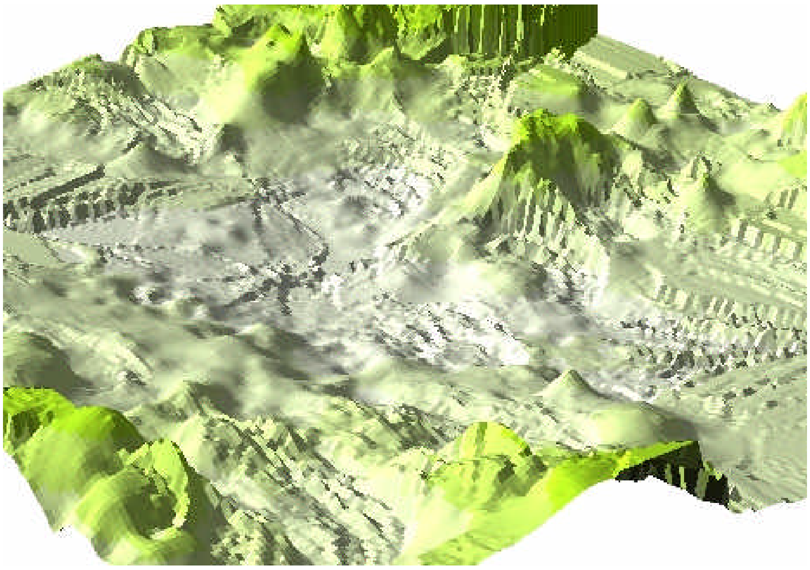

Figure 1, Example of 3D Visualisation

Our research included various spatial analytical techniques such as Triangulated Irregular Networks (TINs), Surface Response Analysis and a number of interpolation options (Inverse Distance Weighting (IDW) Spline and Kriging). Furthermore GIS was used as a stand alone analytical tool to predict residential property values, combined MRA within a GIS, and refining the models.

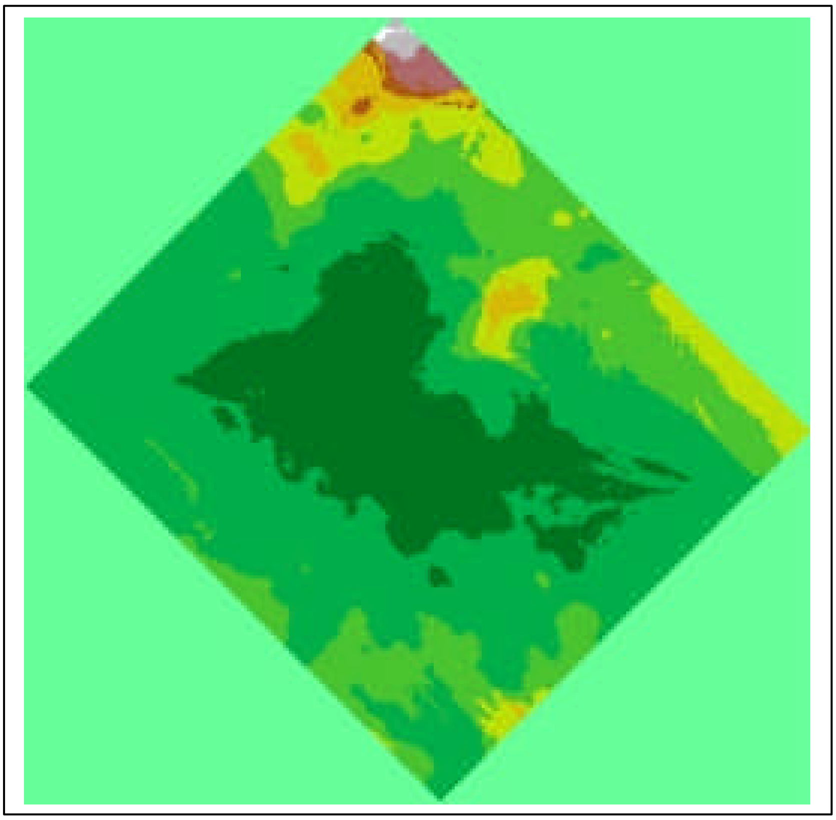

Figure 2, Example of Krigged Surface

Current Work

The applied use of surface response techniques, such as IDW, Spline and Kriging methods, within the wider remit of CAMA generated interest within recent literature. The Kriging interpolation technique worked best in terms of predictive quality and explainability. Crucially, this technique produces a variance grid that can be further utilised within a correlative process. To better understand how a model is performing, one must appreciate two things:

- the data
- the mathematical operations within the model.

The Data

The data used initially in the early research comprised property sales information from a small town (population circa 40,000) within easy travelling distance. This convenience, facilitated data verification and the screening of transactions. Local authority house sales (at discounted price) and first time sales from local builders were extracted, thus cleaning up the sales data. One of the drawbacks with the first choice study area, in geo-spatial analysis terms, was the disintegrated pattern of the residential neighbourhoods. To overcome these initial problems data from Londonderry (population circa 90,000 and the second largest city in Northern Ireland) where residential neighbourhoods exhibited greater spatial integration, was utilised. The Valuation and Lands Agency kindly facilitated data verification by appointing one of their valuation/appraisal staff to verify and screen the data.

Valuation and Lands Agency Data

Data comprising, all residential sales within the local authority area (over a period from Jan 1998 to Dec 1999) was provided. The data set comprised 800 sales before screening and some 650 after screening. The attributes associated with each property included the sale price, date of sale, age of the property, size (gross external area), number of bedrooms, number of bathrooms, number of garages, type of central heating, physical condition, and neighbourhood quality index and group cluster.

Ordnance Survey Northern Ireland

Ordnance Survey Northern Ireland provided the digital topographic map base required for the study area. This comprised some 200 levels of information held as attributes against vector geometry. In addition an ‘associated address database’ is maintained for all addressable properties held as attributes of each building polygon. Each record provides a spatially unique property reference number in the form of a geocode. This is a centroid within each building polygon accurate to within 1m on the ground. Linking the two data sets was achieved by merging via a concatenated postal address and thus provided the ability to plot each of the VLA records.

The Mathematical Operations

Building on our previous work we persevered with our methodological approach and the utilisation of kriging interpolation techniques.

Kriging

Kriging, or the regionalised variable theory, is a geostatistical interpolation method often used in preference over other interpolation techniques due to its ability to describe the certainty with which it computes values. A continuous variable (or attribute), such as price per square meter, will not occur in a regular fashion and thus cannot accurately be modelled by a smooth mathematical model. The variations can be better described by a stochastic surface with the attribute being known as the regionalised variable. (Burrough and McDonnell, 1998.)

Regionalised variable theory assumes three fundamental components:

- A structural component having a constant mean or trend
- A random but spatially correlated component known as the variation of the regionalised variable
- A spatially uncorrelated random noise or residual error term.

This can be shown by the following equation:

$$Z(x) = m(x) + \varepsilon'(x) + \varepsilon''$$

where *m(x)* is a deterministic function describing the structural component, *e'* is the term denoting the locally varying but spatially dependent residuals from *m(x)* and *e''* is a residual, spatially independent Gaussian noise term having zero mean and variance expressed as σ^2. This 'noise term' is more commonly known as the nugget. The information within the variograms can also be useful to help determine where additional sample data should be recorded to improve the predictive accuracy of the model. In terms of CAMA modelling this is crucial information. It may be that sales data is not forthcoming for particular areas and so appropriate remedial action, such as individual assessment, may be useful. This additional information could create a more robust system and should theoretically improve the predictive quality of the model.

Considering an appropriate model

In considering a model process one must first evaluate the data. Two issues are important :

Firstly, the intensity and degree of variation of events across the region,.

Secondly, the question as to what level of this variation is

(a) due to heterogeneity rather than

(b) second order dependence (Bailey and Gatrell, 1995).

Semi-variograms, were used to assess the data prior to creating surfaces and generating residual error. The following examples of semi-variograms illustrate the predictive differences that may be generated by using various methods.

Methodology

All modelling, surface interpolations and analysis was performed using several GIS packages including ArcView GIS, ArcView Spatial Analyst, Idrisi GIS and S-Plus. The methodology comprised several different stages of application but in general terms included the following processes.

- geocode the sales data
- regress data to create a temporal standard
- derive a random sample of 33% for hold out purposes from the original sample.
- assess the data using semi-variograms
- run geo-statistical procedures on data
- generate grids using appropriate Kriging interpolation techniques
- determine neighbourhood clusters and influence centres
- extract values and include as a regressor within MRA
- generate COD and compare results with traditional MRA

Study Area Interpolations

The Derry Study Area exhibits typical results from various Kriging interpolation methods - some areas perform well within the model, other areas perform badly. The results suggest no apparent geographic bias, which we found interesting. To further investigate this amendments were made to the semi-variograms and to the lags of various Kriging interpolations. The 'best fit' semi-variogram used a 1200 lag exponential series. The others models performed quite poorly. A surface was then generated using this semi-variogram and the holdout overlaid. The resulting COD and Adjusted R^2 values were still quite poor in predictive terms.

The results suggested that pure surface models cannot be used to accurately predict residential property value without either (1) removing all outliers or (2) incorporating a neighbourhood based lattice to provide accurate samples on which to build the model. Clearly (1) is unworkable and although there is some merit in (2), in a real world application sense neither method offers a practical solution.

Recent work by McCluskey *et al* has proved that the integration of a location factor into Multiple Regression Analysis improves the predictive quality of the model. This factor has normally been some sort of areal unit such as an administrative area. The distribution of each areal unit will not be homogeneous and so the influence of micro-markets will be inhibited so although this has improved the model, in terms of a pure 'location' factor it is still weak. Research performed using various other forms of stratification, such as Neighbourhoods, experienced the same problems. It is difficult to determine 'appropriate' neighbourhoods without reducing the size of these areal units to cover very small areas. The distribution within the study area showed marked differences between minimum and maximum sales values within one neighbourhood as Figure 3 illustrates.

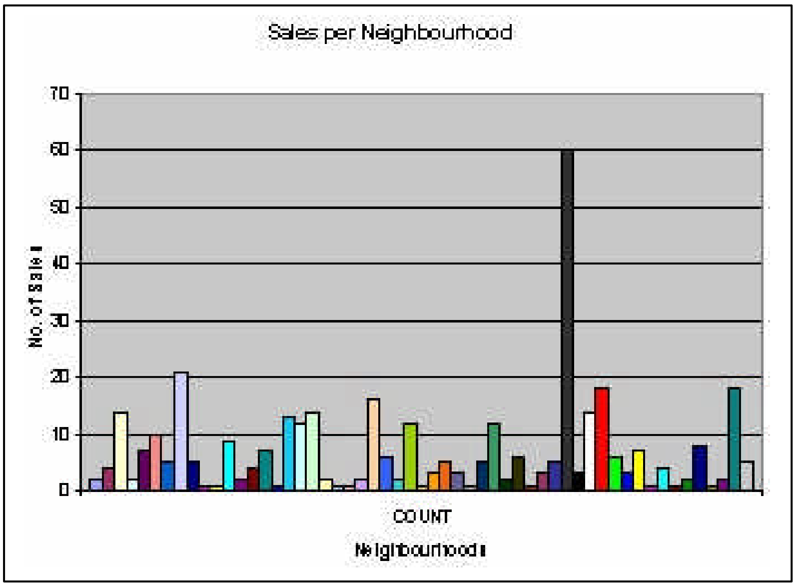

Figure 3. Sales per Neighbourhood.

Furthermore, these issues are compounded by the low number of sales in some units as compared to others. The distribution of the recorded sales information is clustered into micro-neighbourhoods with the majority of sales located in one neighbourhood. The neighbourhoods with a scarcity of sales data may not prove the most robust when used as binary regressors. Clearly excessive stratification requires additional data preparation and the results do not always warrant such efforts.

However, as some regions within the interpolated surface performed very well we decided to attempt further work. We incorporated the use of multiple surfaces using a variety of semi-variogram models for selected regions within the study. As we expected the predictive quality increased. Our work then focused on how the distribution affects the predictive quality of models. The premise of our research being that if the distribution is known then 'distribution aware' models can be used to improve predictive quality. This work utilises empirical modelling procedures through advanced surface interpolation techniques and is still ongoing although initial results are encouraging.

Conclusions and Further Work

The research initially sought to identify an appropriate method by which a location factor might be generated to augment existing multiple regression techniques. The fundamentals of Spatial Statistics and Geo-Spatial Modelling follow the basic premise that objects which are in close proximity to each other are likely to exhibit similar characteristics (a normal distance decay function). Likewise, as the distance between subject and object increases, the likelihood of the characteristics being similar diminishes.

The application of geo-spatial statistical analysis within a CAMA model may provide an efficient link between pure spatial tool sets, such as the integration of a GIS tool-set within a CAMA environment. As suggested previously in this text, it is unlikely that results generated purely from a GIS, such as single surfaces, will provide a sufficient level of predictive accuracy to be used without additional statistical support. Further work is needed to assess the extended use of clustering techniques for the generation of 'distribution aware' models.

Bibliography

Adair, A. S. and McGreal, W. S. (1987) "The application of multiple regression analysis in property valuation", Journal of Valuation, Vol.6, P57-67.

Adair, A. S; Berry, J.N and McGreal, W. S. (1996) "Hedonic modelling, Housing Submarkets and Residential Valuation", Journal of Property Research, Vol 13, P67-83.

Banton, H.S. (1993) "GIS plays a role in real estate appraisal and analysis", GIS World Inc., Vol.6, No.4, P62.

Borst, R.A. (1992) "Artificial Neural Networks: The next Modelling/Calibration Technology for the Assessment Community", Artificial Neural Networks, 1992, P64-94.

Bailey, T.C. and Gatrell, A.C. (1995) Interactive Spatial Data Analysis Longman Scientific and Technical, UK.

Borst, R.A. and McCluskey, W.J. (1997), An Evaluation of MRA, Comparable Sales Analysis and ANNs for the Mass Appraisal of Residential Properties in Northern Ireland, *Assessment Journal*, 4(1), pp. 47-55)

Burrough, P.A. & McDonnell, R.A. (1998) Principles of Geographical Information Systems, Oxford University Press. UK.

Can, A. (1990), The Measurement of Neighbourhood Dynamics in Urban House Prices, *Economic Geography*, Vol. 66, pp. 254-272.

Carter, M., Fairburn, J., Fletcher, M., Gallimore, P. and Mangan, J. (1996). A Model to Predict House Prices Involving Geographic Location. *'Cutting Edge' Property Research Conference of The Royal Institution of Chartered Surveyors, 21-22 September 1996*, University of West of England, Brsitol.

Czernowski, R.M.J.(1990) "Expert systems in Real Estate Valuation", Journal of Valuation, Vol.8, P376-393.

Dubin R.A. (1992), Spatial Autocorrelation and Neighbourhood Quality, *Regional Science and Urban Economics*, Vol. 22, pp. 433-452.

Dubin, R.A. and Sung, C. (1987), Spatial Variation in the Price of Housing: Rent Gradients in Non-monocentric Cities, *Urban Studies*, 24, pp. 193-204.

Fleming, M.C. and Nellis, J.G. (1984) "the Halifax House Price Index Technical Details, Halifax Building Society, Halifax.

Gallimore, P., Fletcher, M. and Carter, M. (1996). Modelling the Influence of Location on Value. *Journal of Property Valuation and Investment*, 14(1), pp.6-19.

Galster, G. (1996), William Grigsby and the Analysis of Housing Sub-markets and Filtering, *Urban Studies*, 33(10), pp. 1797-1805.

LaRose, T.A. (1988). Global Response Surface Analysis Used to Update Appraisals in a Computer Assisted Mass Appraisal Environment. *World Congress III of Computer Assisted Valuation and Land Information Systems, 07-12 August 1988*, Cambridge, Massachusetts.

McCluskey, W; Deddis, W; McBurney, R.D; Mannis, A; Borst, R. (1997) "Interactive application of computer assisted mass appraisal and geographic information systems", Journal of Property Valuation and Investment; pp. 448-465, Vol 15, Issue 5

Oliver, M.A., (1990). Kriging: A Method of Interpolation for Geographic Information Systems. *International Journal of Geographic Information Systems,* 4 (4), pp.313-332

Olmo, J.C., (1995). Spatial Estimation of Housing Prices and Locational Rents. *Urban Studies.* 32 (8), pp.1331-1344.

Pitman, R.H; Seldin, M (1990) "Real estate analyses using geographic data", Real Estate Issues, Vol.15, No.1, p32-38.

Rodriguez, M *et al* (1995) "Using GIS to improve real estate analysis", Journal of Real Restate Research, Vol.10, No. 2.

Ver Hoef, J.M. (1993), Universal Kriging for Ecological data, in Goodchild, M.F., Parks, B.O. and Stegaert, L.T. (Eds.), *Environmental Modelling with GIS*, Oxford University Press, Oxford, pp. 447-453.

Thrall G.I.(1985-1988) Editor, Scientific Geography Series, Newbury Park, California; Sage Publications.

Wyatt, P.J. (1997). "The Development of Property Information System for Real Estate Valuation", *International Journal of Geographical Information Systems*, 11(5), pp. 435-450.