Geoadditive Hedonic Pricing Models

Andrew Chernih *†
Optimal Decisions Group
c/o CVA Level 65 MLC Centre
Sydney, Australia, 2000
Tel: +61 (0) 410 697 411
fax: +61 2 8257 0899
achernih@optimal-decisions.com

Prof. Michael Sherris
Actuarial Studies
Faculty of Commerce and Economics
University of New South Wales
Sydney, Australia, 2052
Tel: + 61 2 9385 2333
m.sherris@unsw.edu.au

April 15, 2004

Abstract

Until recently most analysis of property prices and the explanatory factors for these prices has been based on hedonic pricing models and linear regression. In some cases, krigging has been used to capture spatial dependence. More recently, spatial modelling techniques have become important in the modelling and analysis of issues in real estate valuation and policy development. This paper presents the application of a new approach to dealing with spatially dependent data. The application is to an hedonic analysis of residential property prices comprising over 37,000 houses that sold in Sydney in 2001, which makes it one of the largest hedonic analyses of residential property to date. The study considers the impact of noise and air pollution, closeness to green space, access to main roads, along with many other factors on residential property prices in Sydney. The additional flexibility offered by a semiparametric spatial model permits more accurate estimation of the underlying spatial structure and more reliable results for the impact

*The authors acknowledge the financial and other support of the Department of Environment and Conservation, IAG, Residex, PwC and the Institute of Actuaries of Australia. In particular support from Simon Smith, Dan Tess, Tony Coleman, Mary Haines, Dan Liebke, John Edwards and Grant Billen.

†Corresponding author.

of these factors. The study quantifies the non-linear effect of these factors on property prices. Although a number of factors are found to have similar effects to those found in previous studies, the quantification of the non-linear effects is new and interesting.

Key Words: hedonic price model, geoadditive model, residential property prices

1 Introduction

Hedonic price analysis fits a function of house price (the natural logarithm being a common choice) to a set of covariates, such as proximity to transport, environmental attributes and structural characteristics. Rosen[9] has been credited with pioneering this approach. Previous studies have generally been based on linear models and, with the large number of realted factors usually included in these studies, the impact of multi-collinearity is often not considered carefully.

It has been recognized that there is an additional complication with estimation procedures due to the presence of spatial dependence in observations. Specifically, houses that are in closer proximity are more likely to have similar environmental and accessibility characteristics and houses in the same area are more likely to have been constructed at a similar point in time and hence to have similar structural characteristics. It has been shown by many previous papers (Dubin[4]) that neglect of spatial dependence in pricing models not only affects the magnitudes of the estimates and their significance, but may also lead to serious errors in the interpretation of standard regression diagnostics such as tests for heteroskedasticity (Kim *et al.*[8]).

There have been numerous techniques applied to account for spatial dependence. Dubin[3] applied the geostatistical method of *kriging* to the estimation of the covariance structure in the model, while Can[1] introduced a spatially weighted dependent variable (spatial lag) as well as varying coefficients to capture neighbourhood effects. These spatial econometric techniques were extended in Kim *et al.*[8] where both spatial-lag and spatial-error models are utilised.

Clapp[2] includes non-spatial factors linearly in the hedonic pricing equation and fits local polynomial regression to the resulting spatial surface. This is designed for the purpose of assisting automated valuation algorithms.

This paper presents the application of a new approach to dealing with spatially dependent data. The application is to an hedonic analysis of residential property prices comprising over 37,000 houses that sold in Sydney in 2001, which to our knowledge makes it one of the largest hedonic analyses of residential property to date. The study considers the impact of noise and air pollution, closeness to green space, access to main roads, along with many other factors on residential property prices in Sydney. The additional flexibility offered by a semiparametric spatial model permits more accurate estimation of the underlying spatial structure and more reliable results for the impact of these factors. The study quantifies the non-linear effect of these factors on property prices. Although a number of factors are found to have similar effects to those found in previous studies, the quantification of

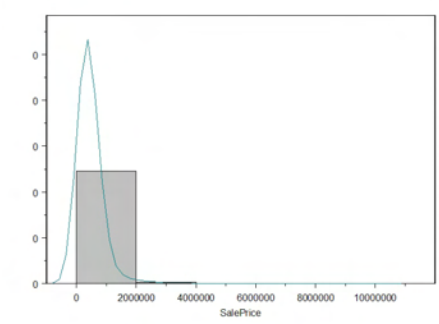

Figure 1: *Histogram of Sale Price*

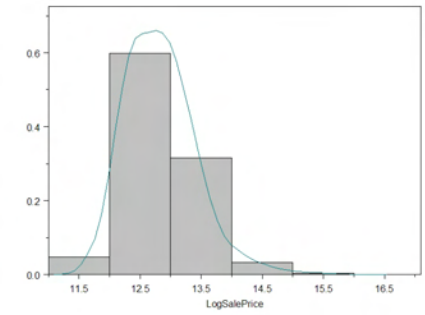

Figure 2: *Histogram of Log of Sale Price*

the non-linear effects is new and interesting.

This paper fits non-spatial factors in both linear and nonlinear ways to examine the need for nonlinear models in hedonic equations. This is important due to the fact that the clear majority of hedonic pricing literature has assumed linear relationships. Secondly rather than assuming a fixed covariance structure, a more flexible semiparametric function is fitted to the underlying spatial structure. The effect of neglecting spatial dependence on the hedonic pricing equation is also examined.

The structure of the paper is as follows: Section 2 describes the data set. Section 3 describes the statistical methodology used, in particular with regards to splines. Section 4 describes the four proposed models, followed by analysis of the complete dataset in Section 5 then partial dataset analysis in Section 6. Section 7 concludes and discusses areas for further research.

2 Data

The complete dataset comprised 37,676 properties which sold in the Sydney Statistical Division (an official geographical region including Sydney) for between $100,000 and $10,000,000 in the calendar year of 2001. These properties were geocoded in Mapinfo, a mapping program. A histogram of the sale amounts is provided in Figure 2, a histogram of the log of the sale amounts is provided in Figure 1. A plot of the distribution of properties showed a general spread across the geographical area under consideration. Although valuations were available for a larger number of properties, this study used actual recorded sale prices.

Table 1 provides summary statistics for certain variables in the dataset. A definition of all variables in used in this study is set out in Appendix 1.

Variable	**Mean**	**Median**	**Standard Deviation**	**Minimum**	**Maximum**
Sale Price	459526.08	355000	442070.19	100455	9970000
Log Sale Price	12.82934	12.77987	0.58688	11.51747	16.11509
LotSize	693.95749	664	175.34151	400	2000
Crime Rate	0.19635	0.20692	0.07554	0.10786	1.08818
Income	1122.6	1100	338.54048	259	2000
National Park	4.77914	4.35716	2.8199	0.043	14.92772
Park	0.38255	0.31198	0.32853	0.00407	5.98709
RailStation	2.672393	1.785	2.69035	0.02922	16.06727
Highway	4.09308	2.68613	4.37822	0.00516	26.56455
Freeway	4.2018	3.10764	3.75419	0.02054	28.92147
MainRoad	0.75043	0.50341	0.73045	0	6.78244
AirNoise	0.61206	0	3.61882	0	35
Foreignerratio	0.63798	0.65335	0.1214	0.1002	0.89834
GPO	22.5553	20.80927	13.13602	0.94095	68.86061
PM_{10}	18.32468	18.72994	1.69415	15.84026	20.85395
NEPH	0.30134	0.30207	0.04404	0.22842	0.38689
Ambulance	3.74778	3.11748	2.66475	0.0342	17.1209
Factory	4.10612	3.505677	2.4849	0.00967	20.106171

Table 1: *Summary Statistics for Certain Variables in Final Dataset*

Lot size information came from the Valuer General, a governmental division. If the recorded lot size for a property did not fall between the values of $400m^2$ and $2000m^2$ then the EM algorithm was used to infer an estimated lot size. This meant that approximately 50% of properties required the EM algorithm to obtain a value for lot size. An analysis of mean and variance for the EM derived sample indicated this provided resonable estimates for the purpose of this study. Lot size was used as a proxy for housing characteristics in this study since details on structural attributes for individual properties were not available for all the properties in the study.

The Environment Protection Authority (now part of The Department of Environment and Conservation) was able to provide data from seventeen monitoring stations across Sydney regarding daily average and daily maximum readings of carbon oxide, NEPH (suspended matter - nephelometer), nitrous oxide, nitrogen dioxide, ozone, PM_{10} (particulate matter with a diameter of under 10μm), $PM_{2.5}$ (particulate matter with a diameter of under 2.5μm) and sulfur dioxide. Values for individual properties were assigned based on the value at the nearest monitoring station. Not all air pollutants were recorded at each monitoring station. Only pollutants which were recorded at 10 or more monitoring stations were considered in the hedonic pricing analysis, which were NEPH, nitrous oxide, nitrogen dioxide, ozone and PM_{10}.

Mapinfo was used to determine the distance from each property to the nearest of various infrastructure and green space features, such as roads, train stations and parks. Neighbourhood information, comprising mean household income and percentage of people born outside of Australia, was available at Collection District level, which is the lowest level of aggregation for Australian Census information.

3 Smoothers

A feature of this study is the application of recently developed techniques to fit non-linear relationships to multivariate data. Application of similar techniques are found in actuarial science where mortality and morbidity tables are constructed taking into account goodness-of-fit and smoothness.

A smoother is a tool for summarising the trend of a response measurement Y as a function of one or more predictor measurements $X_1, \ldots, X_p$. An important property of a smoother is its nonparametric nature. It does not assume a rigid nature for the dependence of Y on $X_1, \ldots, X_p$. We consider two smoothers here, for the case of a univariate explanatory variable and then for the case of bivariate explanatory variables.

3.1 Cubic Smoothing Spline

A smoothing spline is the solution to the following optimisation problem: among all functions $\eta(x)$ with two continuous derivatives, find the one that minimises the penalised least square

$$\sum_{i=1}^{n} (y_i - \eta(x_i))^2 + \lambda \int_a^b \left(\eta''(t)\right)^2 dt$$

where λ is a fixed constant, and $a \leq x_1 \leq \ldots \leq x_n \leq b$. The first term measures closeness to the data while the second term penalises curvature in the function. It can be shown that there exists an explicit, unique minimiser, and that minimiser is a cubic spline with knots at the unique values of x_i.

The parameter λ is the smoothing parameter. Larger values of λ produce smoother curves. Smaller values fit the data more closely in terms of least squares. Selection of λ is discussed below.

3.2 Thin-Plate Smoothing Spline

Thin-plate smoothing splines are the natural bivariate extension of cubic smoothing splines.

Suppose that H_m is a space of functions whose partial derivatives of total order m are in $L_2(E^d)$ where E^d is the domain of $\mathbf{x}$. Consider the data model:

$$y_i = f\left(x_1(i), \ldots, x_d(i)\right) + \epsilon_i, \;\; i = 1, \ldots, n$$

where $f \in \mathcal{H}_m$.

Then f is estimated by minimising the penalised least squares function. Define $\mathbf{x}_i$ as a d-dimensional covariate vector, $\mathbf{z}_i$ as a p-dimensional covariate vector, and y_i as the observation associated with $(\mathbf{x}_i, \mathbf{z}_i)$. Assuming that the relation between $\mathbf{z}_i$ and y_i is linear but the relation between $\mathbf{x}_i$ and y_i is unknown, the data can be fit using a semiparametric model as follows:

$$y_i = f(\mathbf{x}_i) + \mathbf{z}_i\beta + \epsilon_i$$

where f is an unknown function that is assumed to be reasonably smooth, $\epsilon_i, \;\; i = 1, \ldots, n$ are independent, zero-mean random errors, and β is a p-dimensional unknown parametric vector.

Estimate f, for a fixed λ, by minimising the penalised least squares function:

$$\frac{1}{n}\sum_{i=1}^{n} (y_i - f(\mathbf{x}_i) - \mathbf{z}_i\beta)^2 + \lambda \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \sum \frac{m!}{\alpha_1! \ldots \alpha_d!} \left[\frac{\partial^m f}{\partial x_1^{\alpha_1} \ldots x_d^{\alpha_d}}\right]^2 dx_1 \ldots dx_d$$

Further details are found in Wahba[12].

4 Functional Forms

For the hedonic pricing equation, four functional forms are used in this study. These include the standard hedonic pricing models often used in previous studies along with models that allow non-linear relationships between the explanatory variables and property prices as well as those that include an allowance for spatial dependence. The models are set out first followed by the results from fitting the models.

4.1 Model 1

The first is just the standard linear regression, which has been the favoured choice of many previous studies. For a random variable Y, explained in terms of a set of explanatory variables $\{X_i\}_{i=1}^{p}$, the model takes the following form:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon,$$

where $\beta_i, i = 0 \ldots n$ are constants and ϵ are the error or disturbance terms, usually assumed to be $Normal(0, \sigma^2)$.

4.2 Model 2

The second model adjusts the multiple linear regression to include spatial dependence. A model with spatial dependence to explain a random variable Y in terms of a set of explanatory variables $\{X_i\}_{i=1}^{p}$, takes the form:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + f(L_1, L_2) + \epsilon,$$

where $\beta_i, i = 0 \ldots n$ are constants and $f(L_1, L_2)$ is a bivariate thin plate spline fitted to the longitude, denoted by L_1, and to the latitude, denoted by L_2, and ϵ are the error or disturbance terms. An example of a bivariate thin plate spline can be seen in Figure 55 and it will be discussed in further detail later.

4.3 Model 3

A (normal) additive model used to explain a random variable Y in terms of a set of explanatory variables $\{X_i\}_{i=1}^{p}$, takes the form:

$$Y = \beta_0 + \sum_{i=1}^{p} f_i(X_i) + \epsilon$$

where β_0 is a constant and f_i are smooth, but otherwise unspecified, functions of the explanatory variables. These functions are fit using cubic smoothing splines.

Additive models have many attractive features. The joint effect of all the predictor variables upon the response is expressed as a sum of individual effects. These individual effects show how the expected response varies as any single predictor varies with the others held fixed at arbitrary values; because of assumed additivity in the model the effect of one predictor does not depend on the values at which the others are fixed.

Partial prediction plots can be produced for an additive model, which shows the effect on the dependent variable of changes in each independent variable, holding all other independent variables constant. With a linear regression, this would just be a straight line with a gradient equal to the parameter estimate from the fitted model.

Additive models can also be extended to *generalised additive models*, in the same way as a linear model can be extended to a generalised linear model (Hastie and Tibshirani[6]).

4.4 Model 4

The combination of the normal additive model with a bivariate thin plate spline fit to longitude and latitude to account for spatial dependence is the fourth model, and this has been referred to as a *geoadditive* model in Kammann and Wand[7]. The geoadditive model used to explain a random variable Y in terms of a set of explanatory variables $\{X_i\}_{i=1}^{p}$ takes the following form:

$$Y = \beta_0 + \sum_{i=1}^{p} f_i(X_i) + f(L_1, L_2) + \epsilon$$

where β_0 is a constant and $f(L_1, L_2)$ is a bivariate thin plate spline fitted to the longitude, denoted by L_1, and to the latitude, denoted by L_2 and the univariate smooth functions are fit using cubic smoothing splines and ϵ are the error or disturbance terms.

4.5 Model Fitting

In our study we found that the computational requirements of fitting bivariate thin plate splines in SAS did not allow us to fit Models 2 and 4 to the entire dataset. Therefore section 4 will present the results from fitting Models 1 and 3 to the complete dataset and Section 5 will present the results from fitting all 4 models to a smaller sample. The smaller sample was a representative random selection of 1000 properties used to analyse the differences between the 4 models and to understand the impact of the inclusion of a bivariate thin plate spline to account for spatial dependence on the significance and magnitude of factors in the hedonic pricing equation.

4.6 A Note about SAS

The GAM procedure in SAS that was used to fit the linear regression with bivariate smoothing, the normal additive model and the geoadditive model is experimental, and we occasionally found that it does not complete the Analysis of Deviance table which is part of the standard output. The degrees of freedom are still reported for each parameter, and partial plots are produced, however when this happens it is not possible to test whether the nonparametric spline component was significant for several explanatory variables from the output. The Analysis of Deviance showed that for nearly every factor, a smoothing spline was significant with a p-value of less than 0.0001 when the full output was produced. In cases where the Analysis of Deviance table was incomplete, it was assumed that the variables were still significant.

5 Full Dataset Analysis

5.1 Model Selection

In order to develop an hedonic pricing model it is necessary to determine the significant factors and the functional form for the model based on the data. The model was fitted by attempting to adhere to five general criteria:

- *Parsimony*: Models should contain the fewest number of parameters, yet provide a sufficient goodness of fit;
- *Interpretation*: Models should be easily and meaningfully interpretable;
- *Significant Effects*: Factors included in the model should be statistically significant;

- *Goodness of fit*: Models should be selected according to goodness of fit;

- *Meeting of Assumptions*: Model assumptions should be satisfactory.

Factor analysis was used to identify clusters of variables with common characteristics to assist in variable selection. This was used to separate the 45 variables into 8 categories of variables with a similar effect. Adjusted R square was the primary goodness of fit measure used for comparing linear regression models.

PROC GAM in SAS does not produce any goodness of fit measures as part of the default output and given that research into selection of explanatory variables for additive models is still in its early stages (see, eg: Ruppert *et al.*[11]), the same factors from the linear regression were used in the normal additive model. The selection of the smoothing parameter, λ, was done by use of the GCV criterion in SAS. Although this value is not reported in the SAS output, the degrees of freedom, which also indicate non-linearity, are reported.

5.2 Model 1

Using the log of the sale price as the dependent variable, Tables 2-4 gives the details of the fitted linear regression. The significant explanatory variables were found to be neighbourhood mean household income, lot size of property, exposure to aircraft noise, level of air pollution, house price inflation rate, distance to the city as well as distance to the nearest rail station, main road, park, highway, freeway, ambulance station and factory.

Many previous studies have ignored multicollinearity, however there are three possible consequences of multicollinearity. First, it becomes difficult to identify the separate effects of the variables involved precisely. Second, estimates may not appear significantly different from zero, or be of the wrong sign (assuming a prior expectation as to the sign of regression cofficients). Third, estimators may be very sensitive to the addition or deletion of a few observations or the deletion of apparently insignificant variables.

To assess multivariate multicollinearity, one can use the variance inflation factor test. The tolerance for an independent variable is defined as 1R2 for the regression of that independent variable on all the other independent variables, ignoring the dependent variable. Then the variance inflation factor is defined as the reciprocal of tolerance. It can be shown (Fox[5]) that the standard error of the regression coefficient is doubled when the variance inflation factor is

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	8200.57401	630.81339	4974.50	<0.0001
Error	37662	4775.90026	0.12681		
Corrected Total	37675	12976			

Table 2: *Analysis of Variance for Optimal Linear Regression Model - Complete Dataset*

Root MSE	0.3561
Dependent Mean	12.82934
Coeff Var	2.77569
R-Square	0.632
Adjusted R-Square	0.6318

Table 3: *Model Fit Diagnostics for Optimal Linear Regression - Complete Dataset*

4.0. Therefore a variance inflation factor of 4 is an arbitrary but common cut-off criterion for deciding when a given independent variable displays excessive multicollinearity and the same cut-off was used in this study.

The values of the variance inflation factors for this fitted model are given in Table 5.

5.2.1 *Analysis of Fit*

In Figure 3 the quantile-quantile plot can be seen. This clearly violates normality in both tails. This result was quite insensitive to the choice of explanatory variables and raises the issue of the appropriateness of a log-linear regression in hedonic pricing. This indicates the possible need for consideration of a generalised linear model, to allow for non-normal error distributions. The probability-probability plot in Figure 4 supports this conclusion.

Semivariograms are used to examine spatial dependence. The omnidirectional, east-west and north-south semivariograms are shown in Figures 5 - 7 and the correlogram is in Figure 8. The semivariograms all have a similar shape, indicating the possible presence of spatial autocorrelation in model residuals up to the distance at which the sill is reached (the range). Directional semivariograms are similar, indicating isotropic (non direction-dependent) autocorrelation. A similar conclusion follows from observation of the correlogram, which decays to zero at approximately the same distance as the range seen in the semivariograms. Consequently based on this analysis, it appears that spatial autocorrelation is present in the residuals, violating the assumption of independence.

For further information regarding spatial analysis, including semivari-

Variable	*Parameter Estimate*	*Standard Error*	*t value*	*p value*
(Intercept)	13.07178	0.02909	449.38	<0.0001
LotSize	0.00067892	0.00001126	60.29	<0.0001
Main Road	-0.03356	0.00275	-12.19	<0.0001
Inflation	0.39042	0.04322	9.03	<0.0001
Income	0.00047877	0.00000648	73.85	<0.0001
PM_{10}	-0.04155	0.00127	-32.83	<0.0001
GPO	-0.02675	0.00019982	-133.85	<0.0001
RailStation	0.00741	0.00098987	7.49	<0.0001
Park	0.01932	0.00586	3.3	0.001
Highway	0.00945	0.00068251	13.85	<0.0001
Freeway	0.01036	0.00062914	16.47	<0.0001
Ambulance	-0.00181	0.00090032	-2.01	0.0447
AirNoise	-0.00315	0.00052945	-5.95	<0.0001
Factory	0.00489	0.0008743	5.6	<0.0001

Table 4: *Parameter Estimates for Optimal Linear Regression Model - Complete Dataset*

Variable	*Variance Inflation*
Intercept	0
Lot Size	1.15829
Income	1.43117
PM_{10}	1.36581
GPO	2.04702
RailStation	2.10706
Park	1.10093
Highway	2.65288
Freeway	1.65742
Main Road	1.20107
Air Noise	1.09064
Inflation	1.01498
Ambulance	1.71006
Factory	1.40232

Table 5: *Variance Inflation Factors for Optimal Linear Regression Model - Complete Dataset*

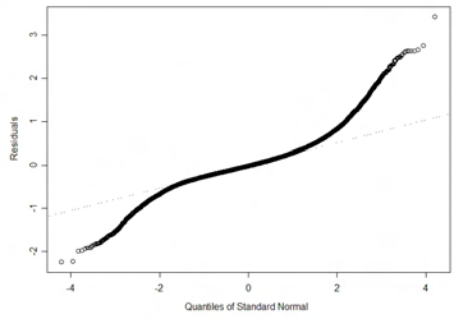

Figure 3: *Quantile-Quantile Plot for the Residuals of the Optimal Linear Regression Model - Complete Dataset*

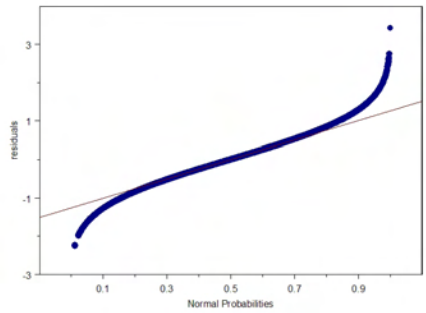

Figure 4: *Probability-Probability Plot for the Residuals of the Optimal Linear Regression Model - Complete Dataset*

ograms and correlograms, the interested reader is referred to the excellent book by Cressie[3].

5.3 Normal Additive Model

Given the same factors that were used in the multiple linear regression, and using the GCV criterion to select the smoothing parameter yielded the following results. Table 6 gives the parameter estimates for the parametric part of the model. Table 7 gives the analysis of deviance output for the nonparametric component of the model and then Figures 9 - 21 show the partial prediction plots for each explanatory variable.

In Figure 22 the quantile-quantile plot can be seen. This clearly violates normality in both tails, as was the case for the linear regression residuals. The probability-probability plot in Figure 23 supports this conclusion. This again questions the normality assumption and indicates the potential need for a generalised additive model, to allow for non-normal error distributions.

The omnidirectional, east-west and north-south semivariograms are shown in Figures 24 - 26 and the correlogram is in Figure 27. The north-south semivariogram appears quite flat, indicating no spatial dependence in this direction, but the same was not true for the omnidirectional and east-west semivariogram. The omnidirectional semivariogram has a slight upward trend, which appears to be driven by the east-west semivariogram (which has the same shape). Once again, it appears that spatial autocorrelation is present in the residuals, violating the assumption of independently distributed residuals.

The analysis of deviance table was not completed for this additive model

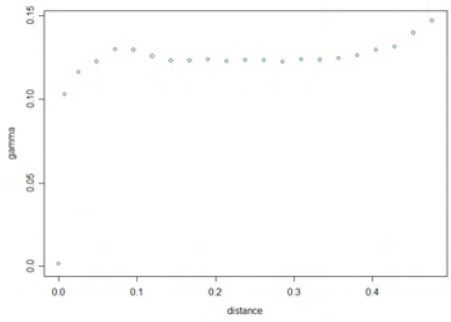

Figure 5: *Omnidirectional semivariogram for residuals of Optimal Linear Regression Model - Complete Dataset*

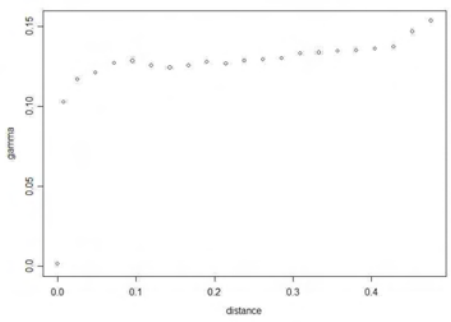

Figure 6: *East-west semivariogram for residuals of Optimal Linear Regression Model - Complete Dataset*

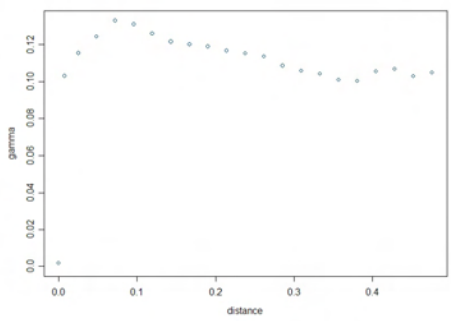

Figure 7: *North-south semivariogram for residuals of Optimal Linear Regression Model - Complete Dataset*

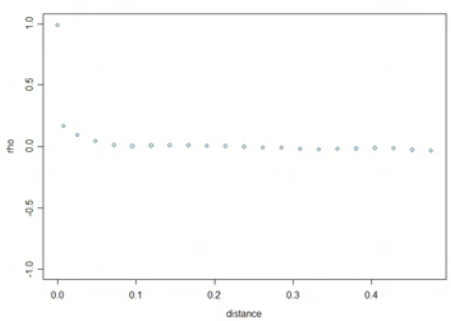

Figure 8: *Correlogram for residuals of Optimal Linear Regression Model - Complete Dataset*

Variable	*Parameter Estimate*	*Standard Error*	*t value*	*p value*
(Intercept)	12.97222	0.02504	518.01	<0.0001
LotSize	0.00068033	0.00000969	70.18	<0.0001
AirNoise	-0.00636	0.0004558	-13.96	<0.0001
Income	0.00036055	0.00000588	64.6	<0.0001
GPO	-0.02959	0.00017203	-172.00	<0.0001
Ambulance	0.01439	0.00077509	18.56	<0.0001
PM_{10}	-0.02624	0.00109	-24.08	<0.0001
RailStation	0.02516	0.00085218	29.53	<0.0001
Park	0.0084468	0.00504	0.17	0.867
Highway	-0.00302	0.00058758	-5.14	<0.0001
MainRoad	-0.0174	0.00237	-7.34	<0.0001
Factory	0.00156	0.00075269	2.08	0.0378
Freeway	0.00440	0.00054163	8.12	<0.0001
Inflation	0.24626	0.03721	6.62	<0.0001

Table 6: *Parameter Estimates for Normal Additive Model - Complete Dataset*

Variable	**Degrees of Freedom**	**Sum of Squares**	**Chi-Sq**	**Pr > Chi-Sq**
LotSize	2820.61937	.	.	.
AirNoise	4	11137	118499.263	< 0.0001
Income	58.34455	.	.	.
GPO	193.10864	5870.453335	62461.1367	<0.0001
Ambulance	51.80469	.	.	.
PM_{10}	5.76812	.	.	.
RailStation	42.65625	.	.	.
Park	25.625	6457.373539	68705.919	<0.0001
Highway	51.5	.	.	.
MainRoad	25.75	.	.	.
Factory	52.28125	.	.	.
Freeway	30.5	.	.	.
Inflation	3.92141	.	.	.

Table 7: *SAS Output Smoothing Model Analysis - Analysis of Deviance - Normal Additive Model - Complete Dataset*

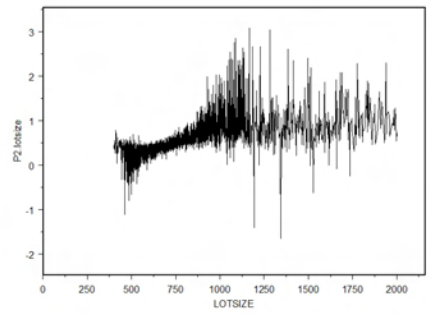

Figure 9: *Partial Prediction Plot for LotSize*

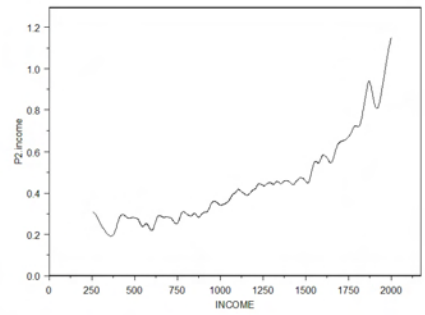

Figure 10: *Partial Prediction Plot for Income*

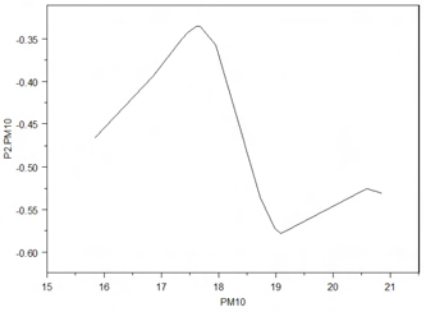

Figure 11: *Partial Prediction Plot for PM_{10}*

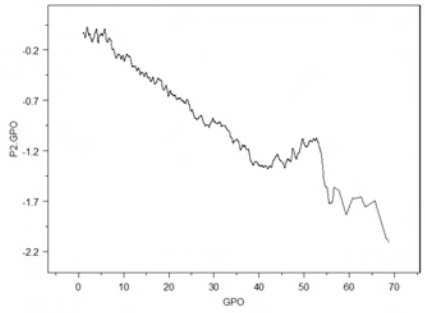

Figure 12: *Partial Prediction Plot for GPO*

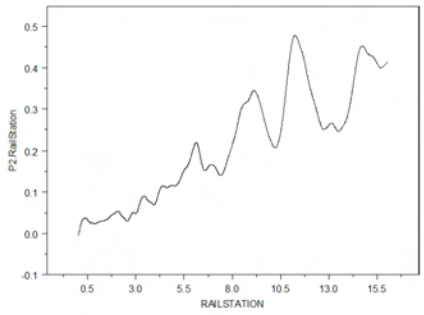

Figure 13: *Partial Prediction Plot for RailStation*

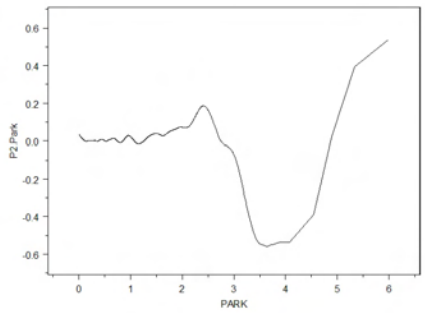

Figure 14: *Partial Prediction Plot for Park*

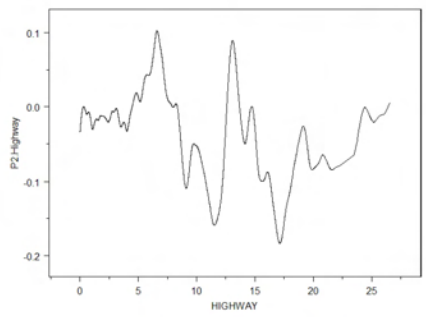

Figure 15: *Partial Prediction Plot for Highway*

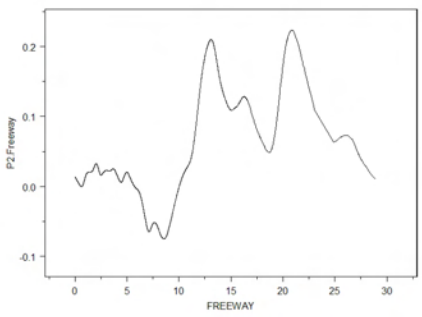

Figure 16: *Partial Prediction Plot for Freeway*

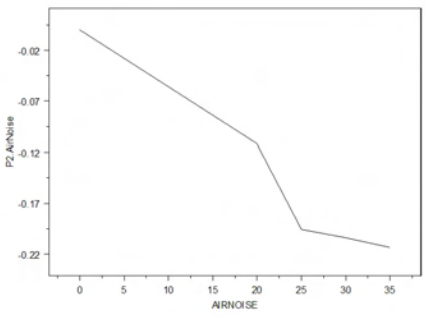

Figure 17: *Partial Prediction Plot for AirNoise*

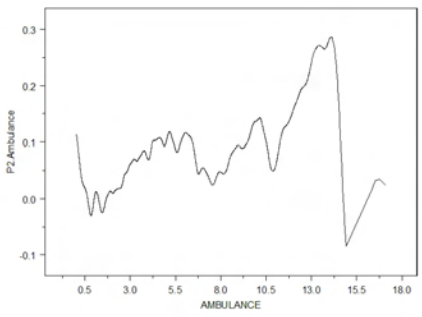

Figure 18: *Partial Prediction Plot for Ambulance*

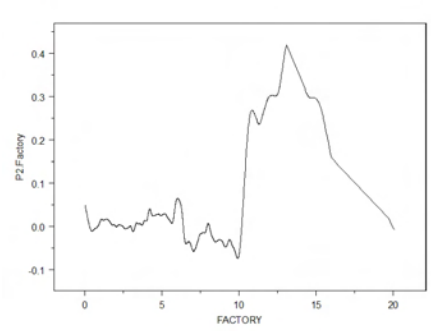

Figure 19: *Partial Prediction Plot for Factory*

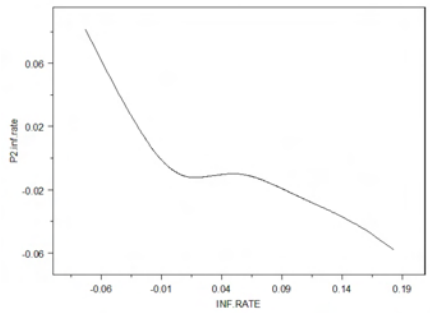

Figure 20: *Partial Prediction Plot for Inflation Rate*

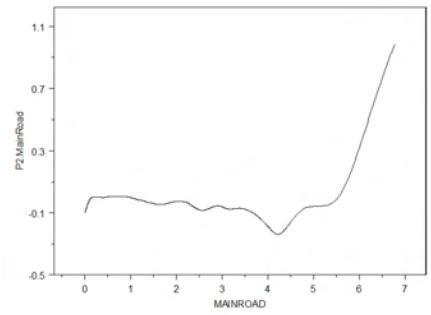

Figure 21: *Partial Prediction Plot for Main Road.*

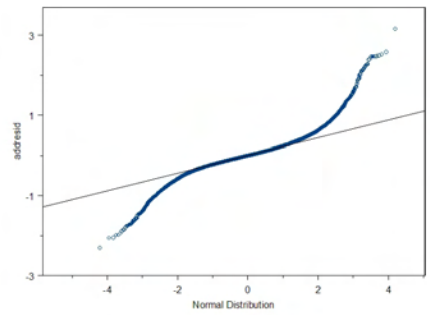

Figure 22: *Quantile-Quantile Plot for the Residuals of the Additive Model - Complete Dataset*

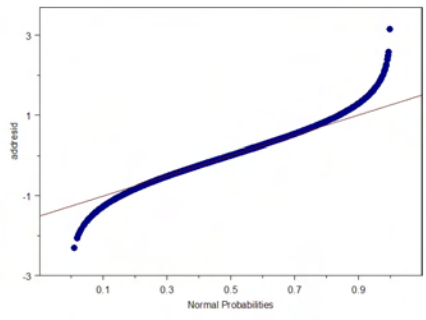

Figure 23: *Probability-Probability Plot for the Residuals of the Additive Model - Complete Dataset*

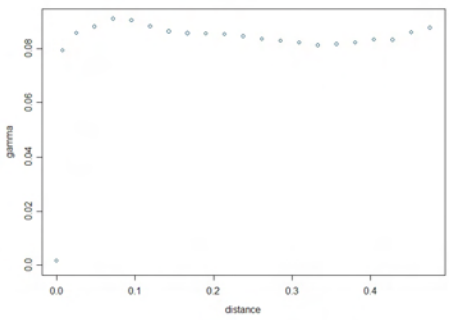

Figure 24: *Omnidirectional semivariogram for residuals of Additive Model - Complete Dataset*

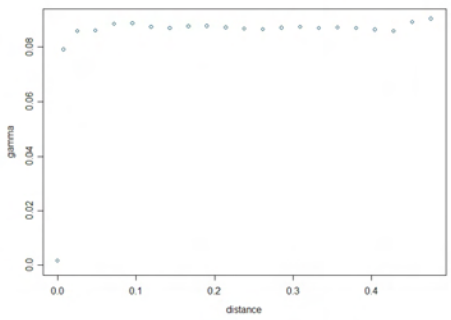

Figure 25: *East-west semivariogram for residuals of Additive Model - Complete Dataset*

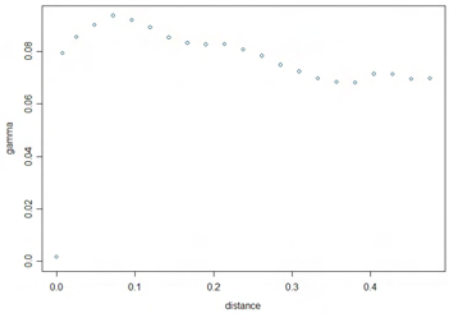

Figure 26: *North-south semivariogram for residuals of Additive Model - Complete Dataset*

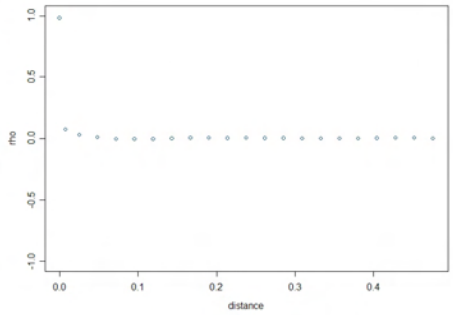

Figure 27: *Correlogram for residuals of Additive Model - Complete Dataset*

fit, and there are no test statistics for the significance of the smoothing parameters for some explanatory variables at the optimal degrees of freedom values. An alternative model was fit by manually selecting the degrees of freedom to be similar to the optimally selected degrees of freedom, such that the analysis of deviance is complete allowing analysis of significance. Details are given in Appendix 2 and this confirms the significance of the smoothing parameters and the significance of an additive model.

5.4 Full Dataset Model Comparison

A comparison of the two models fitted to the complete dataset - that is, models 1 and 3 as defined above - yields some interesting conclusions. The first is that there is significant statistical support based on this data for nonlinear models. Generalised additive models are well suited in terms of their flexibility and interpretability. The second is that spatial dependence appears to be less of an issue in the additive model, as can be seen from the semivariograms and correlograms, suggesting that non-linear effects could be misinterpreted to some extent as spatial dependence.

6 Partial Dataset Analysis

Using the smaller representative sample, a new set of explanatory variables for the linear regression model was determined. The same techniques were

Variable	*Parameter Estimate*	*Standard Error*	*t value*	*p value*
(Intercept)	13.14084	0.17355	75.72	<0.0001
LotSize	0.00067147	0.0000727	9.24	<0.0001
Income	0.00042651	0.00003788	11.26	<0.0001
GPO	-0.02724	0.0011	-24.81	<0.0001
PM_{10}	-0.03691	0.00772	-4.78	<0.0001
RailStation	0.01417	0.00554	2.56	0.0107
Highway	0.02112	0.00382	5.53	<0.0001
NatPark	-0.0185	0.00527	-3.51	0.0005

Table 8: *Parameter Estimates for Optimal Linear Regression Model - Partial Dataset*

used as were applied to the full dataset. These are reported in section 5.5.1. The other three models used the same explanatory variables and the resulting fits compared. Semivariograms and correlograms are reported and used to estimate the effect that including bivariate smoothers has on spatial dependence.

6.1 Model 1

Using the log of the sale price as the dependent variable, Table 8 gives the details of the fitted linear regression. No variance inflation factor value was larger than 2.1 and the adjusted R square was 64.89%.

The omnidirectional, east-west and north-south semivariograms are shown in Figures 28 - 30 and the correlogram is in Figure 31. The semivariograms all have a similar shape, indicating the possible presence of spatial autocorrelation in model residuals up to the distance at which the sill is reached (the range). There is a slight rise at very large distances, but since there are fewer observations at these distances and given the sensitivity of the semivariogram to data gaps, this is not regarded as significant.

Directional semivariograms were similar, indicating isotropic (non direction-dependent) autocorrelation. A similar conclusion follows from observation of the correlogram, which decays to zero at approximately the same distance as the range seen in the semivariograms. Based on this analysis, there is evidence of spatial autocorrelation in the residuals, violating the independence assumption.

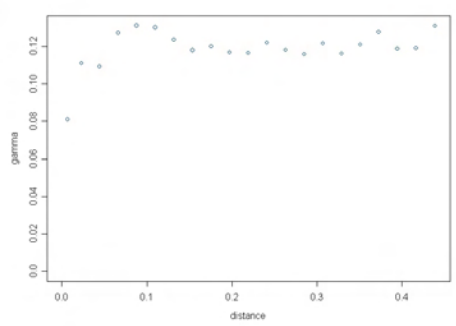

Figure 28: *Omnidirectional semivariogram for residuals of optimal linear regression model - - Partial Dataset*

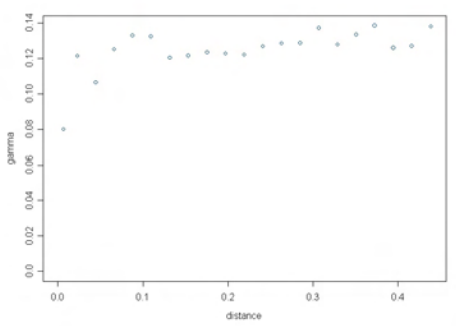

Figure 29: *East-west semivariogram for residuals of optimal linear regression model - - Partial Dataset*

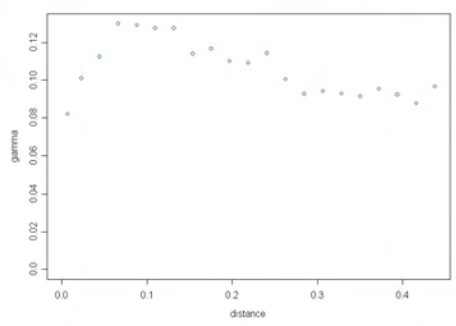

Figure 30: *North-south semivariogram for residuals of optimal linear regression model - - Partial Dataset*

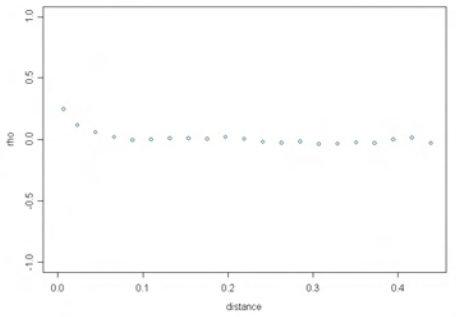

Figure 31: *Correlogram for residuals of optimal linear regression model - Partial Dataset*

Variable	*Parameter Estimate*	*Standard Error*	*t value*	*p value*
(Intercept)	12.8749	0.13525	94.99	<0.0001
LotSize	0.00065529	0.00005665	11.57	<0.0001
Income	0.00023429	0.00002952	7.94	<0.0001
GPO	-0.0288	0.00085567	-33.65	<0.0001
PM_{10}	-0.0088	0.00602	-1.46	0.144
RailStation	0.01471	0.00432	3.41	0.0007
Highway	0.01676	0.00298	5.63	<0.0001
NatPark	-0.0785	0.00411	-1.91	0.0564

Table 9: *Parameter Estimates for Optimal Linear Regression Model with Bivariate Smoothing - Partial Dataset*

Variable	**Degrees of Freedom**	**Sum of Squares**	**Chi-Sq**	**Pr > Chi-Sq**
(XCOORD,YCOORD)	213.47039	64.11275	862.5575	<0.0001

Table 10: *SAS Output Smoothing Model Analysis - Analysis of Deviance - Linear Regression with Bivariate Smoothing - Partial Dataset*

6.2 Model 2

Table 9 gives the linear regression parameter estimates as well as standard errors and associated p-values. Then Table 10 provides analysis of deviance for the bivariate thin plate spline fit to longitude and latitude. It is seen to be highly significant. The plot of this spline is in figure 55.

The omnidirectional, east-west and north-south semivariograms are shown in Figures 32 - 34 and the correlogram is in Figure 35. An analysis of the semivariograms shows that spatial autocorrelation is less noticeable in the residuals after including the bivariate smoothing spline. There is some spatial autocorrelation evident in the north-south semivariogram, but the omnidirectional and east-west semivariogram show minimal spatial autocorrelation, and this finding is supported by the correlogram.

6.3 Model 3

Table 11 gives the parameter estimates for the parametric part of the model. Table 12 gives the analysis of deviance output for the nonparametric component of the model.

The omnidirectional, east-west and north-south semivariograms are shown in Figures 36 - 38 and the correlogram is in Figure 39. An analysis of the

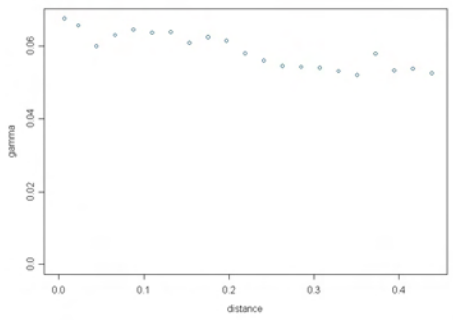

Figure 32: *Omnidirectional semivariogram for residuals of optimal linear regression model with bivariate smoothing - Partial Dataset*

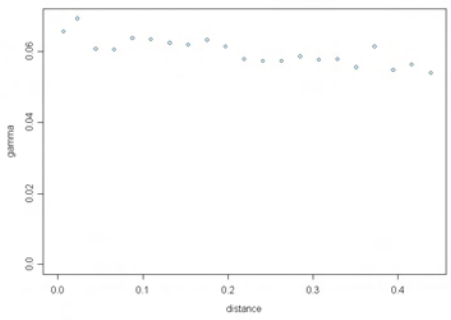

Figure 33: *East-west semivariogram for residuals of optimal linear regression model with bivariate smoothing - Partial Dataset*

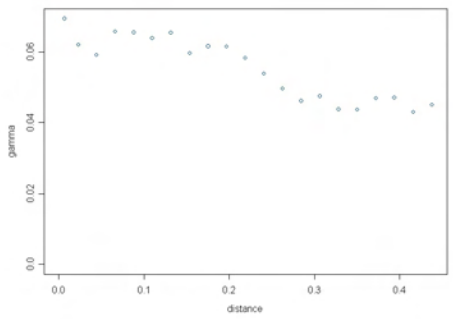

Figure 34: *North-south semivariogram for residuals of optimal linear regression model with bivariate smoothing - Partial Dataset*

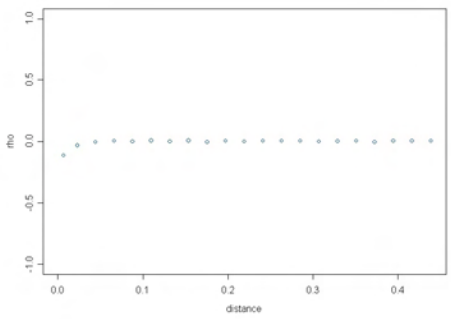

Figure 35: *Correlogram for residuals of optimal linear regression model with bivariate smoothing - Partial Dataset*

Variable	*Parameter Estimate*	*Standard Error*	*t value*	*p value*
(Intercept)	12.80372	0.15521	82.49	<0.0001
LotSize	0.00067923	0.00006501	10.45	< 0.0001
Income	0.00038986	0.00003388	11.51	<0.0001
GPO	-0.02722	0.00098194	-27.72	<0.0001
PM_{10}	-0.01509	0.0069	-2.19	0.0291
RailStation	0.03305	0.00496	6.67	<0.0001
Highway	0.00572	0.00342	1.67	0.0944
National Park	-0.02262	0.00472	-4.80	<0.0001

Table 11: *Parameter Estimates for Normal Additive Model - Partial Dataset*

Variable	**Degrees of Freedom**	**Sum of Squares**	**Chi-Sq**	**Pr > Chi-Sq**
Lotsize	34.20235	10.109605	102.3608	<0.0001
Income	15.49673	7.346238	74.3814	<0.0001
GPO	3.46228	2.271756	23.0018	<0.0001
PM_{10}	6.96644	5.358344	54.2537	<0.0001
RailStation	1.94727	0.271541	2.7494	0.2438
Highway	0.99957	0.271576	2.7497	0.0972
National Park	14.88785	4.725316	47.8443	<0.0001

Table 12: *SAS Output Smoothing Model Analysis - Analysis of Deviance - Normal Additive Model - Partial Dataset*

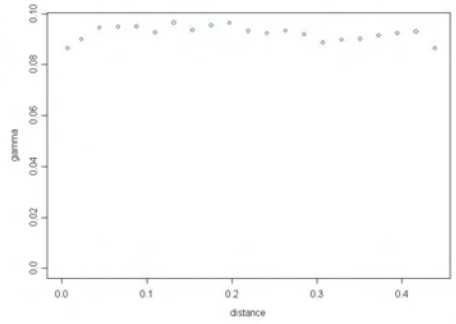

Figure 36: *Omnidirectional semivariogram for residuals of Additive Model - Partial Dataset*

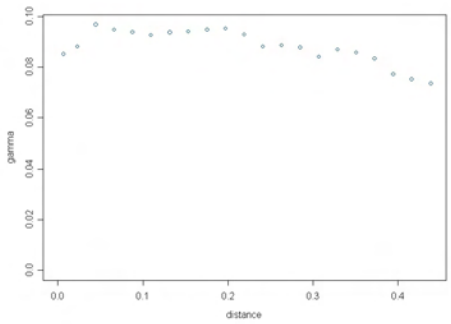

Figure 37: *East-west semivariogram for residuals of Additive Model - Partial Dataset*

three semivariograms shows little evidence of spatial autocorrelation. Once again a large rise is seen at larger distances in the north-south direction, and this may be due to fewer observations at this distance. The absence of spatial autocorrelation is supported by the correlogram, which is approximately zero for all distances.

6.4 Model 4

Table 13 gives the parameter estimates for the parametric part of the model. Table 14 gives the analysis of deviance output for the nonparametric component of the model. Figure 56 provides the plot of the fitted bivariate thin plate spline. It is worth noting that the coefficient of PM_{10} is positive in this model, whereas it was negative in model 2. This provides further evidence of how sensitive the parameter estimates can be to the inclusion of spatial dependence and hence the importance of understanding the role that spatial dependence plays in hedonic analysis.

The omnidirectional, east-west and north-south semivariograms are shown in Figures 40 - 42 and the correlogram is in Figure 43. The findings for the geoadditive model are similar to those for the additive model. Only the north-south semivariogram suggests spatial autocorrelation, although only at larger distances. The addition of bivariate smoothing appears to have decreased the degree of spatial autocorrelation in this direction, as can be seen from the decreased trend in the north-south semivariogram of the geoadditive model as compared to the additive model. It appears again that spatial

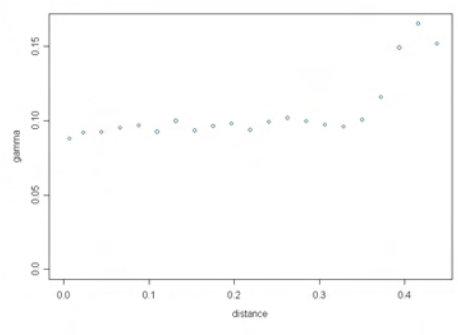

Figure 38: *North-south semivariogram for residuals of Additive Model - Partial Dataset*

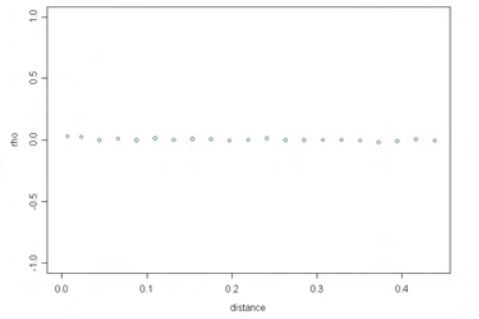

Figure 39: *Correlogram for residuals of Additive Model - Partial Dataset*

Variable	*Parameter Estimate*	*Standard Error*	*t value*	*p value*
(Intercept)	12.49604	0.13192	94.73	<0.0001
LotSize	0.00067942	0.00005525	12.30	<0.0001
Income	0.00026651	0.00002879	9.26	<0.0001
GPO	-0.02810	0.00083456	-33.68	<0.0001
PM_{10}	0.00642	0.00587	1.09	0.2746
RailStation	0.02042	0.00421	4.85	<0.0001
Highway	0.00945	0.00291	3.25	0.0012
NatPark	-0.00418	0.00401	-1.04	0.2978

Table 13: *Parameter Estimates for Geoadditive Model - Partial Dataset*

Variable	**Deg of Freedom**	**Sum of Squares**	**Chi-Sq**	**Pr > Chi-Sq**
Lotsize	38.10058	5.715484	80.1131	<0.0001
Income	15.28466	2.01616	28.2602	0.0224
GPO	1.01129	0.009638	0.1351	0.7176
PM_{10}	3.29792	0.255659	3.5835	0.3562
RailStation	1.00586	0.009622	0.1349	0.7158
Highway	0.99957	0.009639	0.1351	0.7130
National Park	15.72657	1.665948	23.3514	0.0965
(XCOORD,YCOORD)	182.11331	41.574308	582.7411	<0.0001

Table 14: *SAS Output Smoothing Model Analysis - Analysis of Deviance - Geoadditive Model - Partial Dataset*

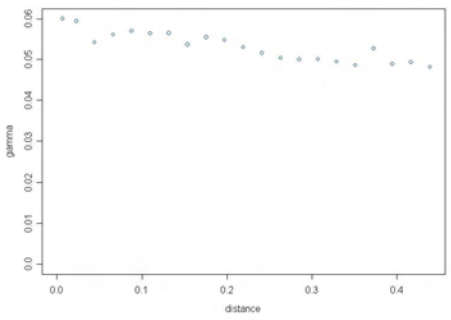

Figure 40: *Omnidirectional semivariogram for residuals of Geoadditive Model - Partial Dataset*

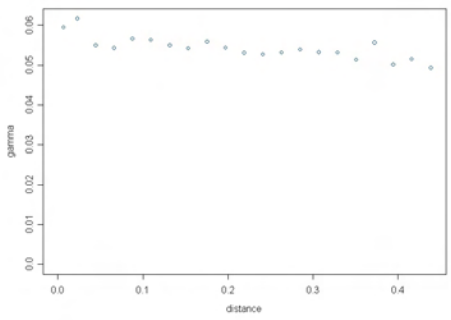

Figure 41: *East-west semivariogram for residuals of Geoadditive Model - Partial Dataset*

autocorrelation has been removed from the residuals, and consequently that the independence of residuals assumption is satisfied.

6.5 Model Comparison

Table 15 shows the percentage change in the coefficients in the linear regression model when a bivariate thin plate spline to longitude and latitude is included in the model. It is important to note that the inclusion of spatial dependence can make some variables insignificant presumably because their effect is captured in the spatial surface.[1] An asterisk denotes those parameters which became statistically insignificant at the 5% level after this inclusion (all parameters were significant at the 5% level in the original linear regression).

This table shows that for those parameters which remained significant, coefficient estimates changed in magnitude by 2.41% to 45.07%. This clearly shows that inclusion of a spatial structure, as well as being statistically significant, will influence the estimates of other explanatory factors in the functional form, supporting the findings of Kim *et al.*[8].

The partial prediction plots are produced for both the additive models and geoadditive models in Figures 44 - 51, with the additive model plot on the left and the geoadditive model plot on the right to make comparison easier. The partial prediction included only the component(s) that were found

[1]The authors thank Greg Taylor for this observation.

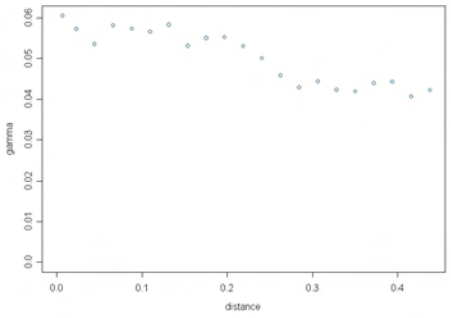

Figure 42: *North-south semivariogram for residuals of Geoadditive Model - Partial Dataset*

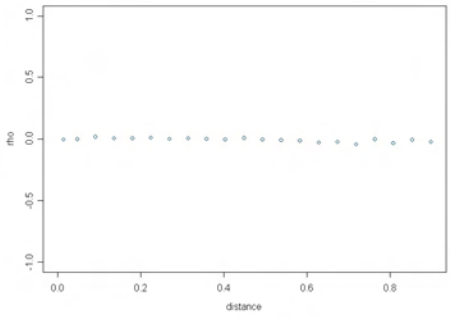

Figure 43: *Correlogram for residuals of Geoadditive Model - Partial Dataset*

Variable	**Percentage Change**
Lotsize	-2.41%
Income	-45.07 %
GPO	5.73 %
PM_{10}*	-76.16%
RailStation	3.81 %
Highway	-20.64%
National Park*	-324.32%

Table 15: *Percentage Changes in Coefficient Estimates for the Linear Regression Model Upon Inclusion of Bivariate Spatial Smoothing - Partial Dataset*

to be significant; that is, if the parametric or nonparametric component was not significant at the 5% level, it was not included in the partial prediction. For the additive model, Highway was not significant either parametrically or nonparametrically and for the geoadditive model PM_{10} and National Park were found to be insignificant parametrically and nonparametrically and consequently these partial plots were not produced.

Consequently the partial prediction plots for PM_{10} and NatPark for the additive model are in Figures 52 - 53 and the partial prediction plot for Highway for the geoadditive model is in Figure 54.

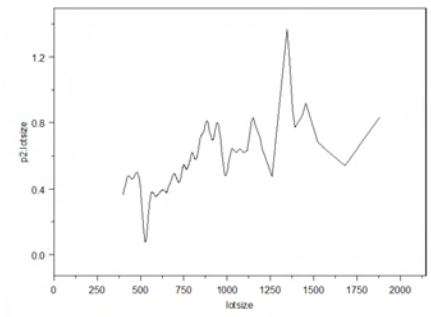

Figure 44: *Partial Prediction Plot for Lot-Size - Additive Model - Partial Dataset*

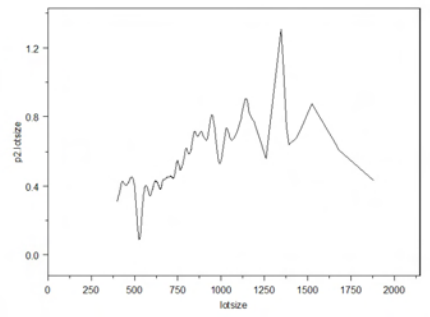

Figure 45: *Partial Prediction Plot for Lot-Size - Geoadditive Model - Partial Dataset*

Some interesting conclusions can be drawn from these graphs. Firstly, as in the linear regression case, the addition of bivariate thin plate splines has made some explanatory variables insignificant. But the effect of those that remain significant can be seen to be nearly unchanged - that is, the partial prediction plots for the additive and geoadditive models are nearly identical. The additive model may capture effects that would otherwise be treated as spatial dependence.

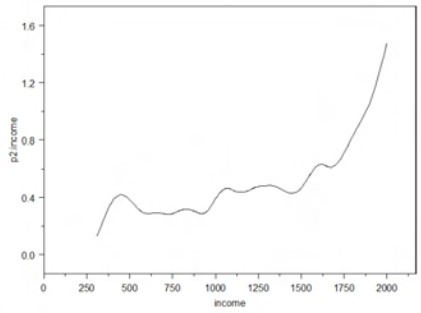

Figure 46: *Partial Prediction Plot for Income - Additive Model - Partial Dataset*

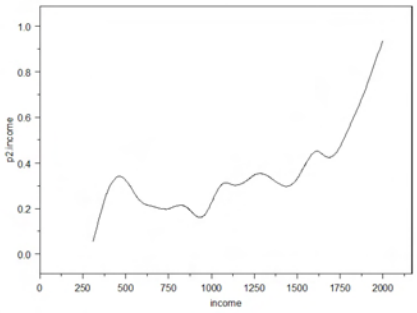

Figure 47: *Partial Prediction Plot for Income - Geoadditive Model - Partial Dataset*

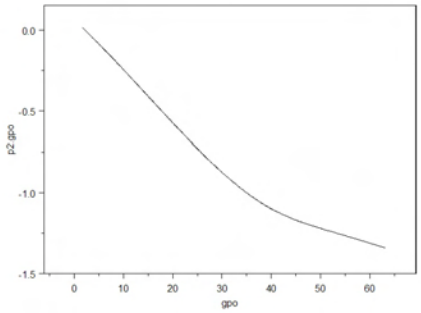

Figure 48: *Partial Prediction Plot for GPO - Additive Model - Partial Dataset*

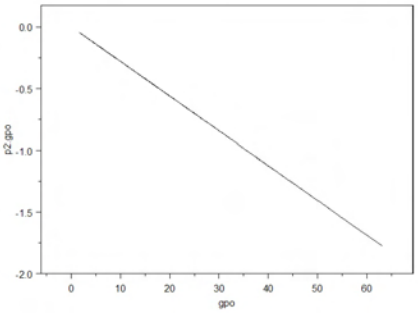

Figure 49: *Partial Prediction Plot for GPO - Geoadditive Model - Partial Dataset*

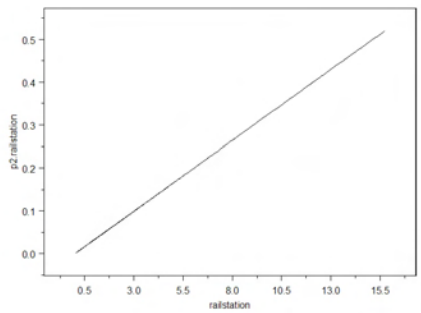

Figure 50: *Partial Prediction Plot for Railstation - Additive Model - Partial Dataset*

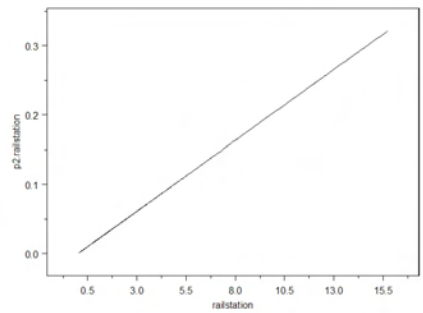

Figure 51: *Partial Prediction Plot for Railstation - Geoadditive Model - Partial Dataset*

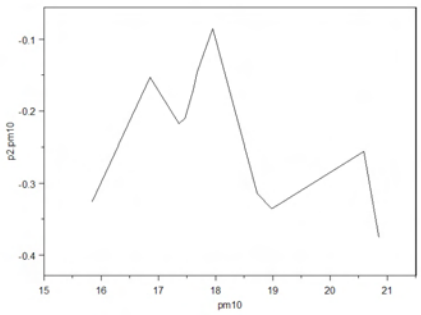

Figure 52: *Partial Prediction Plot for PM10 - Additive Model - Partial Dataset*

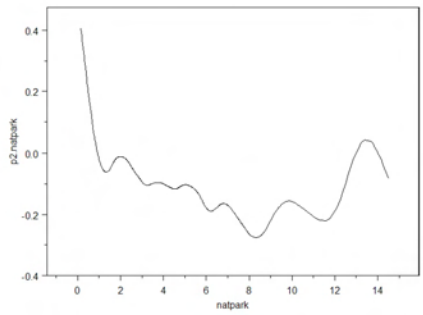

Figure 53: *Partial Prediction Plot for NatPark - Additive Model - Partial Dataset*

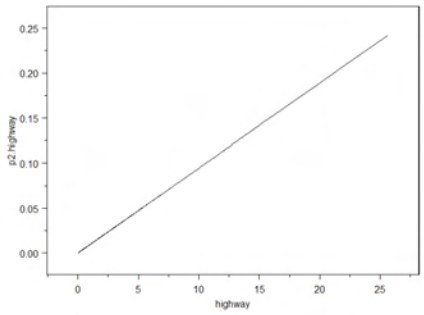

Figure 54: *Partial Prediction Plot for Highway - Geoadditive Model - Partial Dataset*

The partial prediction plots for longitude and latitude for the linear regression model with bivariate smoothing and for the geoadditive model are in Figures 55 and Figure 56.

Increasing longitude corresponds to being further east, and increasing latitude corresponds to a position further north. It is interesting to notice that the two plots are very similar, but the scale for partial LogSalePrice for the geoadditive model is -4 to 2, whilst for the additive model partial plot it was -6 to 3. The noticeably smaller peaks that occur at the boundaries would suggest that given the added flexibility of an additive model over a linear regression, this resulted in a more consistent spatial surface being constructed

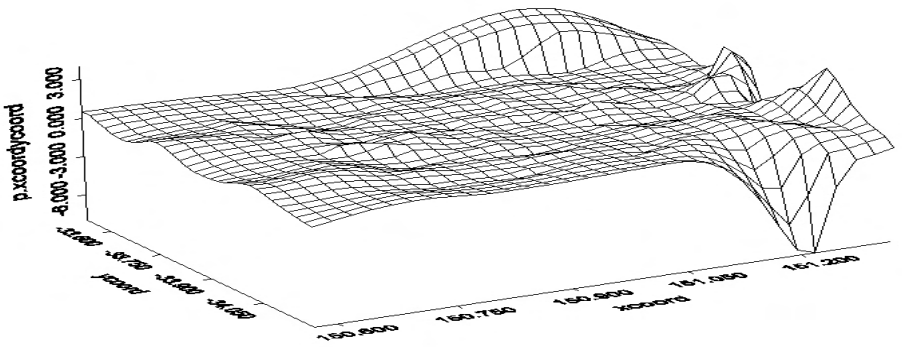

Figure 55: *Partial Prediction Plot for Longitude and Latitude - Linear Regression Model with Bivariate Smoothing - Partial Dataset*

7 Conclusion

This paper has investigated the effects of various factors such as pollution and proximity to parks public and public transport on residential property prices in Sydney. The most significant factors in explaining residential property prices in Sydney in 2001 were found to be lotsize, mean household income, level of aircraft noise, mean levels of PM_{10}, distance from the city and distances from the nearest main road, highway, freeway, rail station, park, national park, ambulance station and factory. This paper considered model issues and the results are considered in more detail, with reference to findings in previous hedonic pricing studies, in a forthcoming companion paper.

There was significant evidence supporting the need for additive models, both from analysing the residuals from the linear regression as well as from the tests of significance of the additive models. The added flexibility permits a more precise understanding of the impact of factors on property prices.

The inclusion of bivariate thin plate splines to account for spatial dependence was found to be highly significant and to have significant changes (in magnitude) of between 2.41% and 45.07% for other explanatory variables in the linear regression model as well as causing other factors to become statistically insignificant. Bivariate thin-plate splines were found to change significance of explanatory variables in the additive model as well, however

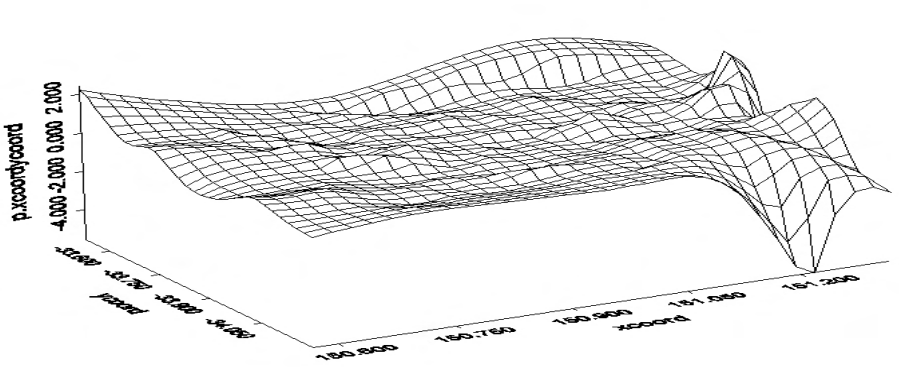

Figure 56: *Partial Prediction Plot for Longitude and Latitude - Geoadditive Model - Partial Dataset*

for those factors which remained significant, the effect was nearly unchanged, indicating increased stability to spatial dependence.

Additive models offer great potential for hedonic pricing, because they allow sophisticated relationships between each factor and sale price, however the issue of interactions becomes more important than in the case of a linear regression model, and more work is required both in respects of model selection and validation and also in what role splines can play in dealing with interactions in the data.

For both the linear regression and additive models, the assumption of normally distributed errors was not satisfied. Future research should consider the use of generalised linear models and generalised additive models to allow for non-normal distributions for the errors.

Bivariate thin plate splines offer a more flexible approach to dealing with spatial dependence and more research is required to investigate how this compares with traditional semivariogram approaches. The penalty parameter λ controls the trade-off between bias and variance. Having a single value of λ may not be suitable for a regression function which is spatially nonhomogeneous with some regions of rapidly changing curvature and other regions of little change in curvature.

Spatially adaptive smoothing can overcome this problem by allowing λ to vary spatially in order to accommodate possible spatial nonhomogeneity

of the regression function. In other words, λ is a function of the independent variable x. Allowing λ to be a function of spatial location can improve mean squared error and the accuracy of inference (Ruppert *et al.*[10]).

The issues of how to best incorporate non-linear effects and spatial dependence, along with multicollinearity, require consideration in any hedonic pricing study. However this is an area of developing research and many aspects remain unexplored. It will be of interest to follow future developments.

Appendix 1

The variables that are used in the full model are listed and defined below.

- *LogSalePrice*: Natural logarithm of the sale price
- *LotSize*: Lot size of the property
- *Longitude*: Longitude position of property
- *Latitude*: Latitude position of property
- *Income*: Mean household income
- *GPO*: Euclidean distance to the General Post Office
- PM_{10}: Mean Daily Value of PM_{10} (particulate matter with a diameter of under 10μm): a measure of air pollution
- *RailStation*: Euclidean distance to the nearest train station
- *Park*: Euclidean distance to the nearest park
- *Highway*: Euclidean distance to the nearest highway
- *Freeway*: Euclidean distance to the nearest freeway
- *Ambulance*: Euclidean distance to the nearest ambulance station
- *AirNoise*: ANEF value
- *Factory*: Euclidean distance to the nearest factory/power station/industrial feature
- *MainRoad*: Euclidean distance to the nearest main road (as defined in Mapinfo)
- *NatPark*: Euclidean distance to the nearest national park

Appendix 2

SAS failed to complete the Analysis of Deviance table for the additive model for the complete dataset, which meant that there were no tests for significance of the nonparametric components of several explanatory variables.

An alternative to allowing SAS to calculate the degrees of freedom for each explanatory variable is to manually select an approximate degrees of freedom to use. In this way, a similar degrees of freedom can be used for the majority of the variables in our additive model, but a complete analysis of deviance table will be produced. This is not designed to replace the optimal fit calculated by SAS, but merely to offer an insight into the significance of the fits. If the fit is highly significant in this output with a similar value of degrees of freedom to that in the optimal fit, it is likely that the fit is significant in the optimal case as well.

The degrees of freedom were selected to be as close to the optimally chosen degrees of freedom whilst also leading to a complete analysis of deviance table. This meant it was necessary to greatly increase the degrees of freedom for Highway. Its significance with such a large value for degrees of freedom means that it exhibits significant non-linearity and it is probable a similar finding would result with the use of the "optimal" degrees of freedom reported earlier.

The results are reported in Tables 16 and 17. The similarity of the results of the parameter estimates in this case to the results in the optimally selected model provides further support for these findings.

8 References

1. A. Can, Specification and estimation of hedonic housing pricing models, *Reg. Sci. Urban Econom.* **22**, 453-474 (1992).

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

3. N.A.C. Cressie. *Statistics for Spatial Data*. Wiley-Interscience (1993).

4. R.A. Dubin, Spatial Autocorrelation and neighborhood quality, *Reg. Sci. Urban Econom.* **22**, 432-452 (1992).

5. J. Fox, *Regression Diagnostics*. Thousand Oaks, CA: Sage Publications, 1991.

Variable	*Parameter Estimate*	*Standard Error*	*t value*	*p value*
(Intercept)	12.79081	0.02593	493.25	<0.0001
LotSize	0.00067895	0.00001010	67.28	<0.0001
AirNoise	-0.00830	0.00047540	-17.46	<0.0001
Income	0.00034130	0.00000590	57.87	<0.0001
GPO	-0.02808	0.00019747	-142.22	<0.0001
Ambulance	0.01784	0.00080863	22.06	<0.0001
PM_{10}	-0.016	0.00117	-13.63	<0.0001
RailStation	0.03805	0.00088785	42.85	<0.0001
Park	0.00883	0.00528	1.67	0.0947
Highway	-0.00824	0.00063160	-13.04	<0.0001
MainRoad	-0.01394	0.00248	-5.63	<0.0001
Factory	0.00365	0.00078450	4.65	<0.0001
Freeway	0.00269	0.00058259	4.61	<0.0001
National Park	-0.00808	0.00078356	-10.32	<0.0001

Table 16: *Parameter Estimates for Normal Additive Model with Manually Selected Degrees of Freedom - Complete Dataset*

Variable	**Degrees of Freedom**	**Sum of Squares**	**Chi-Sq**	**Pr > Chi-Sq**
Lotsize	149.91064	102.592270	1005.1865	< 0.0001
AirNoise	3.85929	0.524291	5.1369	0.2567
Income	10.13757	68.644921	672.5745	< 0.0001
GPO	138.66992	138.414487	1356.1682	< 0.0001
Ambulance	117.53906	25.346991	248.3467	< 0.0001
PM_{10}	8.72063	80.747492	791.1541	< 0.0001
RailStation	114.98828	25.887096	253.6386	< 0.0001
Park	100.52148	15.053080	147.4882	0.0016
Highway	5000.00000	641.528169	6285.6146	< 0.0001
MainRoad	90.90820	27.918981	273.5468	< 0.0001
Factory	200.01819	38.965038	381.7747	<0.0001
Freeway	100.67188	24.803487	243.0251	< 0.0001
National Park	131.90503	29.148066	285.5892	< 0.0001

Table 17: *SAS Output Smoothing Model Analysis - Analysis of Deviance - Normal Additive Model with Manually Selected Degrees of Freedom - Complete Dataset*

6. T.J. Hastie and R.J. Tibshirani, *Generalized Additive Models*. Chapman and Hall, New York (1990).

7. E.E. Kammann and M.P. Wand, Geoadditive Models, *Applied Statistics* **52**, 1-18 (2003).

8. C.W. Kim, T.T. Phipps and L. Anselin, Measuring the benefits of air quality improvement: a spatial hedonic approach. *Journal of Environmental Economics and Management* **45**, 24-39 (2003).

9. S. Rosen, Hedonic Prices and Implicit Markets: Product differentiation in pure competition, *Journal of Political Economy* **82**, 34-55 (1974).

10. D. Ruppert and R.J. Carroll, Spatially-adaptive penalties for spline fitting, *Australian and New Zealand Journal of Statistics* **42**, 205-24 (2000).

11. D. Ruppert, M.P. Wand and R.J. Carroll, *Semiparametric Regression*. Cambridge University Press, Cambridge, first edition (2003).

12. G. Wahba, *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia (1990).

13. M.P. Wand, A comparison of regression spline smoothing procedures, *Computational Statistics* **15**, 443-62 (2000).