Real estate appraisal of land lots using GAMLSS models

LUTEMBERG FLORENCIO[1], FRANCISCO CRIBARI-NETO[2] AND RAYDONAL OSPINA[2,*]

[1]Banco do Nordeste do Brasil S.A, Boa Vista, Recife/PE, 50060–004, Brazil
[2]Departamento de Estatística, Universidade Federal de Pernambuco, Recife/PE, Brazil

Abstract

The valuation of real estates is of extreme importance for decision making. Their singular characteristics make valuation through hedonic pricing methods difficult since the theory does not specify the correct regression functional form nor which explanatory variables should be included in the hedonic equation. In this article, we perform real estate appraisal using a class of regression models proposed by Rigby and Stasinopoulos (2005) called generalized additive models for location, scale and shape (GAMLSS). Our empirical analysis shows that these models seem to be more appropriate for estimation of the hedonic prices function than the regression models currently used to that end.

Keywords: Cubic splines · GAMLSS regression models · Hedonic prices function · Nonparametric smoothing · Semiparametric models.

Mathematics Subject Classification: Primary 62P25 · Secondary 62P20 · 62P05.

1. INTRODUCTION

The real estate, apart from being a consumer good that provides comfort and social status, is one of the economic pillars of all modern societies. It has become a form of stock capital, given the expectations of increasing prices, and a means of obtaining financial gains through rental revenues and sale profits. As a consequence, the real estate market value has become a parameter of extreme importance.

The estimation of a real estate value is usually done using a hedonic pricing equation according to the methodology proposed by Rosen (1974). This is seen as a heterogeneous good comprised of a set of characteristics and it is then important to estimate an explicit function, called hedonic price function, that determines which are the most influential attributes, or attribute "package", when it comes to determining its price. However, the estimation of a hedonic equation is not a trivial task since the theory does not determine the exact functional form nor the relevant conditioning variables.

*Corresponding author. Raydonal Ospina. Departamento de Estatística, Universidade Federal de Pernambuco, Cidade Universitária, Recife/PE, 50740–540, Brazil. Email: raydonal@de.ufpe.br

The use of classical regression methodologies, such as the classical normal linear regression model (CNLRM), for real estate appraisal can lead to biased, inefficient and/or inconsistent estimates given the inherent characteristics of the data (for example, nonnormality, heteroskedasticity and spatial correlation). The use of generalized linear models (GLM) is also subject to shortcomings, since the data may come from a distribution outside the exponential family and the functional relationship between the response and some conditioning variables may not be the same for all observations. There are semiparametric and nonparametric hedonic price estimations available in the literature, such as Pace (1993), Anglin and Gencay (1996), Gencay and Yang (1996), Thorsnes and McMillen (1998), Iwata et al. (2000) and Clapp et al. (2002). We also highlight the work of Martins-Filho and Bin (2005), who modeled data from the real estate market in Multnomah County (Oregon-USA) non-parametrically. We note that the use of nonparametric estimation strategies require very large datasets in order to avoid the "curse of dimensionality". Overall, however, most hedonic price estimations are based on traditional methodologies such as the classical linear regression model and the class of generalized linear models.

This article proposes a methodology for real estate mass appraisal[1] based on the class of GAMLSS models. The superiority GAMLSS modeling relative to traditional methodologies is evidenced by an empirical analysis that employs data on urban land lots located in the city of Aracaju, Brazil. We perform a real estate evaluation in which the response variable is the unit price of land lots and the independent variables are reflect the land lots structural, location and economic characteristics. We estimate the location and scale effects semiparametrically in such a way that some covariates (the geographical coordinates of the land lot, for instance) enter the predictor nonparametrically and their effects are estimated using smoothing splines (see Silverman, 1984; Eubank, 1999) whereas other regressors are included in the predictor in the usual parametric fashion. The model delivers a fit that is clearly superior to those obtained using the usual approaches. In particular, we note that our fit yields a very high pseudo-R^2.

The paper unfolds as follows. In Section 2, we briefly present the class of GAMLSS models and highlight its main advantages. In Section 3, we describe the data used in the empirical analysis. In Section 4, we present and discuss the empirical results. Finally, Section 5 closes the paper with some concluding remarks.

2. GAMLSS Modeling

Rigby and Stasinopoulos (2005) introduced a general class of statistical models called "additive models for location, scale and shape" (GAMLSS). It encompasses both parametric and semiparametric models, and includes a wide range of continuous and discrete distributions for the response variable. It also allows the simultaneous modeling of several parameters that index the response distribution using parametric and/or nonparametric functions. With GAMLSS models, the distribution of the response variable is not restricted to the exponential family and different additive terms can be included in the regression predictors for the parameters that index the distribution, like smoothing splines and random effects, which yields extra flexibility to the model. The model is parametric in the sense that the specification of a distribution for the response variable is required and at the same time it is semiparametric because one can model some conditioning effects through nonparametric functions.

The probability density function of the response variable y is denoted as $f(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\top}$ is a p-dimensional parameter vector. It is assumed to belong to a wide

---

[1]Evaluation of a set of real properties through methodology and procedures common to all of them.

class of distributions that we denote by $\mathcal{D}$. This class of distributions includes continuous and discrete distributions as well as truncated, censored and finite mixtures of distributions. In the GAMLSS regression framework the p parameters that index $f(y|\boldsymbol{\theta})$ are modeled using additive terms.

Let $\mathrm{y} = (y_1, \ldots, y_n)^\top$ be the vector of independent observations on the response variable, each y_i having probability density function $f(y_i|\boldsymbol{\theta^i})$, for $i = 1, \ldots, n$. Here, $\boldsymbol{\theta^i} = (\theta_{i1}, \ldots, \theta_{ip})^\top$ is a vector of p parameters associated with the explanatory variables and with random effects. When the covariates are stochastic, $f(y_i|\boldsymbol{\theta^i})$ is taken to be conditional on their values. Additionally, for $k = 1, \ldots, p$, $g_k(\cdot)$ is a strictly monotonic link function that relates the kth parameter $\boldsymbol{\theta}_k$ to explanatory variables and random effects through the additive predictor

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \boldsymbol{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \boldsymbol{Z}_{jk} \boldsymbol{\gamma}_{jk}, \tag{1}$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are $n \times 1$ vectors, $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{J_k' k})^\top$ is a vector of parameters of length J_k' and $\boldsymbol{X}_k$ and $\boldsymbol{Z}_{jk}$ are fixed (covariate) design matrixes of orders $n \times J_k'$ and $n \times q_{jk}$, respectively. Finally, $\boldsymbol{\gamma}_{jk}$ is a q_{jk}-dimensional random variable. Model given in Equation (1) is called GAMLSS; see Rigby and Stasinopoulos (2005).

In many practical situations it suffices to model four parameters ($p = 4$), usually location ($\theta_1 = \mu$), scale ($\theta_2 = \sigma$), skewness ($\theta_3 = \nu$) and kurtosis ($\theta_4 = \tau$); the latter two are said to be shape parameters. Thus, we have the models

$$\text{Location and scale parameters} \begin{cases} g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j1} \boldsymbol{\gamma}_{j1}; \\ \\ g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j2} \boldsymbol{\gamma}_{j2}; \end{cases}$$

$$\text{Shape parameters} \begin{cases} g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3 \boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j3} \boldsymbol{\gamma}_{j3}; \\ \\ g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4 \boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j4} \boldsymbol{\gamma}_{j4}. \end{cases}$$

It is also possible to add to the predictor functions h_{jk} that involve smoothers like cubic splines, penalized splines, fractional polynomials, loess curves, terms of variable coefficients, and others. Any combination of these functions can be included in the submodels for μ, σ, ν and τ. As Akantziliotou et al. (2002) pointed out, the GAMLSS framework can be applied to the parameters of any population distribution and generalized to allow the modeling of more than four parameters.

GAMLSS models can be estimated using the `gamlss` package for R (see Ihaka and Gentleman, 1996; Cribari-Neto and Zarkos, 1999), which is a free software; see `http://www.R-project.org`. Practitioners can then choose from more than 50 response distributions.

2.1 ESTIMATION

Two aspects are central to the GAMLSS additive components fitting, namely: the backfitting algorithm and the fact that quadratic penalties in the likelihood function follow from the assumption that all random effects in the linear predictor are normally distributed.

Suppose that the random effects $\boldsymbol{\gamma}_{jk}$ in model given in Equation (1) are independent and normally distributed with $\gamma_{jk} \sim \mathrm{N}_{q_{jk}}(\boldsymbol{0}, \boldsymbol{G}_{jk}^{-1})$, where $\boldsymbol{G}_{jk}^{-1}$ is the $q_{jk} \times q_{jk}$ (generalized)

inverse of the symmetric matrix $\boldsymbol{G}_{jk} = \boldsymbol{G}_{jk}(\boldsymbol{\lambda}_{jk})$. Rigby and Stasinopoulos (2005) noted that for fixed values of $\boldsymbol{\lambda}_{jk}$, one can estimate $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_{jk}$ by maximizing the penalized log-likelihood function

$$\ell_p = \ell - \frac{1}{2} \sum_{k=1}^{p} \sum_{j=1}^{J_k} \boldsymbol{\gamma}_{jk}^{\top} \boldsymbol{G}_{jk} \boldsymbol{\gamma}_{jk},$$

where $\ell = \sum_{i=1}^{n} \log\{f(y_i|\boldsymbol{\theta}^i)\}$ is the log-likelihood function of the data given $\boldsymbol{\theta}^i$, for $i = 1, \ldots, n$. This can be accomplished by using a backfitting algorithm; for details, see Rigby and Stasinopoulos (2005, 2007), Hastie and Tibshirani (1990) and Härdle et al. (2004).

2.2 Model selection and diagnostic

GAMLSS model selection is performed by comparing various competing models in which different combinations of the components $\mathcal{M} = \{\mathcal{D}, \mathcal{G}, \mathcal{T}, \boldsymbol{\lambda}\}$ are used, where $\mathcal{D}$ specifies the distribution of the response variable, $\mathcal{G}$ is the set of link functions $(g_1, \ldots, g_p)$ for the parameters $(\theta_1, \ldots, \theta_p)$, $\mathcal{T}$ defines the set of predictor terms $(t_1, \ldots, t_p)$ for the predictors $(\eta_1, \ldots, \eta_p)$ and $\boldsymbol{\lambda}$ specifies the set of hiperparameters.

In the parametric GAMLSS regression setting, each nested model $\mathcal{M}$ can be assessed from its fitted global deviance (GD), given by GD $= -2\ell(\hat{\boldsymbol{\theta}})$, where $\ell(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \ell(\hat{\boldsymbol{\theta}}^i)$. Two nested competing GAMLSS models $\mathcal{M}_0$ and $\mathcal{M}_1$, with fitted global deviances GD_0 and GD_1 and error degrees of freedom (df), namely df_{e0} and df_{e1}, respectively, can be compared using the generalized likelihood ratio test statistic $\Lambda = \text{GD}_0 - \text{GD}_1$, which is asymptotically distributed as χ^2 with $d = \text{df}_{e0} - \text{df}_{e1}$ df under $\mathcal{M}_0$. For each model $\mathcal{M}$ the number of error df, namely df_e is $\text{df}_e = n - \sum_{k=1}^{p} \text{df}_{\theta k}$, where $\text{df}_{\theta k}$ are the df used in the predictor of the model for the parameter θ_k, for $k = 1, \ldots, p$.

When comparing nonnested GAMLSS models (including models with smoothing terms), the generalized Akaike information criterion (GAIC) (see Akaike, 1983) can be used to penalize overfittings. This is achieved by adding to the fitted global deviances a fixed penalty $\#$ for each effective df that are used in the model, that is, $\text{GAIC}(\#) = \text{GD} + \#\text{df}$, where GD is the fitted global deviance. One then selects the model with the smallest $\text{GAIC}(\#)$ value.

To assess the overall adequacy of the fitted model, we propose the randomized quantile residual; see Dunn (1996). This is a randomized version of the residual proposed by Cox and Snell (1989) defined as

$$r_i^q = \Phi^{-1}(u_i), \quad i = 1, \ldots, n,$$

where $\Phi(\cdot)$ denotes the standard normal distribution function, u_i is a uniform random variable on the interval $(a_i, b_i]$, with $a_i = \lim_{y \uparrow y_i} F(y_i|\boldsymbol{\theta}^i)$ and $b_i = F(y_i|\boldsymbol{\theta}^i)$. A plot of these residuals against the index of the observations (i) should show no detectable pattern. A detectable trend in the plot of some residual against the predictors may be suggestive of link function misspecification.

Also, normal probability plots with simulated envelopes (see Atkinson, 1985) or worm plots (see Buuren and Fredriks, 2001) are a helpful diagnostic tool. The worm plots are useful for analyzing the residuals in different regions (intervals) of the explanatory variable. If no explanatory variable is specified, the worm plot becomes a detrended normal QQ plot of the (normalized quantile) residuals. When all points lie inside the (dotted) confidence bands (the two elliptical curves) there is no evidence of model misspecification.

In the context of a fully parametric model GAMLSS we can use pseudo R^2 measures.

For example, $R_p^{2*} = 1 - \log \widehat{L} / \log \widehat{L}_0$ (see McFadden, 1974) and $R_{\mathrm{LR}}^2 = 1 - (\widehat{L}_0/\widehat{L})^{2/n}$ (see Cox and Snell, 1989, pp. 208–209), where $\widehat{L}_0$ and $\widehat{L}$ are the maximized likelihood functions of the null (intercept only) and fitted (unrestricted) models, respectively. The ratio of the likelihoods or log-likelihoods may be regarded as a measure of the improvement, over the model with $\boldsymbol{\theta^i}$ parameters achieved by the model under investigation.

Our proposal, however, is to compare the different models using the pseudo-R^2 given by the square of the sample correlation coefficient between the response and the fitted values. Notice that by doing so we can consider both fully parametric models and models that include nonparametric components. We can also compare the explanatory power of a GAMLSS model to those of GLM and CNLRM models. This is the pseudo-R^2 we use, which was introduced by Ferrari and Cribari-Neto (2004) in the context of beta regressions and it is a straightforward generalization of the R^2 measure used in linear regression analysis.

3. Data Description

The data contain 2,109 observations on empty urban land lots located in the city of Aracaju, capital of the state of Sergipe (SE), Brazil, and comes from two sources: (i) data collected by the authors from real estate agencies, advertisements on newspapers and research on location (land lots for sale or already sold); (ii) data obtained from the "Departamento de Cadastro Imobiliário da Prefeitura de Aracaju". Observations cover the years 2005, 2006 and 2007. Each land lot price was recorded only once during that period. It is also noteworthy that the land lots in the sample are geographically referenced relative to the South American Datum[1] and have their geographical positions (latitude, longitude) projected onto the Universal Transverse Mercator (UTM) coordinate system[2].

The sample used to estimate the hedonic prices equation (i.e., the equation of hedonic prices of urban land lots in Aracaju-SE) contains, besides the year of reference, information on the physical (area, front, topography, pavement and block position), location (neighborhood, geographical coordinates, utilization coefficient and type of street in which the land lot is located) and economic (nature of the information that generated the observation, average income of the head of household of the censitary system, where the land is located and the land lot price) characteristics of the land lots. In particular, we use the variables

- YEAR (YR): qualitative ordinal variable that identifies the year in which the information was obtained. It assumes the values 2005, 2006 (YR06) and 2007 (YR07). It enters the model through dummy variables;
- AREA (AR): continuous quantitative variable, measured in m^2 (square meters), relative to the projection on a horizontal plane of the land surface;
- FRONT (FR): continuous quantitative variable, measured in m (meters), concerning the projection of the land lot front over a line which is perpendicular to one of the lot boundaries, when both are oblique in the same sense, or to the "chord", in the case of curved fronts;
- TOPOGRAPHY (TO): nominal qualitative variable that relates to the topographical conformations of the land lot. It is classified as "plain" when the land acclivity is smaller than 10% or its declivity is smaller than 5%, and as "rough" otherwise. It is a dummy variable that equals 1 for "plain" and 0 "rough";

---

[1]The South American Datum (SAD) is the regional geodesic system for South America and refers to the mathematical representation of the Earth surface at sea level.
[2]Cilindrical cartographic projection of the terrestrial spheroid in 60 secant cylinders at Earth level alongside the meridians in multiple zones of 6 degrees longitude and stretching out 80 degrees South latitude to 84 degrees North latitude.

- PAVEMENT (PA): nominal qualitative variable that indicates the presence or absence of pavement (concrete, asphalt, among others) on the street in which the main land lot front is located. It enters the model as a dummy variable that equals 1 when the land lot is located on a paved street and 0 otherwise;

- SITUATION (SI): nominal qualitative variable used to differentiate the disposition of the land lot on the block. It is classified as "corner lot" or "middle lot". It is a dummy variable that assumes value 1 for corner lots and 0 for all other land lots;

- NEIGHBORHOOD (NB): nominal qualitative variable referring to the name of the neighborhood where the land lot is located. It was categorized as valuable (highly priced) neighborhoods and other neighborhoods, with the variable shown as VN and regarded as a dummy (1 for valuable neighborhoods). The neighborhoods were also grouped as belonging or not belonging to the city South Zone, dummy denoted by SZ (1 for South Zone);

- LATITUDE (LAT) and LONGITUDE (LON): continuous quantitative variables corresponding to the geographical position of the land lot at the point $z = (\text{LAT}, \text{LON})$, where LAT and LON are the coordinates measured in UTM;

- UTILIZATION COEFFICIENT (UC): discrete variable given by a number that, when multiplied by the area of the land lot, yields the maximal area (in square meters) available for construction. UC is defined in an official urban development document. It assumes the values $3.0, 3.5, \ldots, 5.5, 6.0$;

- STREET (STR): ordinal qualitative variable used to differentiate the land lot location relative to streets and avenues. It is classified as "minor arterial" (STR1), "collector street" (STR2) and "local street" according to the importance of the street where the land lot is located. It enters the model as dummy variables;

- NATURE OF THE INFORMATION (NI): nominal qualitative variable that indicates whether the observation is derived from "offer", "transaction" or from the Aracaju register office (real state sale taxes). It enters the model through dummy variables;

- SECTOR (ST): discrete quantitative proxy variable of macrolocation used to socioeconomically distinguish the various neighborhoods, represented by the average income of the head of household, in minimum wages, according to the IBGE census (2000). The neighborhood average income functions as a proxy to other characteristics, such as urban amenities. It assumes the values: $1, \ldots, 18$;

- FRONT IN HIGHLY VALUED NEIGHBORHOODS (FRVN): continuous quantitative variable that assumes strictly positive values and corresponds to the interaction between FR and VN variables. It is included in the model to capture the influence of land lots front dimensions in "valuable" neighborhoods;

- UNIT PRICE (UP): continuous quantitative variable that assumes strictly positive values and corresponds to the land lot price divided by its area, measured in R\$$/\text{m}^2$ (reais per square meter).

In real estate appraisals (and specifically in land lots valuations), the interest typically lies in modeling the unit price as a function of the underlying structural, location and economic characteristics of the real estate. We then use UP as the dependent variable (response). The independent variables relate to the location (NB, VN, SZ, LAT, LON, ST, UC and STR), physical (AR, FR, TO, SI and FRVN) and economic (NI) land lot characteristics; we also account for the year of data collection.

Figure 1 presents box-plots of UP, AR and FR and Table 1 displays summary statistics on those variables. The box-plot of UP shows that its distribution is skewed and that there are several extreme observations. Notice from Table 1 that the sample values of UP range from R\$ $2.36/\text{m}^2$ to R\$ $800.00/\text{m}^2$ and that 75% of the land lots have unit prices smaller than R\$ $82.82/\text{m}^2$.

We note that 263 extreme observations have been identified from the box-plot of AR (see Figure 1). These observations are not in error, they appear as outlying data points in the plot because the variable assumes a quite wide range of values: from $41\,\mathrm{m}^2$ to $91,780\,\mathrm{m}^2$, that is, the largest land lot is nearly two thousand times larger than the smallest one.
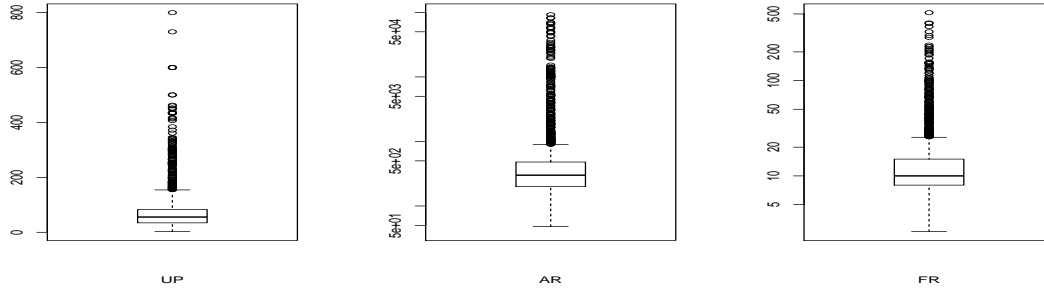

Figure 1. Box-plots of UP, AR and FR.

Table 1. Descriptive statistics.

Variable	Mean	Median	Standard error	Minimum	Maximum	Range
UP	72.82	55.56	70.28	2.36	800.00	797.64
LAT	710100.00	710300.00	2722.34	701500.00	714600.00	13100.00
LON	8787000.00	8786000.00	6638.77	8769000.00	8798000.00	29000.00
AR	1355.00	300.00	6063.53	48.00	91780.00	91732.00
FR	18.13	10.00	30.54	2.60	516.00	513.40

In order to investigate how UP relates to some explanatory variables, we produced dispersion plots. Figure 2 contains the pairwise plots: (i) UP × LAT; (ii) UP × LON; (iii) log(UP) × log(AR); (iv) log(UP) × log(FR); (v) UP × ST and (vi) UP × UC. It shows that there is a direct relationship between UP and the corresponding regressor in (i), (ii), (v) and (vi), whereas in (iii) and (iv) the relationship is inverse. Thus, there is a tendency for the land lot unit price to increase with latitude, longitude, sector and also with the utilization coefficient, and to decrease as the area and the front size increase. We note that the inverse relationship between unit price and front size was not expected. It motivated the inclusion of the covariate FRVN in our analysis.

It is not clear from Figure 2 whether the usual assumptions of normality and homoskedasticity are reasonable. As noted by Rigby and Stasinopoulos (2007), transformations of the response variable and/or of the explanatory variables are usually made in order to minimize deviations from the underlying assumptions. However, this practice may not deliver the expected results. Additionally, the resulting model parameters are not typically easily interpretable in terms of the untransformed variables. A more general modeling strategy is thus called for.

4. Empirical Modeling

In what follows, we estimate the hedonic price function of land lots located in Aracaju using the highly flexible class of GAMLSS models. At the outset, however, we estimate standard linear regression and generalized linear models. We use these fits as benchmarks for our estimated GAMLSS hedonic price function.
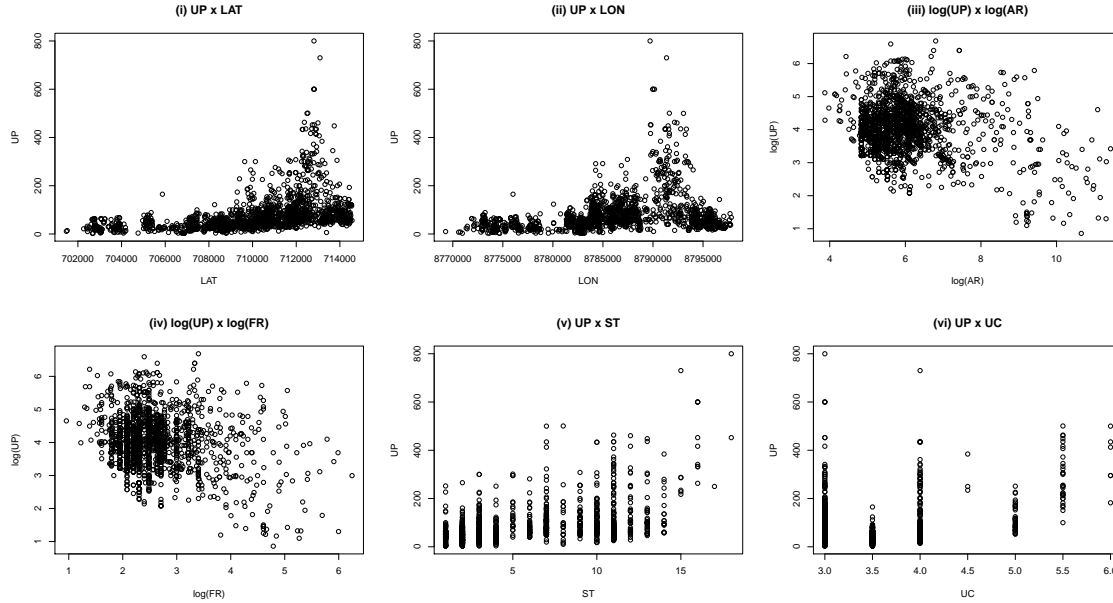
Figure 2. Dispersion plots.

4.1 Data modeling based on the CNLRM

Table 2 lists the classical normal linear regressions that were estimated. The transformation parameter of the Box-Cox model was estimated by maximizing the profile log-likelihood function: $\hat{\lambda} = 0.1010$. All four models are heteroskedastic and there is strong evidence of nonnormality for the first two models. The coefficients of determination range from 0.54 to 0.66. Since the error variances are not constant, we present in Table 3 the estimated parameters of Model E, which yields the best fit, along with heteroskedasticity-robust HC3 standard errors; see Davidson and MacKinnon (1993). Notice that all covariates are statistically significant at the 5% nominal level, except for LAT (p-value = 0.1263), which suggests that pricing differentiation mostly takes place as we move in the North-South direction.

4.2 Hedonic GLM function

Table 4 displays the maximum likelihood fit of the generalized linear model that we call Model A given by

$$g(UP^*) = \beta_0 + \beta_2\text{LON} + \beta_3\log(\text{AR}) + \beta_4\text{UC} + \beta_5\log(\text{ST}) + \beta_6\text{STR1} + \beta_7\text{STR2}$$
$$+\beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{YR06} + \beta_{14}\text{YR07}$$
$$+\beta_{15}\text{SZ} + \beta_{16}\log(\text{FRVN}), \tag{2}$$

where $UP^* = \mathbb{E}[UP] = \mu$, $UP \sim \text{Gamma}(\mu, \sigma)$ and $\eta = \log(\mu)$. We try a number of different models, and this one (gamma response and log link) yielded the best fit. We also note that all regressors are statistically significant at the 1% nominal level, except for LAT (p-value = 0.5295), which is why we dropped this covariate from the model.

Table 2. Fitted models via CNLRM.

Model	Equation	Considerations
B	$\text{UP} = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LON} + \beta_3\text{AR} + \beta_4\text{UC} + \beta_5\text{ST} + \beta_6\text{STR1} + \beta_7\text{STR2} + \beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{YR06} + \beta_{14}\text{YR07} + \beta_{15}\text{SZ} + \beta_{16}\text{FRVN} + \epsilon$	The null hypotheses that the errors are homoskedastic and normal are rejected at the 1% nominal level by the Breusch-Pagan and Jarque-Bera tests, respectively. The explanatory variables proved to be statistically significant at the 1% nominal level (z-tests). Also, $\overline{R}^2 = 0.539$, AIC = 22304 and BIC = 22406.
C	$\log(\text{UP}) = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LON} + \beta_3\text{AR} + \beta_4\text{UC} + \beta_5\text{ST} + \beta_6\text{STR1} + \beta_7\text{STR2} + \beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{YR06} + \beta_{14}\text{YR07} + \beta_{15}\text{SZ} + \beta_{16}\text{FRVN} + \epsilon$	The null hypotheses that the errors are homoskedastic and normal are rejected at the 1% nominal level by the Breusch-Pagan and Jarque-Bera tests, respectively. All explanatory variables proved to be statistically significant at 1% the nominal level (z-tests). Also, $\overline{R}^2 = 0.599$, AIC = 2912 and BIC = 3014.
D	$\log(\text{UP}) = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LON} + \beta_3\log(\text{AR}) + \beta_4\text{UC} + \beta_5\log(\text{ST}) + \beta_6\text{STR1} + \beta_7\text{STR2} + \beta_8\text{SI} + \beta_9\text{PA} + \beta_{10}\text{TO} + \beta_{11}\text{NIO} + \beta_{12}\text{NIT} + \beta_{13}\text{YR06} + \beta_{14}\text{YR07} + \beta_{15}\text{SZ} + \beta_{16}\log(\text{FRVN}) + \epsilon$	The Jarque-Bera test does not reject the null hypothesis of normality at the usual nominal levels, but the Breusch-Pagan test rejects the null hypothesis of homoskedasticity at the 1% nominal level. All explanatory variables are statistically significant at the 1% nominal level, except for the LAT variable (p-value = 0.0190). Also, $\overline{R}^2 = 0.651$, AIC = 2619 and BIC = 2721.
E	$\frac{\text{UP}^\lambda - 1}{\lambda} = \beta_0 + \beta_1\text{LAT} + \beta_2\text{LON} + \beta_3\log(\text{AR}) + \beta_4\text{UC} + \beta_5\log(\text{ST}) + \beta_6\text{STR1} + \beta_7\text{STR2} + \beta_8\text{PA} + \beta_9\text{TO} + \beta_{10}\text{NIO} + \beta_{11}\text{NIT} + \beta_{12}\text{YR06} + \beta_{13}\text{YR07} + \beta_{14}\log(\text{FRVN}) + \epsilon$	Normality is not rejected by the Jarque-Bera test, but the Breusch-Pagan test rejects the null hypothesis of homoskedasticity at the 1% nominal level. All covariates proved to be statistically significant at the 1% nominal level, except for the LAT variable (p-value = 0.0881). Also, $\overline{R}^2 = 0.657$, AIC = 4290 and BIC = 4392.

Table 3. Hedonic price function estimated via CNLRM – Model E.

	Estimate	Standard error	z-statistic	p-value
(Intercept)	-162.6307	34.1920	-4.756	0.0000
LAT	1.85e-05	1.21e-05	1.529	0.1263
LON	1.74e-05	4.60e-06	3.798	0.0001
log(AR)	-0.3507	0.0192	-18.236	0.0000
log(ST)	0.4423	0.0332	13.297	0.0000
UC	0.2651	0.0412	6.429	0.0000
STR1	0.4874	0.0717	6.789	0.0000
STR2	0.1678	0.0675	2.485	0.0130
SI	0.1119	0.0405	2.757	0.0058
PA	0.3853	0.0302	12.767	0.0000
TO	0.4905	0.0798	6.145	0.0000
NIO	0.5994	0.0592	10.131	0.0000
NIT	0.5111	0.0131	3.886	0.0000
YR06	0.2560	0.0351	7.289	0.0000
YR07	0.6450	0.0345	18.645	0.0000
SZ	0.7221	0.0474	15.239	0.0000
log(FRVN)	1.2041	0.0137	8.797	0.0000

4.3 GAMLSS HEDONIC FIT

4.3.1 LOCATION PARAMETER MODELING (μ)

Since UP (the response) only assumes positive values, we consider the distributions lognormal (LOGNO), inverse Gaussian (IG), Weibull (WEI) and gamma (GA). As noted earlier, we use pseudo-R^2 given by

$$\text{pseudo-}R^2 = [\text{correlation (observed values of UP, predicted values of UP)}]^2, \qquad (3)$$

to measure the overall goodness-of-fit.

Table 4. Hedonic price function estimated via GLM – Model A given in Equation (2).

	Estimate	Standard error	z-statistic	p-value
(Intercept)	−151.8019	15.7792	−9.620	0.0000
LON	1.77e-05	1.80e-06	9.851	0.0000
log(AR)	−0.2276	0.0108	−21.120	0.0000
UC	0.1272	0.0231	5.515	0.0000
log(ST)	0.2880	0.0193	14.954	0.0000
STR1	0.3562	0.0395	9.021	0.0000
STR2	0.1419	0.0408	3.482	0.0005
SI	0.0945	0.0255	3.707	0.0002
PA	0.2324	0.0220	10.556	0.0000
TO	0.3139	0.0503	6.236	0.0000
NIO	0.4208	0.0348	12.087	0.0000
NIT	0.3779	0.0642	5.884	0.0000
YR06	0.1947	0.0242	8.035	0.0000
YR07	0.4551	0.0242	18.780	0.0000
SZ	0.4716	0.0310	15.220	0.0000
log(FRVN)	0.7467	0.0622	11.997	0.0000

Table 5. Fitted models via GAMLSS.

Model	$\mathcal{D}$	$\mathcal{G}$	Equation	Considerations
E	LOGNO	logarithmic	$UP = \beta_0 + cs(LAT) + cs(LON) + cs(\log(AR)) + cs(UC) + cs(ST) + \beta_1 STR1 + \beta_2 STR2 + \beta_3 SI + \beta_4 PA + \beta_5 TO + \beta_6 NIO + \beta_7 NIT + \beta_8 YR06 + \beta_9 YR07 + \beta_{10} SZ + cs(\log(FRVN))$	All regressors are significant at the level 1% significance level (z-tests). Also, AIC = 19155, BIC = 19359 and GD = 19083. Pseudo-R^2= 0.739.
F	IG	logarithmic	$UP = \beta_0 + cs(LAT) + cs(LON) + cs(\log(AR)) + cs(UC) + cs(ST) + \beta_1 STR1 + \beta_2 STR2 + \beta_3 SI + \beta_4 PA + \beta_5 TO + \beta_6 NIO + \beta_7 NIT + \beta_8 YR06 + \beta_9 YR07 + \beta_{10} SZ + cs(\log(FRVN))$	All regressors are significant at the 1% significance level (z-test). Also, AIC = 19845, BIC = 20048 and GD = 19773. Pseudo-R^2= 0.678.
G	WEI	logarithmic	$UP = \beta_0 + cs(LAT) + cs(LON) + cs(\log(AR)) + cs(UC) + cs(ST) + \beta_1 STR1 + \beta_2 STR2 + \beta_3 SI + \beta_4 PA + \beta_5 TO + \beta_6 NIO + \beta_7 NIT + \beta_8 YR06 + \beta_9 YR07 + \beta_{10} SZ + cs(\log(FRVN))$	All regressors proved to be significant at the 1% significance level (z-tests). Also, AIC = 19260, BIC = 19463 and GD = 19188. Pseudo-R^2= 0.748.
H	GA	logarithmic	$UP = \beta_0 + cs(LAT) + cs(LON) + cs(\log(AR)) + cs(UC) + cs(ST) + \beta_1 STR1 + \beta_2 STR2 + \beta_3 SI + \beta_4 PA + \beta_5 TO + \beta_6 NIO + \beta_7 NIT + \beta_8 YR06 + \beta_9 YR07 + \beta_{10} SZ + cs(\log(FRVN))$	All regressors are significant at the 1% significance level (z-tests). Also, AIC = 19062, BIC = 19337 and GD = 19134. Pseudo-R^2= 0.746.

Table 6. Hedonic price function estimated via GAMLSS – Model I.

	Estimative	Standard error	z-statistic	p-value
(Intercept)	−165.4000	16.1300	−10.251	0.0000
cs(LAT)	5.17e-05	6.22e-06	8.307	0.0000
cs(LON)	1.51e-05	2.13e-06	7.071	0.0000
cs(log(AR))	−0.2317	0.0096	−24.074	0.0000
cs(ST)	0.0465	0.0037	12.416	0.0000
cs(UC)	0.1223	0.0206	5.947	0.0000
STR1	0.3133	0.0349	8.963	0.0000
STR2	0.0926	0.0364	2.545	0.0100
SI	0.0920	0.0227	4.054	0.0000
PA	0.1891	0.0195	9.670	0.0000
TO	0.2662	0.0474	5.951	0.0000
NIO	0.4135	0.0395	13.362	0.0000
NIT	0.3485	0.0571	6.102	0.0000
YR06	0.1645	0.0215	7.632	0.0000
YR07	0.4358	0.0215	20.235	0.0000
cs(log(FRVN))	0.6513	0.0569	11.443	0.0000
SZ	0.3875	0.0299	12.935	0.0000

The models listed in Table 5 include smoothing cubic splines (cs) with 3 effective df for the covariates LAT, LON, log(AR), UC, ST and log(FRVN). Other smoothers (such as loess and penalized splines), as well as different combinations of $\mathcal{D}$ (see Rigby and Stasinopoulos, 2007), such as BCPE, BCCG, LNO, BCT, exGAUSS, among others, and $\mathcal{G}$, such as identity, inverse, reciprocal, among others, were considered. However, they did not yield superior fits. We also note that Model I yields the smallest values of the three model selection criteria. Table 6 contains the a summary of the model fit.

The use of three effective df in the smoothing functions delivered a good model fit. However, in order to determine whether a different number of effective df delivers superior fit, we used two criteria, namely: the AIC (objective) and visual inspection of the smoothed curves (subjective); visual inspection aimed at avoiding overfitting. We then arrived at Model J. It also uses cubic spline smoothing (cs), but with a different number of effective df in the smoothing functions; see Table 7. Notice that there was a considerable reduction – relative to Model I– in the AIC, BIC and GD values (18822, 19212 and 18684, respectively) and that there is a better agreement between observed and predicted response values.

Table 7. Hedonic price function estimated via GAMLSS – Model J.

	Estimative	Standard error	z-statistic	p-value
(Intercept)	-130.1000	14.8100	-8.787	0.0000
cs(LAT, df = 10)	5.92e-05	5.71e-06	10.354	0.0000
cs(LON, df = 10)	1.05e-05	1.96e-06	5.352	0.0000
cs(log(AR), df = 10)	-0.2559	8.83e-03	-28.963	0.0000
cs(ST, df = 8)	0.0373	3.44e-03	10.831	0.0000
cs(UC, df = 3)	0.1769	0.0188	9.370	0.0000
STR1	0.2571	0.0320	8.012	0.0000
STR2	0.0728	0.0334	2.180	0.0293
SI	0.1029	0.0208	4.940	0.0000
PA	0.1436	0.0179	7.999	0.0000
TO	0.1822	0.0410	4.436	0.0000
NIO	0.4173	0.0284	14.690	0.0000
NIT	0.3388	0.0524	6.462	0.0000
YR06	0.1373	0.0198	6.941	0.0000
YR07	0.4190	0.0197	21.190	0.0000
cs(log(FRVN), df = 10)	0.6599	0.0522	12.630	0.0000
SZ	0.5119	0.0275	18.613	0.0000

Figure 3 contains plots of the smoothed curves from Model J. The dashed lines are confidence bands based on pointwise standard errors. Panels (I), (II), (III), (IV), (V) and (VI) reveal that the effects/impacts of LAT, LON, log(AR), ST, UC and log(FRVN) are typically increasing, increasing/decreasing, decreasing, increasing, increasing and increasing, respectively, with increases in latitude, longitude, log area, socioeconomic indicator, utilization coefficient and log land front in highly priced neighborhoods. (Panel (II) alternately shows local increasing and decreasing trends.) Some of these effects were also suggested by the estimated coefficients of the CNLRM and GLM models. Here, however, one obtains a somewhat more flexible global picture, as we see.

In Panel (I), one notices that as the latitude increases the "contribution" of the LAT covariate between the 702000 and 709000 latitudes (approximately) –neighborhoods that belong to the expansion zone of the city– is negative, whereas starting from position 709000 (approximately) –South Zone and downtown area– the price effect is positive. Additionally, we note that, in certain ranges, increases in latitude lead to drastic changes in the slope of the smoothed curve, e.g., between the 708000 and 710000 positions, whereas in other areas, for instance between the 706000 and 708000 latitudes –the Mosqueiro neighborhood–, an increase in latitude leads to an uniform negative effect.

Panel (II) shows that as longitude increases to position 8780000 the "contribution" of the LON covariate is positive and nearly uniform, which almost exclusively covers observations relative to the Mosqueiro neighborhood. Starting at the 8785000 position there is a remarkable change in the slope of the fitted curve, which is triggered by the location of the most upper class neighborhoods: from 8785000 to 8794000. After the 8794000 position, the effect remains positive, but is decreasing and, eventually, it becomes negative.

We see in Panel (III) that as the area (in logs) increases the "contribution" of the log(AR) covariate, for land lots with log areas between 4 and 5 (respectively), is clearly positive.
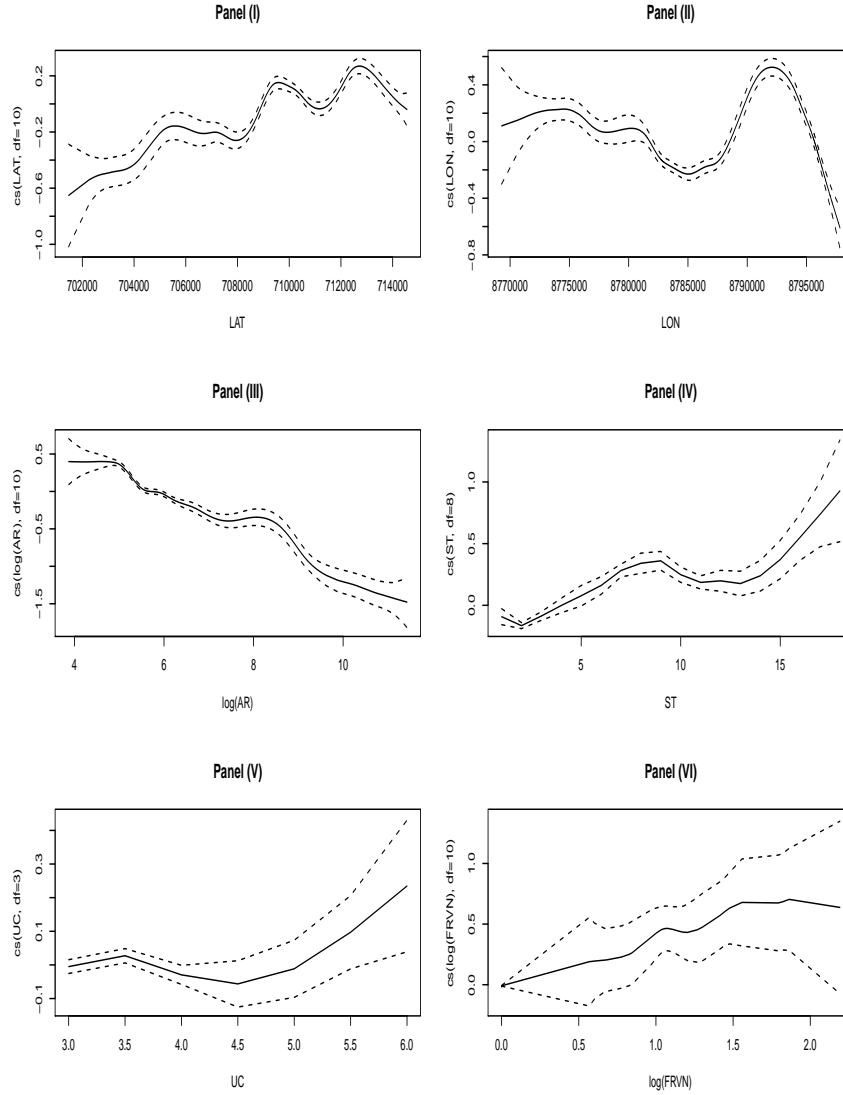
Figure 3. Smoothed additive terms – Model J.

The effect is negative for land lots with log areas in excess of 5.

In Panel (IV), it is possible to notice that as we move up in the socioeconomic scale the "contribution" of the ST covariate, in the range from 1 to 4 minimum wages, is negative, even though the there is an increasing trend. For land lots located in neighborhoods that correspond to more than 4 minimum wages, the effect is always positive; from 10 to 15 minimum wages the effect is uniform.

We note from Panel (V) that, contrary to what one would expect, the "contribution" of the UC covariate is not positive. In the range from 3.0 to 5.0, the fitted curve displays small oscillations, alternating in the positive and negative regions. The positive effect only holds for utilization coefficients greater than 5.0.

Notice from Panel (VI) that as the front land lot (in logs) increases in highly priced neighborhoods the "contribution" of the log(FRVN) covariate is mostly increasing and positive. However, in the 1.5 to 2.0 interval the positive effect is approximately uniform.

4.4 COMPARING MODELS

In order to compare the best estimated models via CNLRM (Model E), GLM (Model A given in Equation (2)) and GAMLSS (Model J) we use the AIC and BIC. Note the criteria only is used to compare models that use the response (UP) in the same measurement scale, i.e., Models A and J. We also compare the different models using the pseudo-R^2 given in Equation (3).

We present in Table 8 a comparative summary of the three models. We note that Model J is superior to the two competing models. Not only it has the smallest AIC and BIC values (in comparison to Model A, but it also has a much larger pseudo-R^2. The GAMLSS pseudo-R^2 exceeds 0.80, which is notable.

Table 8. Comparative summary of the CNLRM, GLM and GAMLSS estimated models.

Model	Class	AIC	BIC	Pseudo-R^2
E	(CNLRM)	4290	4392	0.667
A	(GLM)	19486	19581	0.672
J	(GAMLSS)	18822	19212	0.811

4.5 DISPERSION PARAMETER MODELING (σ)

After a suitable model for the prediction of μ was selected, we carried out a likelihood ratio test to determine whether the GAMLSS scale parameter σ is constant for all observations. The null hypothesis that σ is constant was rejected at the usual nominal levels. We then built a regression model for such a parameter. To that end, we used stepwise covariate selection, considered different link functions (such as identity, inverse, reciprocal, etc.) and included smoothing functions (such as cubic splines, loess and penalized splines) in the linear predictor, just as we had done for the location parameter. We used the AIC for selecting the smoothers and for choosing the number of df of the smoothing functions together with visual inspection of the smoothed curves.

We present in Table 9 the GAMLSS hedonic price function parameter estimates obtained by jointly modeling the location (μ) and dispersion (σ) effects; Model K. The model uses the gamma distribution for the response and the log link function for both μ and σ. We note that Model K contains parametric and nonparametric terms, and for that reason it is said to be a linear additive semiparametric GAMLSS.

We note from Table 9 that the parameter estimates of the location submodel in Model K are similar to the corresponding estimates from Model J, in which σ was taken to be constant; see Table 7. It is noteworthy, nonetheless, that there was a sizeable reduction in the AIC, BIC and GD values (18607, 19065 and 18445, respectively) and also an improvement in the residuals as evidenced by the worm plot; see Figures 4 and 5.

Only two covariates were selected for the σ regression submodel in Model K, namely: ST and log(AR). The former (ST) entered the model in the usual parametric fashion whereas the latter (log(AR)) entered the model nonparametrically through a cubic spline smoothing function with ten effective df. We note that the positive sign of the log(AR) coefficient indicates that the UP dispersion is larger for land lots with larger areas whereas the negative sign of the ST coefficient indicates that the dispersion is inversely related to the socioeconomic neighborhood indicator.

It is noteworthy that the pseudo-R^2 of Model K is quite high (0.817) and that all of explanatory variables are statistically significant at the 1% nominal level which is not all that common in large sample cross sectional analyses, especially in real estate appraisals. Overall, the variable dispersion GAMLSS model is clearly superior to the alternative models. The good fit of Model K can be seen in Figure 6 where we plot the observed response values against the predicted values from the estimated model. Note that the 45º line in this plot indicates perfect agreement between predicted and observed values.

Table 9. Hedonic price function estimated via GAMLSS – Model K.

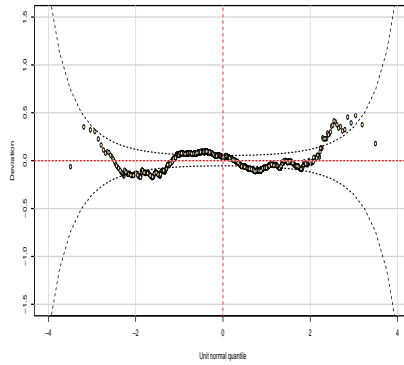|	Estimative	Standard error	z-statistic	p-value
		μ Coefficients		
(Intercept)	−95.1300	14.2700	−6.665	0.0000
cs(LAT, df = 10)	5.94e-05	5.37e-06	11.053	0.0000
cs(LON, df = 10)	6.45e-06	1.86e-06	3.460	0.0000
cs(log(AR), df = 10)	−0.2087	0.0104	−20.138	0.0000
cs(ST, df = 8)	0.0321	0.0030	10.666	0.0000
cs(UC, df = 3)	0.2095	0.0161	13.006	0.0000
STR1	0.2039	0.0298	6.838	0.0000
STR2	0.0729	0.0276	2.635	0.0084
SI	0.7136	0.0192	3.705	0.0000
PA	0.1653	0.0157	10.465	0.0000
TO	0.1778	0.0370	4.799	0.0000
NIO	0.3722	0.0251	14.799	0.0000
NIT	0.2790	0.0468	5.957	0.0000
YR06	0.1255	0.0175	7.144	0.0000
YR07	0.4195	0.0177	23.622	0.00
cs(log(FRVN), df = 10)	0.6809	0.0403	16.88	0.0000
SZ	0.4824	0.0241	20.001	0.0000
		σ coefficients		
(Intercept)	−1.6838	0.0839	−20.072	0.0000
cs(log(AR), df = 10)	0.1370	0.0143	9.593	0.0000
ST	−0.0391	0.0040	−9.632	0.0000

Figure 4. Worm plot – Model J.

Model K is given by

$$\log(\mu) = \beta_0 + \text{cs(LAT, df} = 10) + \text{cs(LON, df} = 10) + \text{cs(log(AR), df} = 10) +$$
$$\text{cs(UC, df} = 3) + \text{cs(ST, df} = 8) + \beta_1 \text{STR1} + \beta_2 \text{STR2} + \beta_3 \text{SI} +$$
$$\beta_4 \text{PA} + \beta_5 \text{TO} + \beta_6 \text{NIO} + \beta_7 \text{NIT} + \beta_8 \text{YR06} + \beta_9 \text{YR07} + \beta_{10} \text{SZ} +$$
$$\text{cs(log(FRVN), df} = 10),$$
$$\log(\sigma) = \gamma_0 + \gamma_1 \text{ST} + \text{cs(log(AR), df} = 10),$$

in which the response (UP) follows a GA distribution with location and scale parameters μ and σ, respectively. This model proved to be the best model for hedonic prices equation estimation of urban land lots in Aracaju.
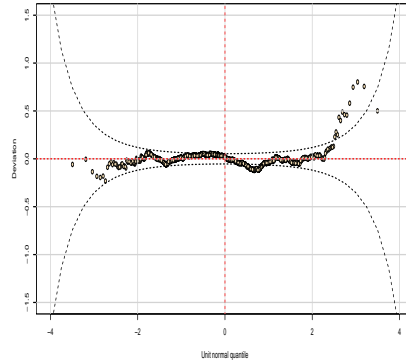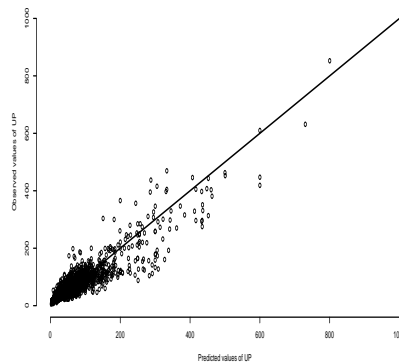
Figure 5. Worm plot – Model K.

Figure 6. Observed values × predicted values of UP – Model K).

5. Concluding Remarks

Real state appraisal is usually performed using the standard linear regression model or the class of generalized linear models. In this paper, we introduced real state appraisal based on the class of generalized additive models for location, scale and shape, GAMLSS. Such a class of regression models provides a flexible framework for the estimation of hedonic price functions. It even allows for some conditioning variables to enter the model in a nonparametric fashion. The model also accommodates variable dispersion and can be based on a wide range of response distributions. Our empirical analysis was carried out using a large sample of land lots located in the city of Aracaju (Brazil). The selected GAMLSS model displayed a very high pseudo-R^2 (approximately 0.82) and yielded an excellent fit. Moreover, the inclusion of nonparametric additive terms in the model allowed for the estimation of the hedonic price function in a very flexible way. We showed that the GAMLSS fit was clearly superior to those based on the standard linear regression and on a generalized linear model. We strongly recommend the use of GAMLSS models for real state appraisal.

REFERENCES

Akaike, H., 1983. Information measures and model selection. Bulletin of the International Statistical Institute, 50, 277–290.

Akantziliotou, C., Rigby, R.A., Stasinopoulos, D.M., 2002. The R implementation of generalized additive models for location scale and shape. In Stasinopoulos, M., Touloumi, G., (eds.). Statistical Modelling in Society: Proceedings of the 17th International Workshop on Statistical Modelling. Chania, Greece, pp. 75–83.

Anglin, P., Gencay, R., 1996. Semiparametric estimation of hedonic price functions. Journal of Applied Econometrics, 11, 633–648.

Atkinson, A.C. (1985) Plots, Transformations and Regression. Oxford University Press, New York.

Buuren, S., Fredriks, A.M., 2001. Worm plot: a simple diagnostic device for modeling growth reference curves. Statistics in Medicine, 20, 1259–1277.

Clapp, J.M., Kim, H.J., Gelfand, A., 2002. Predicting spatial patterns of house prices using LPR and Bayesian smoothing. Real Estate Economics, 30, 505–532.

Cox, D.R., Snell, E.J., 1989. Analysis of Binary Data. Chapman and Hall, London.

Cribari-Neto, F., Zarkos, S.G., 1999. R: yet another econometric programming environment. Journal of Applied Econometrics, 14, 319–329.

Davidson, R., MacKinnon, J.G., 1993. Estimation and Inference in Econometrics. Oxford University Press, New York.

Dunn, P.K., Smyth, G.K. (1996). Randomised quantile residuals. *Journal of Computational and Graphical Statistics.* 5, 236-244.

Eubank, R., 1999. Nonparametric Regression and Spline Smoothing. Second edition. Marcel Dekker, New York.

Ferrari, S.L.P., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. Journal of Applied Statistics, 31, 799–815.

Gencay, R., Yang, X., 1996. A forecast comparison of residential housing prices by parametric and semiparametric conditional mean estimators. Economic Letters, 52, 129–135.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semiparametric Models. Springer-Verlag, Berlin.

Hastie, T.J., Tibshirani, R.J., 1990. Generalizeds Additive Models. Chapman and Hall, London.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5, 299–314.

Iwata, S., Murao, H., Wang, Q., 2000. Nonparametric assessment of the effects of neighborhood land uses on the residential house values. In Fomby, T., Carter, H.R., (eds.). Advances in Econometrics: Applying Kernel and Nonparametric Estimation to Economic Topics. JAI Press, New York.

Martins-Filho, C., Bin, O., 2005. Estimation of hedonic price functions via additive nonparametric regression. Empirical Economics, 30, 93–114.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., (ed.). Frontiers in Econometrics. Academic Press, New York, pp. 105–142.

Pace, R.K., 1993. Non-parametric methods with applications to hedonic models. Journal of Real Estate Finance and Economics, 7, 185–204.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape (with discussion). Applied of Statistics, 54, 507–554.

Rigby, R.A., Stasinopoulos, D.M., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, 23, 1–46.

Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation perfect competition. Journal of Political Economy, 82, 34–55.

Silverman, B.W., 1984. Spline smoothing: the equivalent kernel method. The Annals of
    Statistics, 12, 896–916.
Thorsnes, P., McMillen, D.P., 1998. Land value and parcel size: a semiparametric analysis.
    Journal of Real Estate Finance and Economics, 17, 233–244.