Spatial Dependence, Housing Submarkets, and House Prices

Steven C. BOURASSA
University of Louisville, USA
Eva CANTONI
University of Geneva, Department of Econometrics
Martin HOESLI
Universtiy of Geneva, HEC, FAME and University of Aberdeen

Research Paper N° 151
June 2005

FAME - International Center for Financial Asset Management and Engineering

UNIVERSITÉ DE GENÈVE

Spatial Dependence, Housing Submarkets, and House Prices

Steven C. Bourassa[1] Eva Cantoni[2] Martin Hoesli[3]

This draft: June 11, 2005

Abstract

This paper compares the impacts of alternative models of spatial dependence on the accuracy of house price predictions in a mass appraisal context. Explicit modeling of spatial dependence is characterized as a more fluid approach to defining housing submarkets. This approach allows the relevant "submarket" to vary from house to house and for transactions involving other dwellings in each submarket to have varying impacts depending on distance. We compare the predictive ability of different specifications of both geostatistical and lattice models as well as a simpler model based on submarkets with fixed boundaries. We conclude that – for our data – no spatial statistics method does as well in terms of predictive ability as a simple OLS model that includes a series of dummy variables defining submarkets. However, of the spatial statistics methods, geostatistical models provide more accurate predictions than lattice models. We argue that this is due to the fact that the kriging procedure used to make predictions in a geostatistical framework directly incorporates spatial information about nearby properties. That is not possible in a lattice framework due to the reliance on a matrix of weights that incorporates relationships only for the sample of properties that transact.

Key Words: spatial dependence, hedonic price models, geostatistical models, lattice models, mass appraisal, housing submarkets

JEL Codes: C21, R31

---

[1] To whom correspondence should be addressed. *KHC Real Estate Research Professor, School of Urban and Public Affairs, University of Louisville, 426 W. Bloom Street, Louisville, Kentucky 40208, phone: (502) 852 5720, fax: (502) 852 4558,* email: steven.bourassa@louisville.edu

[2] *Department of Econometrics, University of Geneva, 40 boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland*, email: eva.cantoni@metri.unige.ch

[3] *HEC and FAME, University of Geneva, 40 boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland, and University of Aberdeen Business School, Edward Wright Building, Aberdeen AB24 3QY, Scotland*, email: martin.hoesli@hec.unige.ch

Executive Summary

House prices are customarily modeled using hedonic regression models whereby the price is explained by structural and locational attributes. As with all regression models, errors should be independent from one another, else parameter estimates will be inefficient and confidence intervals will be incorrect. The price of a house is likely, however, to be related to the price of adjacent properties. If a hedonic model cannot perfectly capture the effects of location then the residuals of adjacent properties will be correlated. The aim of this paper is to analyze how best to take into account this spatial dependence in a mass appraisal context.

We investigate whether spatial statistical models perform better than an OLS model with neighborhood dummy variables. The comparison is therefore about whether the structure of the errors has to be modeled or whether neighborhood variables can be used. This is also an issue of ease of use as the latter approach is simpler to implement. This is not the first study on this topic, but previous analyses have generally relied on a limited subset of the available spatial techniques or have used a small sample of properties.

We use a rich database of over 4,800 residential sales in Auckland, New Zealand. Two variations each of two OLS, four geostatistical, and two lattice models are considered. Our results suggest that the geostatistical methods perform better than the simple OLS model, but that a simple adjustment of predictions using the average residuals in neighborhoods (submarkets) is almost as good. When submarket dummy variables are added to the OLS model, the predictions are more accurate than the predictions generated with the geostatistical methods. The lattice models perform poorly, in some cases worse than the unadjusted OLS predictions, and we conclude that such models are not suited for ex-sample prediction purposes. Overall, we find that valuer-defined submarkets are more useful in a mass appraisal context than the more fluid concept of submarkets implied by formal modeling of the spatial dependence of residuals.

1. Introduction

House prices are customarily modeled using hedonic regression models whereby the price is explained by structural and locational attributes. As with all regression models, errors should be independent from one another, else parameter estimates will be inefficient and confidence intervals will be incorrect. The independence assumption is unlikely to be valid in a standard ordinary least squares (OLS) context, as house price residuals have been shown to exhibit spatial dependence in spite of efforts to model locational effects accurately (Pace, Barry and Sirmans, 1998). This obviously creates problems as such models are used for house price index construction (Can and Megbolugbe, 1997) and also for mass appraisal (Basu and Thibodeau, 1998; Bourassa, Hoesli and Peng, 2003). Basu and Thibodeau (1998), for instance, argue that spatial dependence exists because nearby properties will often have similar structural features (they were often developed at the same time) and also share locational amenities. More generally, LeSage and Pace (2004) discuss both theoretical and statistical reasons that would explain why data from several fields of study would be prone to spatial dependence.

Such dependence can be treated in two ways. We assume a general model,

$$Y = m(X) + e \ , \tag{1}$$

where Y is a vector of transaction prices, X is a matrix of values for residential property characteristics, and e is an error term. One approach is to model $\mu(X)$ so that residuals over space do not exhibit any pattern. This usually implies including geographical coordinates or other spatial indicators as regressors, parametrically or even nonparametrically (Colwell, 1998; Clapp, 2003; Fik, Ling and Mulligan, 2003). One such approach includes as a regressor the weighted average of recent sales prices for nearby properties (Can and Megbolugbe, 1997). An alternative approach is to model e, that is, to assume not only that $E(e) = 0$, but also that $E(ee') = \Omega$, which is a matrix with at least some nonzero off-diagonal elements. Ripley (1981) and Cressie (1993) provide a discussion of relevant spatial statistics methods for modeling e. These include geostatistical models such as those applied in real estate by Dubin (1998) or Basu and

Thibodeau (1998) and the lattice models that have been refined and applied by Pace and his colleagues (e.g., Pace and Barry, 1997a).

Theoretically speaking, the assumptions behind the two classes of spatial statistics differ in terms of the definition of the domain over which spatial locations are permitted to vary (see Cressie, 1993, pp. 8-9, for details). In the case of lattice models, which include simultaneous autoregressive (SAR) and conditional autoregressive (CAR) variants, locations are restricted to a discrete set of points. In contrast, geostatistical models permit an infinite number of locations within a given geographical area. This has implications for the way predictions based on each type of model take into account spatial information. Given their constraints, the lattice models seem less suited than the geostatistical models for prediction purposes. Whether this is of practical relevance is an empirical question that we will address here.

Our focus is to compare the utility of spatial statistical methods relative to each other and to simpler OLS methods in a mass appraisal context. For testing purposes, we use a large sample of 4,880 residential sales in Auckland, New Zealand. For each method, 100 random samples each containing 80% of the transactions are generated to estimate the predictive ability (and variability) of each technique for the 20% ex-samples. We estimate four geostatistical models, one each based on exponential and spherical variograms and then robust versions of the same models. We then estimate two lattice models, SAR and CAR. We compare predictions from these models with each other and with the predictions from two OLS models. The predictions from one OLS model are adjusted by the unweighted average residuals for valuer-defined submarkets, while the predictions from the other OLS model are not. Finally, we add a set of submarket dummy variables to each of the models (geostatistical, lattice, and OLS) in order to assess the impacts of simple controls for neighborhood effects on the accuracy of predictions.[4] We are particularly interested in comparing the predictions based on the OLS model that

---

[4] An alternative approach would be to incorporate variables measuring neighborhood characteristics (as in Dubin, 1988, for example). Such data would typically be available for small areas defined for census purposes. However, census areas are less likely to correspond to housing submarkets than are the valuer-defined areas used here.

incorporates submarket dummies with those based on the spatial statistical models without submarket dummies.

Previous research has either focused on a limited subset of the available spatial techniques or used a small sample of properties. Using a small sample from Baltimore, Dubin (1988) compares ex-sample predictions using OLS and a geostatistical technique and concludes that the geostatistical approach is superior even when some neighborhood (census block group) characteristics are included as explanatory variables. Basu and Thibodeau (1998), for instance, compare the predictive ability of OLS and one geostatistical technique, concluding that the latter is superior for six of eight regions in Dallas. Dubin, Pace and Thibodeau (1999) compare regression coefficients across OLS and four different spatial methods (including both geostatistical and lattice models) using a small simulated example for which the true parameters are known. Most of the spatial models performed better than OLS with respect to parameter estimation. Militino, Ugarte and García-Reinaldos (2004) apply several models – including CAR, SAR, and geostatistical – to 293 transactions from Pamplona, Spain, but do not attempt ex-sample predictions.

A recent study by Case et al. (2004) is in some ways similar to the present one. They apply OLS and several spatial statistical methods to a very large sample of about 50,000 transactions from Fairfax County, Virginia, using out-of-sample prediction accuracy for comparison purposes. In their final results, two of the three spatial methods produced more accurate ex-sample predictions than an OLS model that included median residuals for a small number of nearest neighbors. Although these authors estimated an OLS model with neighborhood (census tract) dummy variables, they did not then use that model for prediction purposes. Also, unlike the present paper, they performed their predictions using only one split of the data. This means that their results may depend on the particular split.

The paper is structured as follows. We first discuss the relationship between the ideas of spatial dependence and housing submarkets. These ideas are very closely related; thinking in terms of housing submarkets is helpful in conceptualizing the problem that spatial dependence models seek to rectify. Section 3 contains a presentation of the spatial statistical methods that are used in the paper, while section 4 outlines our research design.

We discuss our empirical analysis in the following section. Section 6 concludes the paper.

2. Spatial Dependence and Housing Submarkets

The concepts of spatial dependence and housing submarkets are closely related. The submarket concept relies on the idea of substitutability. Substitutes are pairs of goods for which an increase in the price of one leads to an increase in the demand for the other. Pairs of goods with similar characteristics are likely to be substitutes. In equilibrium, prices equalize across substitutes. Within housing submarkets, prices of houses are similar because submarkets contain close substitutes. Implicit prices of the characteristics of houses are similar for the same reason.

Spatial dependence or autocorrelation refers to the existence of covariance in the errors in the context of hedonic price estimation for residential property markets. Given the similarities in the prices of housing characteristics within a submarket, errors are more likely to be correlated within submarkets than across submarkets. Therefore, controlling for submarkets in hedonic equations can substantially reduce estimation errors. This can be accomplished in a variety of ways. Simple methods include incorporating a series of dummy variables for the submarkets, estimating a separate equation for each submarket, or adjusting predicted values using the errors within each submarket.[5]

Controlling for submarkets in hedonic price equations assumes either that one has a predefined set of submarkets or that one is going to use some method to define them. Predefined submarkets are typically geographical areas, such as those defined by real estate agents (e.g., Palm, 1978) or by valuers (e.g., Bourassa, Hoesli and Peng, 2003). Alternatively, submarkets can be defined in terms of the characteristics of dwellings, neighborhoods, or census units. Statistical techniques, such as principal components and

[5] Bourassa, Hoesli and Peng (2003) show that the latter method, although quite simple, results in significant improvements in the accuracy of predictions based on a market-wide hedonic equation, thus we test for its impact here.

cluster analysis, can be used to combine similar dwellings or neighborhoods into submarkets, which may or may not be geographical areas (e.g., Bourassa et al., 1999). Ugarte, Goicoa and Militino (2004) demonstrate the use of mixture models which both estimate hedonic equations and classify transactions into submarkets which are not geographical areas. However, there is some evidence to suggest that geographical submarkets are more meaningful and therefore useful for improving prediction accuracy (Bourassa, Hoesli and Peng, 2003).

Spatial statistical methods allow for a more fluid concept of submarkets than is permitted by the fixed definitions based on geographical areas or housing or neighborhood characteristics. In effect, methods such as the lattice or geostatistical approaches applied here allow for the relevant submarket to vary from property to property. The relationships between the focal property in a submarket and nearby properties are captured in a matrix of weights in the case of lattice models or by a distance function based on a fitted variogram (or semivariogram) in the case of geostatistical models. This more fluid approach to modeling the relationships between properties would seem *a priori* to allow for more effective reduction of prediction errors due to spatial dependence.

Whether these spatial statistical approaches are more effective than simpler methods for improving mass appraisal accuracy is the empirical question that we address here. It is useful in this context to consider Can's (1992) distinction between *adjacency* and *neighborhood* effects. The lattice and geostatistical methods focus on adjacency effects, or the external effects of nearby properties on the property in question. The simpler methods mentioned above, such as controlling for location within a relatively homogeneous geographical area defined by valuers for appraisal purposes, imply a focus on neighborhood effects. Thus our empirical question is whether adjacency or neighborhood effects predominate. In other words, is it more important to account for each property's situation within the boundaries of relatively homogeneous neighborhoods that are recognized as such in a particular market or to account for the relationships between each property and its neighbors?

3. Alternative Methods for Modeling Spatial Dependence

In this section, we present two modeling approaches for spatial data that we use in this paper: lattice and geostatistical models. We refer the reader to Ripley (1981) and Cressie (1993) for a more complete and detailed description of the statistical aspects of these models. In a nutshell, the lattice approach models the covariance matrix of the errors parametrically, whereas the geostatistical approach builds upon a direct (nonparametric) estimation of the covariance matrix of the underlying process. Moreover, the underlying assumptions of the two approaches differ (see the discussion in Section 1).

3.1. Lattice Models

In this subsection, we assume that the data are issued from equation (1), with $E(e) = 0$ and $E(ee') = \Omega$. Lattice models assume $m(X) = Xb$ and parameterize the covariance function of the error term of the model by assuming either that $\Omega^{-1} = s^2(I - fC)$ (CAR models) or that $\Omega^{-1} = s^2(I - aD)'(I - aD)$ (SAR models), where C and D represent spatial weight matrices that specify the dependence among observations. These matrices satisfy the conditions that their rows sum to 1, that their diagonal is 0 (an observation does not impact its own prediction) and that $0 \leq a, f < 1$. Many choices for specifying the weight matrices are available in the literature (see Getis and Aldstadt, 2004, for a review), but some of them show only small practical differences, such as the results in Militino, Ugarte and García-Reinaldos (2004).

The estimates of the parameters a (or f) and b are then obtained by maximizing the log-likelihood

$$\ln L = const. + \frac{1}{2}\ln\left|\Omega^{-1}\right| - \frac{1}{2}\left((Y - Xb)'\Omega^{-1}(Y - Xb)\right). \tag{2}$$

The most important computational issue here is the evaluation of the log-determinant ($\ln\left|\Omega^{-1}\right|$), which is infeasible by standard methods for large sample sizes, given that Ω is of size n by n. Pace and Barry (1997a, 1997b) have derived approximations to these terms that are implemented in their Matlab code.[6] In this paper, we use their code to fit CAR and SAR models, with the Delauney spatial weight matrix (Cressie, 1993, p. 374).

Predictions are computed simply as $\hat{Y} = X\hat{b}$. Ripley (1981, p. 90) gives a formula to compute fitted values (that is predictions for in-sample observations) for SAR models (see also Pace and Gilley, 1997, 1998). This formula borrows strength from the information provided by neighboring observations through the spatial weight matrix D. We are interested in ex-sample predictions, therefore ruling out the use of this formula.

3.2. Geostatistical Models

The modeling approach developed in this section is based on the assumption that the observed data at a location s is a realization of a random process $\{Y(s) : s \in F\}$, which is supposed to satisfy a second-order stationarity assumption, that is, for which $E(Y(s)) = m$ for all $s \in F$ (constant mean) and $Cov(Y(s_1), Y(s_2)) = C(s_1 - s_2)$ for all $s_1, s_2 \in F$. In effect, the covariance between locations depends only on the distance between them. $C(\cdot)$ is called the covariogram.

The geostatistical approach attempts to model the covariance matrix directly through a procedure based on three steps: (1) the computation of an empirical variogram, (2) the parametric modeling of this variogram, and (3) kriging (that is, prediction). The only information needed to perform these three steps is the notion of variogram defined as a function of the distance h between locations:

$$2g(h) = Var(Y(s+h) - Y(s)), \tag{3}$$

where $g(h)$ is called the semivariogram.

The classical and most popular estimator of the variogram is obtained by the method of moments and was first proposed by Matheron (1962):

$$2\hat{g}(h) = \frac{1}{|N(h)|} \sum_{N(h)} \left(Y(s_i) - Y(s_j)\right)^2, \tag{4}$$

[6] This code is available at http://www.spatial-statistics.com.

where $N(h) = \{(i, j) : s_i - s_j = h\}$ and $|N(h)|$ is the number of distinct elements of $N(h)$. For a given distance h, this variogram estimator is a variance estimator over all pairs of observations that are at a distance h apart. Note that when data are irregularly spaced, the variogram is usually smoothed by summing over pairs of points that lie in a tolerance region. $\hat{g}(h)$ is an unbiased estimator of $g(h)$, but is badly affected in presence of outliers because of the $(\cdot)^2$ term in the sum. Therefore, Cressie and Hawkins (1980) have defined a more robust estimator:

$$2\tilde{g}(h) = \left\{ \frac{1}{|N(h)|} \sum_{N(h)} |Y(s_i) - Y(s_j)|^{1/2} \right\}^4 \Bigg/ \left(0.457 + \frac{0.494}{|N(h)|} \right), \tag{5}$$

which achieves robustness through $|Y(s_i) - Y(s_j)|$. In the presence of outlying observations this estimator is more stable.

The second step of the procedure consists of fitting a parametric model to the empirical variogram (either classical or robust). The most popular variogram models include the exponential variogram defined by

$$g(h;J) = \begin{cases} 0 & \text{if } h = 0 \\ c_0 + c_e\left(1 - \exp(-\|h\|/a_e)\right) & \text{if } h \neq 0 \end{cases}, \tag{6}$$

where $J = (c_0, c_e, a_e)'$ with $c_0 \geq 0$, $c_e \geq 0$ and $a_e \geq 0$, and the spherical variogram defined by

$$g(h;J) = \begin{cases} 0 & \text{if } h = 0 \\ c_0 + c_s\left\{(3/2)(\|h\|/a_s) - (1/2)(\|h\|/a_s)^3\right\} & \text{if } 0 < \|h\| \leq a_s \\ c_0 + c_s & \text{if } \|h\| > a_s \end{cases}, \tag{7}$$

where $J = (c_0, c_s, a_s)'$, with $c_0 \geq 0$, $c_s \geq 0$ and $a_s \geq 0$. The parameter c_0 is the limit of $g(h)$ when $h \to 0$ and is called the "nugget effect". The other parameters in J control the functional form of $g(h;J)$ (see Cressie, 1993, pp. 61-63, for details). The parametric variograms can be fitted to data by several procedures, which include – among others – (restricted) maximum likelihood and generalized least squares.

Given a fitted variogram, the procedure goes on to compute the prediction at a point s_0 as a linear combination of the responses, that is,

$$\hat{Y}(s_0) = l'Y = \sum_{i=1}^{n} l_i Y(s_i), \tag{8}$$

where $l = (l_1, \ldots, l_n)'$ is obtained by minimizing the mean squared error

$$E(Y(s_0) - \sum_{i=1}^{n} l_i Y(s_i))^2 . \tag{9}$$

The solution for l depends on $g(s_0 - s_i)$ for all $i = 1, \ldots, n$, and on $g(s_i - s_j)$ for all $1 \leq i,\ j \leq n$. $\hat{Y}(s_0)$ is the best linear unbiased predictor. The solution obtained is an exact interpolation at the sample points, that is, $\hat{Y}(s_i) = Y(s_i)$ for all $i = 1, \ldots, n$. Note also that the formula above allows the computation of predictions at both sampled and unsampled locations.

When the process is not stationary, a preliminary step can be performed to capture the trend. For instance, one can first fit a regression model and then compute the variogram on the residuals of this regression. The final predictions are then computed by adding the kriging predictions on residuals to the fitted values of the regression. This can introduce some bias, and a GLS iterative procedure could be constructed based on the covariance matrix resulting from a variogram model (see Basu and Thibodeau, 1998). The bias is of order $1/n$, making this issue of little concern in our case (and in residential real estate analysis in general) given that sample sizes are often quite large.

To implement the geostatistics approach, we used the S+Spatial Toolbox of the commercial software Splus. Several other free or commercial software packages are available.

4. Research Design

The main objective of the paper is to contrast the out-of-sample accuracy in house price predictions of several alternative specifications. We consider 8 different techniques, each with and without submarket dummy variables, for a total of 16 techniques. These include two OLS approaches, four types of geostatistical models and two types of lattice models. We perform hedonic regressions for each of the 16 techniques using 100 randomly selected samples of our data each containing 80% of the observations. Hence the methods are compared based on the same 100 samples of data.

For each model and for each of the 100 splits, out-of-sample predictions are generated for the remaining 20% of data. The OLS predictions are performed with and without taking spatial dependence into account. The method of accounting for spatial dependence in this case is to adjust the predictions by the unweighted average residuals for each of the submarkets. We thus have 16 sets of predictions and we calculate the proportion of predictions that are within 10% and 20% of the sale prices. This is done for each of the 100 samples and the median and distributions of the proportions are calculated for each model. These form the basis for our comparisons.

The source of data for this study is the official database of all real estate transactions in New Zealand. We use data pertaining to detached and semi-detached dwellings. We focus on sales in the City of Auckland in 1996. A total of 4,880 transactions were retained for the analysis.[7] The database contains the date of sale, the sale price, and such information as: exact location, floor area, age, wall material and condition, and quality of the principal structure. The land area is provided for 65% of the units. The units for which no land area is provided are generally "cross-leased", which means that the land is owned collectively by the owners of the dwellings on that site. The collective owners lease a fraction of the land to each individual owner for a "peppercorn", or nominal, rent. For all such cross-leased dwellings, we set land area equal to zero and set a dummy variable equal to one. Supplementary information used for mass appraisal purposes is also available. These data include important characteristics such as water views, and the quality of landscaping and of the neighborhood. For the dependent variable in our hedonic models we use the sale price net of the value of any chattels.

---

[7] A sale was removed from the sample if it fell into one of the following categories: (a) the property had a land area larger than 0.25 hectares (this excluded properties that may have been sold primarily for redevelopment purposes); (b) the property had a floor area either less than 30 square meters (probably due to an error in data entry); (c) the transaction was flagged as not being "arm's length"; or (d) the property was located on an island.

The use of GIS also allowed the data set to be supplemented with the measure of the distance between each property and the central business district (CBD). In addition, we use geographical areas (sales groups) defined by official appraisers in New Zealand as our spatial submarkets. Finally, quarterly time dummy variables are included in each model.

Some variables were transformed before entering into the estimations. The dependent variable, house price, both land and floor area, and distance to the CBD are as natural logarithms. [8] In addition, both age and age squared are included in the model as the relation between house value and age is expected to follow a U-shaped curve. Table 1 contains the means of the raw independent and dependent variables used in the analyses.

The submarket dummy variables identify location within one of 34 geographical areas within the city. These submarkets, referred to as "sales groups", were defined by valuers for mass appraisal purposes and are considered to be relatively homogeneous submarkets. For estimation purposes, we combined three sales groups located in or near the CBD because they had relatively few transactions; these form the default category.

5. Empirical Analysis

Table 2 contains examples of hedonic regression results using OLS with and without submarket dummy variables for a random sample of 80% of the data. The adjusted R^2 statistic increases from 0.722 to 0.798 when the 33 submarket dummy variables are added to the model. The results are consistent with expectations. The logarithms of land and floor area are positively related to sale price, as is the square of the age of the property. Age itself is negatively related to the sale price. The quality and condition of

---

[8] The OLS predictions are calculated as $\exp(\widehat{\ln Y})$, although the correct transformation would be $\exp(\widehat{\ln Y} + 0.5\hat{s}^2)$. Because we are unable to implement equivalent transformations for predictions based on the other methods, we do not add $0.5\hat{s}^2$ before taking the antilogs of the OLS predictions. Given the large sample size, this has only a trivial impact on the results.

the properties are also important. The logarithm of distance to the CBD is negatively related to sale price and is highly significant. The sale price is approximately 10% higher for properties with a water view. Good landscaping, the number of attached garages, and to a lesser extent, a driveway, significantly affect dwelling prices. The quality of the neighborhood is very important, and higher quality levels are associated with higher prices. In the model with submarket dummy variables, 25 out of 33 are significant at the 95% confidence level. When such variables are included in the model, there is a decline in the percentage price impact of being in better neighborhoods. This would be expected as submarket variables will capture part of the variation in neighborhood quality.

We perform two exploratory analyses to determine whether spatial dependence exists in our data. First, we depict the error structure of the OLS regression that does not include submarket variables by constructing a semivariogram (Figure 1). We also investigate the median of the residuals for various x and y coordinates, respectively (Figure 2). For this purpose, we divide the city into a grid of 19 cells from west to east by 13 cells from south to north (the west-east dimension is greater than the south-north dimension in Auckland). The semivariogram shows that there is covariance in the error structure and that this declines with distance (the semivariogram increases with distance). Figure 2 shows that the median of residuals is not constant across geographical areas. In particular, residuals tend to be negative to the west and south and positive to the east and north. Hence, spatial dependence clearly exists in the error structure of the OLS model.

The next step is to analyze how best to account for such dependence to obtain more accurate house price predictions. Figure 3a shows boxplots of the proportion of predictions within 10% of sales price, while Figure 3b shows the proportion of predictions within 20%. Results for six methods are shown: OLS with and without adjustment by the average residuals in submarkets, the exponential and robust exponential variogram models and the CAR and SAR models. Predictions with the spherical models are not reported as those predictions are very similar to predictions with the exponential variogram model. Both parts of Figure 3 depict results with and without submarket dummy variables. The body of each boxplot is constructed from the first to the third quartiles, while the whiskers are set at ±1.5 times the inter-quartile range from the median. However, if no observation exists at that distance, the whisker is set at the closest observation towards the body of the boxplot. If there are observations outside of the whiskers, each of these is depicted by a line. Within the body of each boxplot, the median over the 100 splits appears as a bar, while the 95% confidence interval is depicted by the indented and unshaded area in the center of the plot.

When the OLS model without submarket dummy variables is considered, the median of predictions within 10% of sales price across the 100 splits is 39.8%.[9] The geostatistical models yield significantly more accurate predictions (median of approximately 44%). However, a simple adjustment to the OLS predictions using the unweighted average residuals in submarkets yields predictions that are only marginally less accurate than the predictions generated using the geostatistical models. The CAR and SAR methods produce predictions that are worse than the unadjusted OLS predictions. It is possible that the error structure contained in O is not reflected in the data. This suggests that the lattice models, although useful in improving the efficiency of parameter estimates, are not well suited for mass appraisal purposes. Similar results are obtained for predictions within 20% of actual sales price. Some 73.8% of adjusted OLS predictions are within 20%, which is only marginally less than with the geostatistical models (approximately 75.5%). Again, CAR and SAR models yield the least accurate predictions.

When submarket variables are added to the OLS model, the median of predictions within 10% rises to 46.8% and that of predictions within 20% to 77.9%.[10] The geographical subdivisions that are used by appraisers work significantly better in improving the accuracy of house price predictions than do the spatial statistical models. Geographical submarkets are more important in predicting house prices than the more fluid approach which permits “submarkets” to vary from house to house. Hence, the valuers’ geographical definition of submarkets is useful for mass appraisal purposes.

---

[9] For comparison purposes, Fik, Ling and Mulligan (2003) note that Freddie Mac prefers to have at least 50% of predictions within 10% of the actual values.

[10] By definition, the adjustment using average residuals in submarkets is ineffective when submarket dummy variables are included in the model.

This conclusion is of practical importance, as a hedonic model with submarket dummy variables is substantially easier to implement than spatial statistical methods.

This conclusion could be a consequence of the extensive set of variables in this data set and/or the fact that the variables are measured relatively accurately. To test the sensitivity of our results to the number of attributes considered, we re-estimated the models without several property characteristics.[11] The same conclusion holds true when these variables are deleted. The second conjecture could only be tested on another data set where variables would be measured with less accuracy. The good results obtained for OLS predictions with submarket dummy variables could also be due to these geographical areas being very well defined and hence capturing much of the spatial dependence in house prices in Auckland. Further work could focus on data sets for cities in which such requirements are not met, or the Auckland data could be used in combination with arbitrarily defined submarkets.

When submarket dummy variables are included in all estimations, the gestatistical models yield slightly higher percentages of predictions within the 10% and 20% limits than the OLS model (80.1% versus 77.9%). It is likely that the geostatistical estimations that incorporate submarket variables are superior to the corresponding OLS model because, to use Can's (1992) terminology, the former capture adjacency effects as well as neighborhood effects. The CAR and SAR models, however, offer little or no increase in prediction accuracy above the OLS model.

6. Conclusions

The price of a house is likely to be related to the price of adjacent properties. If a hedonic model cannot perfectly capture the effects of location then the residuals of adjacent properties will be correlated. The aim of this paper is to analyze how best to take into account this spatial dependence in a mass appraisal context. We investigate

---

[11] In the list of variables in Table 2, we deleted variables from "Walls in good condition" to "Average quality of the principal structure" and from "Water view" to "Very good neighborhood."

whether spatial statistical models perform better than an OLS model with neighborhood dummy variables. The comparison is therefore about whether the structure of the errors has to be modeled or whether neighborhood variables can be used. This is also an issue of ease of use as the latter approach is simpler to implement. This is not the first study on this topic, but previous analyses have generally relied on a limited subset of the available spatial techniques or have used a small sample of properties. The recent paper by Case et al. (2004) is in some ways similar to this paper, but no test of the ex-sample effectiveness of the OLS model with submarket dummy variables is conducted.

We use a rich database of over 4,800 residential sales in Auckland, New Zealand. Two variations each of two OLS, four geostatistical, and two lattice models are considered. Our results suggest that the geostatistical methods perform better than the simple OLS model, but that a simple adjustment of predictions using the average residuals in neighborhoods (submarkets) is almost as good. When submarket dummy variables are added to the OLS model, the predictions are more accurate than the predictions generated with the geostatistical methods. The lattice models perform poorly, in some cases worse than the unadjusted OLS predictions, and we conclude that such models are not suited for ex-sample prediction purposes.

Overall, we find that valuer-defined submarkets are more useful in a mass appraisal context than the more fluid concept of submarkets implied by formal modeling of the spatial dependence of residuals. This appears to differ from Case et al.'s (2004) conclusions, although our methods are not the same as theirs and, as noted above, they do not make predictions from their OLS estimation with census tract dummies.

This work could be expanded in a number of ways. First, we could compare our results with those obtained using various methods that focus on measuring the impact of location more effectively in $\mu(X)$. Our best model yields a proportion of predictions within 10% of house values that is just shy of 50%. Fik, Ling and Mulligan (2003) show that when x and y coordinates are interacted with a limited set of independent variables in a hedonic equation of residential units in Tucson, the proportion of predictions within 10% of sales price is 65%. The use of such a model would imply reducing the number of hedonic attributes considered, however. Second, the models that we consider could be applied in the context of a city where property attributes are measured less accurately than is the case in Auckland. Finally, these methods could be analyzed using submarkets that are defined less homogenously. It could be that spatial statistical methods yield better forecasts in such an environment.

References

Basu, A. and T. G. Thibodeau (1998). "Analysis of Spatial Autocorrelation in House Prices," *Journal of Real Estate Finance and Economics*, 17(1), 61-85.

Bourassa, S. C., F. Hamelink, M. Hoesli and B. D. MacGregor (1999). "Defining Housing Submarkets," *Journal of Housing Economics*, 8(2), 160-183.

Bourassa, S. C., M. Hoesli and V. C. Peng (2003). "Do Housing Submarkets really Matter?" *Journal of Housing Economics*, 12(1), 12-28.

Can, A. (1992). "Specification and Estimation of Hedonic Housing Price Models," *Regional Science and Urban Economics*, 22(3), 453-474.

Can, A. and I. Megbolugbe (1997). "Spatial Dependence and House Price Index Construction," *Journal of Real Estate Finance and Economics*, 14(1/2), 203-222.

Case, B., J. Clapp, R. Dubin and M. Rodriguez (2004). "Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models," *Journal of Real Estate Finance and Economics*, 29(2), 167-191.

Clapp, J. M. (2003). "A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation," *Journal of Real Estate Finance and Economics*, 27(3), 303-320.

Colwell, P. F. (1998). "A Primer on Piecewise Parabolic Multiple Regression Analysis via Estimations of Chicago CBD Land Prices", *Journal of Real Estate Finance and Economics*, 17(1), 87-97.

Cressie, N. (1993). *Statistics for Spatial Data*, Wiley: New York.

Cressie, N. and D. M. Hawkins (1980). "Robust Estimation of the Variogram, I," *Journal of the International Association for Mathematical Geology*, 12(2), 115-125.

Dubin, R. A. (1988). "Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms," *Review of Economics and Statistics*, 70(3), 466-474.

Dubin, R. A. (1998). "Predicting House Prices Using Multiple Listings Data," *Journal of Real Estate Finance and Economics*, 17(1), 35-59.

Dubin, R., R. K. Pace and T. G. Thibodeau (1999). "Spatial Autoregression Techniques for Real Estate Data," *Journal of Real Estate Literature*, 7(1), 79-95.

Fik, T. J., D. C. Ling and G. F. Mulligan (2003). "Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach," *Real Estate Economics*, 31(4), 623-646.

Getis, A. and J. Aldstadt (2004). "Constructing the Spatial Weights Matrix Using a Local Statistic," *Geographical Analysis*, 36(2), 90-104.

LeSage, J. P. and R. K. Pace (2004). "Introduction," in *Spatial and Spatiotemporal Econometrics*, LeSage, J. P. and R. K. Pace (Eds), Advances in Econometrics, Volume 18, Oxford: Elsevier, 1-32.

Matheron, G. (1962). *Traité de Géostatistique Appliquée, Tome I*, Mémoires du Bureau de Recherches Géologiques et Minières, No. 14, Paris: Editions Technip.

Militino, A. F., M. D. Ugarte and L. García-Reinaldos (2004). "Alternative Models for Describing Spatial Dependence among Dwelling Selling Prices," *Journal of Real Estate Finance and Economics*, 29(2), 193-209.

Pace, R. K. and R. Barry (1997a). "Quick Computation of Regressions with a Spatially Autoregressive Dependent Variable," *Geographical Analysis*, 29(3), 232-247.

Pace, R. K. and R. Barry (1997b). "Sparse Spatial Autoregressions," *Statistics and Probability Letters*, 33(2), 291-297.

Pace, R. K., R. Barry, J. M. Clapp and M. Rodriguez (1998). "Spatiotemporal Autoregressive Models of Neighborhood Effects," *Journal of Real Estate Finance and Economics*, 17(1), 15-33.

Pace, R. K., R. Barry and C. F. Sirmans (1998). "Spatial Statistics and Real Estate," *Journal of Real Estate Finance and Economics*, 17(1), 5-13.

Pace, R. K. and O. W. Gilley (1997). "Using the Spatial Configuration of the Data to Improve Estimation," *Journal of Real Estate Finance and Economics*, 14(3), 333-340.

Pace, R. K. and O. W. Gilley (1998). "Generalizing the OLS and Grid Estimators," *Real Estate Economics*, 26(2), 331-347.

Palm, R. (1978). "Spatial Segmentation of the Urban Housing Market," *Economic Geography*, 54(3), 210-221.

Ripley, B. (1981). *Spatial Statistics*, New York: Wiley.

Ugarte, M. D., T. Goicoa and A. F. Militino (2004). "Searching for Housing Submarkets using Mixtures of Linear Models," in *Spatial and Spatiotemporal Econometrics*, LeSage, J. P. and R. K. Pace (Eds), Advances in Econometrics, Volume 18, Oxford: Elsevier, 259-276.

Table 1. Sample statistics

Variable	Mean
Net sale price (NZ$)	328,398
Age of dwelling	46
Land area (square meters)	55
Cross-leased or strata-titled	
Floor area (square meters)	144
Detached houses (proportion)	1.00
Wall condition (proportion)	
Good	0.39
Average	0.58
Bad	0.03
Roof material (proportion)	
Tile	0.41
Metal	0.55
Other	0.04
Wall material (proportion)	
Wood	0.63
Brick	0.13
Fibrolite	0.06
Other	0.18
Quality of the principal structure (proportion)	
Superior	0.19
Average	0.76
Poor	0.05
Distance to CBD (km)	6.8
Water view (proportion)	0.09
Modernization (proportion)	0.26
Landscaping (proportion)	
Good	0.16
Average	0.79
Poor	0.05
Driveway	0.85
Quality of the neighborhood (proportion)	
Very good	0.03
Good	0.20
Average	0.68
Poor	0.09
Number of attached garages	0.75
Sample size	*4,880*

Table 2. Examples of results for OLS estimations without and with submarket dummy variables

Variables	Estimates for equation without submarket dummies	Estimates for equation with submarket dummies (estimates for submarket dummies are preceded by the sales group number)	
Intercept	12.816***	11.344***	(4) -0.206***
Log of floor area	0.690***	0.457***	(5) -0.053
Log of land area	1.898***	2.560***	(6) -0.009
Cross-leased or strata-titled	0.099***	0.128***	(7) 0.226***
Age of dwelling	-0.003***	-0.004***	(8) 0.031
Age of dwelling squared	4.720×10^{-5}***	5.096×10^{-5}***	(9) 0.176***
Walls in good condition	0.088***	0.083***	(10) 0.165***
Walls in average condition	0.063***	0.052***	(12) -0.062
Dwelling with a tile roof	-0.030	-0.028	(13) 0.103***
Dwelling with a metal roof	-0.068***	-0.041**	(14) 0.214***
Dwelling with wooden walls	-0.015	-0.006	(15) -0.059
Dwelling with brick walls	-0.056***	-0.019	(16) -0.379***
Dwelling with fibrolite walls	-0.102***	-0.049***	(17) -0.237***
Superior quality of the principal structure	0.231***	0.124***	(18) -0.410***
Average quality of the principal structure	0.094***	0.050***	(19) -0.322***
Log of distance to the CBD	-0.172***	-0.137***	(22) -0.647***
Quarter 2	0.008	0.014	(23) -0.117***
Quarter 3	-0.020**	-0.019**	(24) -0.006
Quarter 4	0.015	0.017	(25) 0.177***
Water view	0.103***	0.079***	(26) 0.181***
Modernization	0.034***	0.029***	(27) -0.096**
Average landscaping	0.026	0.013	(28) -0.315***
Good landscaping	0.077***	0.060***	(29) -0.220***
Driveway	0.019	0.010	(30) -0.141***
Average neighborhood	0.098***	0.021	(31) -0.119***
Good neighborhood	0.231***	0.067***	(32) -0.130***
Very good neighborhood	0.323***	0.205***	(33) -0.189***
Number of attached garages	0.039***	0.036***	(34) -0.010
			(35) -0.074*
			(37) -0.233***
			(38) -0.268***
			(39) -0.225***
			(53) -0.274***
Adjusted R^2	*0.722*	*0.798*	

Note: The symbols *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

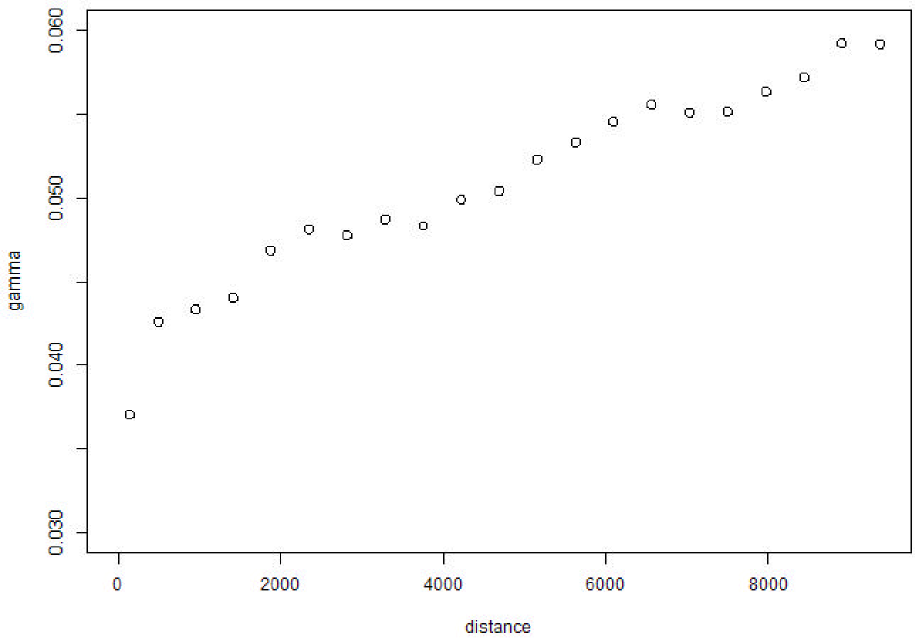

Figure 1. Empirical semivariogram

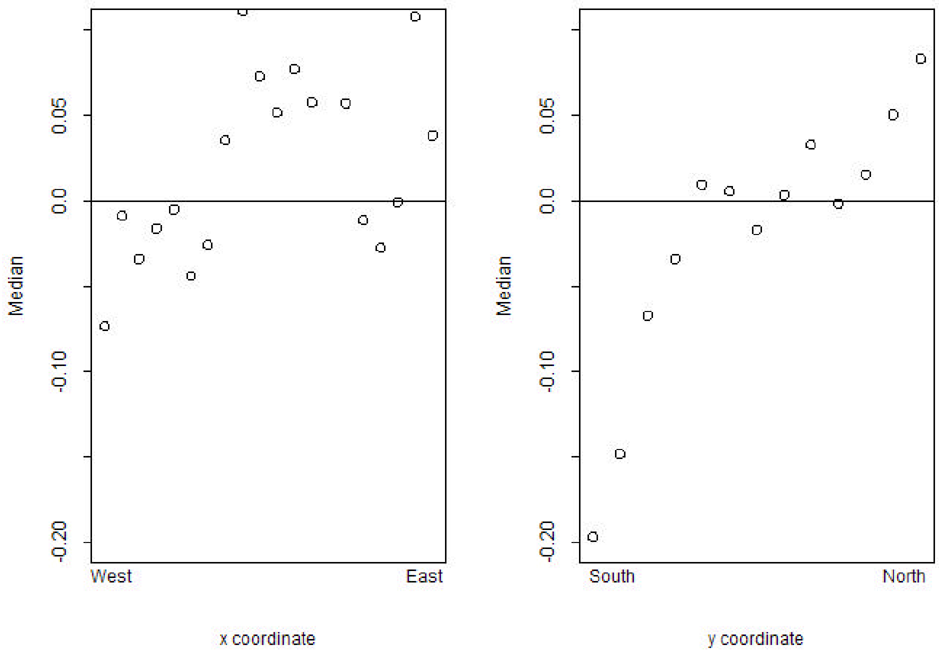

Figure 2. Median values of residuals

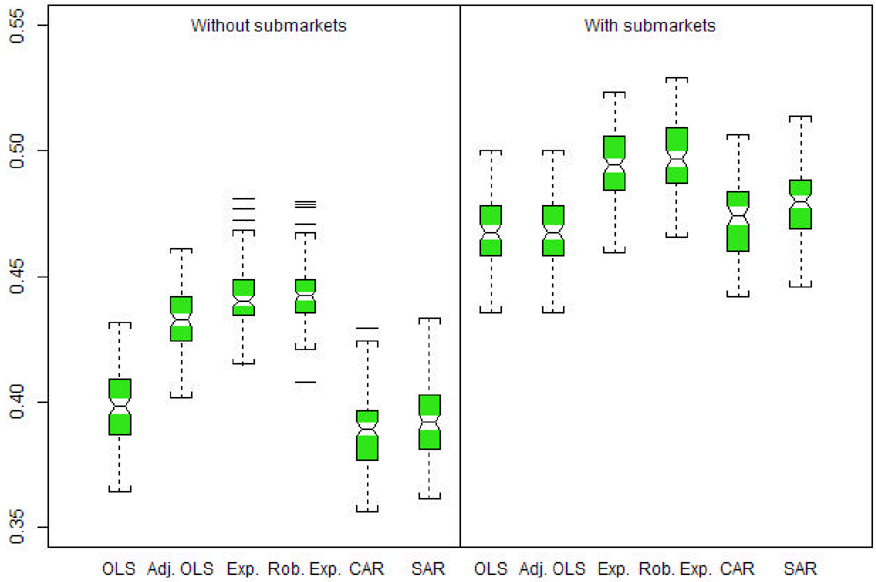

Figure 3a. Proportion of predictions within 10% of actual value

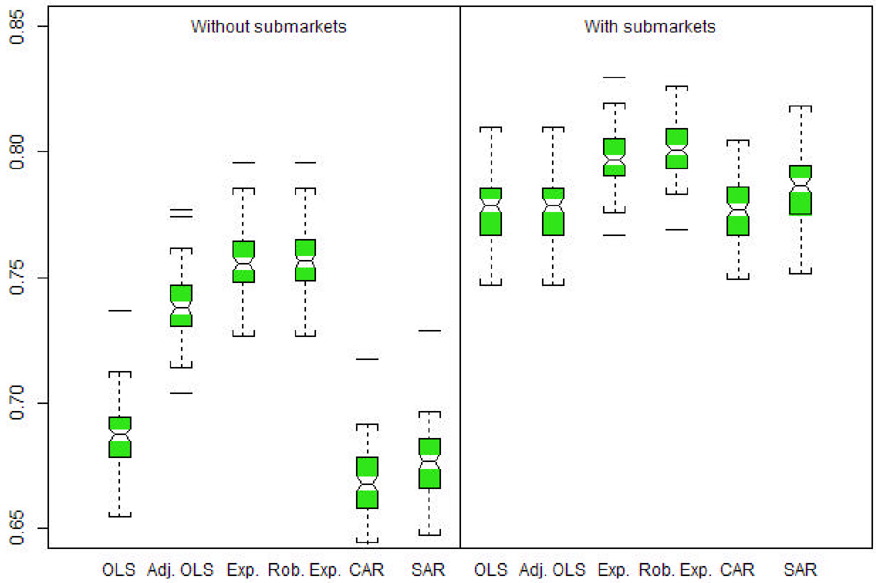

Figure 3b. Proportion of predictions within 20% of actual value

The FAME Research Paper Series

*The International Center for Financial Asset Management and Engineering (FAME) is a private foundation created in 1996 on the initiative of 21 leading partners of the finance and technology community, together with three Universities of the Lake Geneva Region (Switzerland). FAME is about **Research, Doctoral Training**, and **Executive Education** with "interfacing" activities such as the FAME lectures, the Research Day/Annual Meeting, and the Research Paper Series.*

*The **FAME Research Paper Series** includes three types of contributions: First, it reports on the research carried out at FAME by students and research fellows; second, it includes research work contributed by Swiss academics and practitioners interested in a wider dissemination of their ideas, in practitioners' circles in particular; finally, prominent international contributions of particular interest to our constituency are included on a regular basis. Papers with strong practical implications are preceded by an Executive Summary, explaining in non-technical terms the question asked, discussing its relevance and outlining the answer provided.*

Martin Hoesli is acting Head of the Research Paper Series. Please email any comments or queries to the following address: Martin.Hoesli@hec.unige.ch.

The following is a list of the 10 most recent FAME Research Papers. For a complete list, please visit our website at www.fame.ch under the heading 'Faculty and Research, Research Paper Series, Complete List'.

N°150 **Estimation of Jump-Diffusion Precesses via Emirical Characteristic Function**
Michael Rockinger, HEC, University of Lausanne and FAME; Maria Semenova, HEC, University of Lausanne and FAME

N°149 **Suggested vs Actual Institutional Allocation to Real Estate in Europe: A Matter of Size?**
Martin Hoesli, HEC, University of Geneva and FAME and University of Aberdeen; Jon Lekander, Aberdeen Property Investors, Stockholm

N°148 **Monte Carlo Simulations for Real Estate Valuation**
Martin Hoesli, HEC, University of Geneva and FAME and University of Aberdeen; Elion Jani, HEC, University of Geneva, André Bender, HEC, University of Geneva and FAME

N°147 **Equity and Neutrality in Housing Taxation**
Philippe Thalmann, Ecole Polytechnique Fédérale de Lausanne

N°146 **Order Submission Strategies and Information: Empirical Evidence from NYSE**
Alessandro Beber, HEC, University of Lausanne and FAME; Cecilia Caglio, U.S. Security and Exchange Commission

N°145 **Kernel Based Goodness-of-Fit Tests for Copulas with Fixed Smoothing Parameters**
Olivier SCAILLET, HEC, University of Geneva and FAME

N°144 **Multivariate wavelet-based shape preserving estimation for dependent observations**
Antonio COSMA, Instituto di Finanza, University of Lugano, Lugano, Olivier SCAILLET, HEC, University of Geneva and FAME, Geneva, Rainier von SACHS, Institut de statistique, Université catholique de Louvain

N°143 **A Kolmogorov-Smirnov type test for shortfall dominance against parametric alternatives**
Michel DENUIT, Université de Louvain, Anne-Cécile GODERNIAUX, Haute Ecole Blaise Pascal Virton, Olivier SCAILLET, HEC, University of Geneva and FAME

N°142 **Times-to-Default: Life Cycle, Global and Industry Cycle Impacts**
Fabien COUDREC, University of Geneva and FAME, Olivier RENAULT, FERC, Warwick Business School

N°141 **Understanding Default Risk Through Nonparametric Intensity Estimation**
Fabien COUDREC, University of Geneva and FAME

International Center FAME - Partner Institutions

The University of Geneva

The University of Geneva, originally known as the Academy of Geneva, was founded in 1559 by Jean Calvin and Theodore de Beze. In 1873, The Academy of Geneva became the University of Geneva with the creation of a medical school. The Faculty of Economic and Social Sciences was created in 1915. The university is now composed of seven faculties of science; medicine; arts; law; economic and social sciences; psychology; education, and theology. It also includes a school of translation and interpretation; an institute of architecture; seven interdisciplinary centers and six associated institutes.

More than 13’000 students, the majority being foreigners, are enrolled in the various programs from the licence to high-level doctorates. A staff of more than 2’500 persons (professors, lecturers and assistants) is dedicated to the transmission and advancement of scientific knowledge through teaching as well as fundamental and applied research. The University of Geneva has been able to preserve the ancient European tradition of an academic community located in the heart of the city. This favors not only interaction between students, but also their integration in the population and in their participation of the particularly rich artistic and cultural life. ***http://www.unige.ch***

The University of Lausanne

Founded as an academy in 1537, the University of Lausanne (UNIL) is a modern institution of higher education and advanced research. Together with the neighboring Federal Polytechnic Institute of Lausanne, it comprises vast facilities and extends its influence beyond the city and the canton into regional, national, and international spheres.

Lausanne is a comprehensive university composed of seven Schools and Faculties: religious studies; law; arts; social and political sciences; business; science and medicine. With its 9’000 students, it is a medium-sized institution able to foster contact between students and professors as well as to encourage interdisciplinary work. The five humanities faculties and the science faculty are situated on the shores of Lake Leman in the Dorigny plains, a magnificent area of forest and fields that may have inspired the landscape depicted in Brueghel the Elder's masterpiece, the Harvesters. The institutes and various centers of the School of Medicine are grouped around the hospitals in the center of Lausanne. The Institute of Biochemistry is located in Epalinges, in the northern hills overlooking the city. ***http://www.unil.ch***

The Graduate Institute of International Studies

The Graduate Institute of International Studies is a teaching and research institution devoted to the study of international relations at the graduate level. It was founded in 1927 by Professor William Rappard to contribute through scholarships to the experience of international co-operation which the establishment of the League of Nations in Geneva represented at that time. The Institute is a self-governing foundation closely connected with, but independent of, the University of Geneva.

The Institute attempts to be both international and pluridisciplinary. The subjects in its curriculum, the composition of its teaching staff and the diversity of origin of its student body, confer upon it its international character. Professors teaching at the Institute come from all regions of the world, and the approximately 650 students arrive from some 60 different countries. Its international character is further emphasized by the use of both English and French as working languages. Its pluralistic approach - which draws upon the methods of economics, history, law, and political science - reflects its aim to provide a broad approach and in-depth understanding of international relations in general. ***http://heiwww.unige.ch***

FAME

INTERNATIONAL CENTER FOR
FINANCIAL ASSET MANAGEMENT AND ENGINEERING

40, Bd. du Pont d'Arve
PO Box, 1211 Geneva 4
Switzerland
Tel (++4122) 312 09 61
Fax (++4122) 312 10 26
http: //www.fame.ch
E-mail: admin@fame.ch

UNIVERSITÉ DE GENÈVE

THE GRADUATE INSTITUTE OF
INTERNATIONAL STUDIES

UNIVERSITE DE LAUSANNE