

# Journal of Engineering Research

## MASS EVALUATION WITH GENERALIZED ADDITIVE MODEL AND GEOSTATISTICS

---

*Antônio Augusto Ferreira de Oliveira*  
Fortaleza Municipal Finance Secretariat

*Terezinha Ferreira de Oliveira*  
``Universidade Federal do Pará`` – UFPA

*Sandro Ricardo Vasconcelos Bandeira*  
Fortaleza Municipal Finance Secretaria

All content in this magazine is licensed under a Creative Commons Attribution License. Attribution-Non-Commercial-Non-Derivatives 4.0 International (CC BY-NC-ND 4.0).



**Abstract:** This work presents a methodology for mass evaluation using Generalized Additive Models and geostatistics for urban land data in the city of Fortaleza. The Generalized Additive Model confirmed some hypotheses regarding the behavior of real estate market prices, such as an increase in average unit value when the lot is within a gated community, with a water supply available and when located on a paved road. With this model, the non-linearity of unit prices with the tested area, utilization index, distance to main roads, commercialization density, income and IPTU base value of the current PGV was also revealed. However, the interaction of the model's terrain coordinates was not sufficient to eliminate the spatial autocorrelation of the residues, so a spherical variogram was adjusted over them, with subsequent interpolation by ordinary kriging, obtaining a residue surface to be added to the final prediction. From this procedure, there was an improvement in all performance measures used in this work, mainly regarding the COD, which increased from 34.40% to 13.06% after interpolation, which fully meets the IAAO recommendations for evaluation models in large scale. It is concluded that the proposed methodology, using Generalized Additive Models and interpolation of residues with geostatistical techniques, is very promising in mass evaluation, but must be used very sparingly, as measurement errors in the collection of market data can bias the final prediction.

**Keywords:** Generalized additive models (GAMs); geostatistics; kriging, ordinary; mass evaluation.

## INTRODUCTION

The purpose of mass valuation for tax purposes is to determine the market value for a large amount of property in a region for real estate taxation. Therefore, the estimates generated by these must be precise to result in uniform and equitable assessments.

Traditional property valuation approaches are based on hedonic price models, which generally make use of multiple linear regression using ordinary least squares (OLS). There are many assumptions for using OLS models, and their violations in urban property valuations cause biases and inconsistencies in OLS estimators, which leads to poor specification of the functional form of the model. In the same vein, spatial autocorrelation between observations and/or model residues can have serious consequences for the understanding of the phenomenon (ANSELIN; REY, 2014; WOOLDRIDGE, 2016).

Furthermore, it is common in real estate market data to observe asymmetry to the right of observed prices and, therefore, the use of logarithmic transformation in them as a way of establishing supposedly linear relationships between the response variable and its predictors. However, these transformations are often inefficient as they do not effectively portray the underlying relationship.

An alternative to this problem would be the use of Generalized Additive Models (GAM) with a distribution that more closely resembles the functional form of real estate market data, such as the gamma distribution and logarithmic link function (VEIE; PANDURO, 2013) and subsequent treatment of the spatial autocorrelation of residues obtained with geostatistical techniques to propose more precise generic value plans (PGV) to be used in the taxation of urban property and territorial tax (IPTU).

In this context, this work is based on a sample composed of 1,924 urban land data

collected from the Fortaleza real estate market, as well as assessments and declarations for ITBI launch carried out at the Fortaleza Finance Secretariat (SEFIN) to apply the aforementioned methodology. The “land” typology was chosen for this study given its difficulty in modeling, either due to the lack of reliable data available, as well as because this typology is necessary in the taxation of each property in the municipal register, regardless of whether it is built or not, in use. of the evolutionary method, usually used in most Brazilian municipalities in IPTU (urban property and territorial tax) taxation.

## MATERIALS AND METHODS

### STUDY AREA AND DATA DESCRIPTION

The Municipality of Fortaleza is the capital of the state of Ceará, with an estimated population of 2.6 million people, according to IBGE projections for the year 2018, making it the fifth most populous city in Brazil. It has a territorial area of 314.93 km<sup>2</sup>, which makes it the densest among all capitals with 7,786.44 inhabitants/km<sup>2</sup>, according to the 2010 census.

According to the SEFIN municipal real estate registry (CIM), the total number of lots is approximately 385 thousand, of which around 74 thousand are unbuilt (empty land). The number of municipal registrations subject to IPTU (urban property and territorial tax) taxation is approximately 779 thousand. Although the number of territorial properties represents less than 10% of the total registrations, it must be noted that IPTU (urban property and territorial tax) taxation in Fortaleza is carried out using the evolutionary method, where the composition of the market value, the basis for calculating this tax, is given by the sum the value of the territorial part and the depreciated built part.

The 1,924 data for sample composition for this work were made available by SEFIN in the period from July 2019 to July 2020. The sample covers 111 neighborhoods in the municipality, out of a total of 121, and its spatial distribution can be visualized through Map 1:

The explained variable was the unit price of land in R\$/m<sup>2</sup>, representing the division of the observed price (R\$) divided by its area (m<sup>2</sup>). The explanatory variables used for modeling are found in Table 1:

### GENERALIZED ADDITIVE MODELS (GAM)

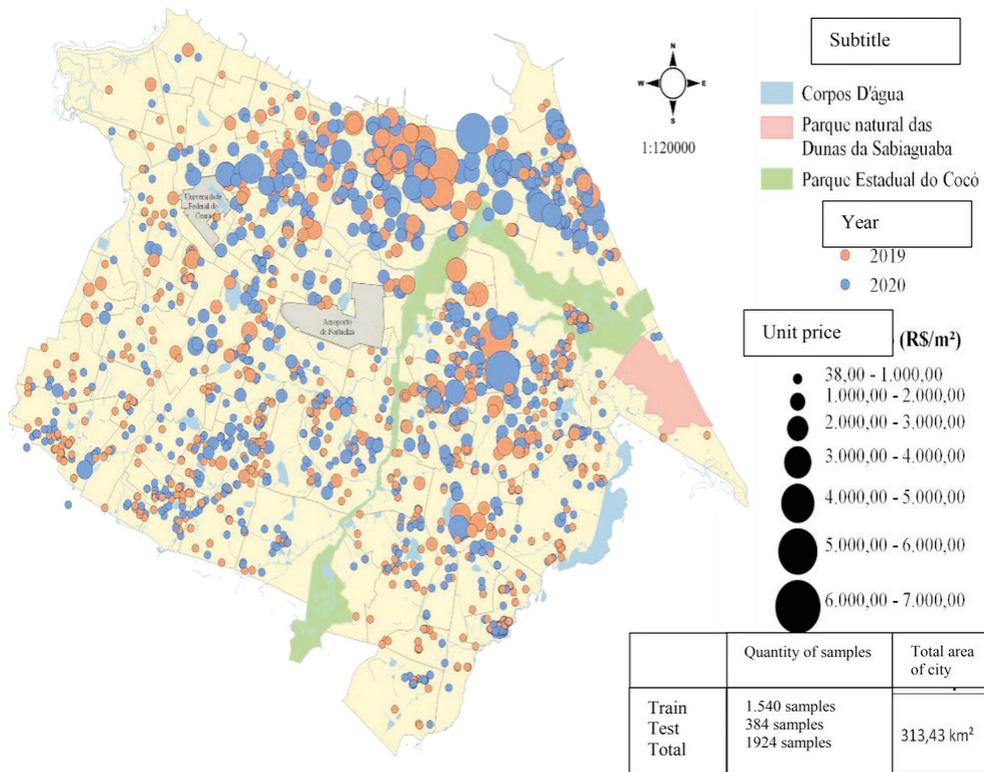
Generalized additive models (GAM) are semi-parametric extensions of generalized linear models (GLM) that use a link function to establish a relationship between the mean of the response variable and a “smoothing” function for the explanatory variables (HASTIE; TIBSHIRANI; FRIEDMAN, 2008). The conditional mean  $\mu(X)$  of the response  $Y$  is related to an additive function of the predictors through a link function  $g$  defined in equation (1):

$$g[\mu(X)] = E(Y|X_1, X_2, \dots, X_n) = \alpha + f_1(X_1) + f_2(X_2) + \dots + f_n(X_n) \quad (1)$$

where  $X_1, X_2, \dots, X_n$  represent the explanatory variables and  $Y$  the response variable. The  $f_j$  are non-parametric functions that can be fitted using a set of base splines (e.g. cubic smoothing or a smoother kernel). Formally, these smoothing functions are given by the sum of  $k$  base splines ( $b_h$  in equation 2) multiplied by their respective coefficients to be estimated ( $\beta_h$  in equation 2) (VEIE; PANDURO, 2013):

$$f = \sum_{h=1}^k \beta_h b_h \quad (2)$$

The additivity of this type of model facilitates interpretation and allows the establishment of non-linear and non-monotone relationships between the response variable and the set of



Map 1 - Spatial distribution of the sample.

Source: own elaboration.

Variables Description		Minimum	Average	Maximum
<i>V unit</i>	Unit value of land in R\$/m <sup>2</sup> .	38,40	865,06	7.000,00
<b>Structural variables</b>				
<i>numfr</i>	Number of fronts	1,00	1,43	9,00
<i>test</i>	Main test length in m.	3,00	28,21	880,00
<i>area</i>	Area of land in m <sup>2</sup> .	23,04	2.206,57	330.000,00
<b>Spatial and location variables</b>				
Variables Description		Minimum	Average	Maximum
<i>xcen</i>	Centroid X of the lot relative to the centroid of the municipality itself and divided by 1,000 (SIRGAS 2000 datum, UTM spindle 24S).	- 1,16	0,09	1,26
<i>ycen</i>	Centroid Y of the lot relative to the centroid of the municipality itself and divided by 1,000 (SIRGAS 2000 datum, UTM spindle 24S).	- 1,11	- 0,09	0,92
<i>distvp</i>	Distance in m to the nearest main road.	4,46	279,67	1.532,50
<i>dscm</i>	Marketing density of lots that are in the same	0,01	0,23	1,00
<i>vbt</i>	section of the street of the current lot. IPTU land reference value.	3,02	56,51	1.532,69
<i>income</i>	Income of the person responsible kriguada IBGE 2010.	0,40	3,46	18,89
<i>idhed</i>	IDH, education dimension by neighborhood.	0,65	0,80	0,95
<i>iaeq</i>	Equivalent utilization index.	0,01	1,66	3,00
<i>lotcnd</i>	0- outside the condominium; 1-in condominium.	0		1
<i>esq</i>	0- it's not a corner; 1-it's a corner.	0		1
<i>water</i>	0- no water; 1-with water.	0		1
<i>esg</i>	0- no sewage; 1-with sewage.	0		1

<i>pav</i>	0- no asphalt paving (or concrete) in the public area; 1-with asphalt paving (or concrete)	0	1
<b>Information variables</b>			
<i>a2020</i>	0- information from the year 2019; 1-information from the year 2020.	0	1
<i>Tr</i>	0- It is not a transaction; 1-Transaction.	0	1
<i>of</i>	0- It is not an offer; 1-Offer.	0	1

Table 1 – Description of the variables tested/used in the model.

Source: own elaboration.

explanatory variables (ibid). For continuous variables, smoothing functions were used, allowing the observation of their relationships with the unit price (penultimate term of the equation:

3), as it will be seen in section 3. As for the dichotomous variables (second term of equation 3), they remained in their original form, allowing their marginal effect to be captured directly.

It is also worth highlighting the use of the link function in logarithmic form for the GAM, as this is widely used in hedonic analyses. The incorporation of UTM coordinates in the model, in order to capture the effect of location on the behavior of observed prices, followed the methodology of Veie and Panduro (2013), that is, through the introduction of the smoothing function with the interaction of location coordinates space. Thus, the final GAM model had the following form:

$$\ln(vunit) = \beta_0 + \beta_i D_i + f_1(xcen, ycen, k) + f_j(X_j, k) + \varepsilon_i \quad (3)$$

, where  $D_i$  are the set of dummy explanatory variables, and  $X_j$  the set of explanatory (and continuous) variables that will be smoothed by the sum of a number:  $k$  of base splines.

It is also worth noting that the specification of a GAM model requires the choice of a distribution family of exponential functions (gamma, Gaussian, inverse Gaussian, poisson, etc.), with the gamma distribution being chosen because property prices are always

positive with variance increasing with the price (ibid). The “mgcv” package in R (WOOD, 2010) was used to optimize the smoothing level of the predictors (k value) based on the minimization of the Bayesian Information Criterion – BIC, as well as the Effective Degrees of Freedom (EDF). The method for estimating predictors and coefficients of the model’s fixed factors was the Restricted Maximum Likelihood (REML) (ibidem).

The GAM model was estimated on the 80% data and prediction was made on the remaining 20%. The performance measures for choosing the model and final prediction used in this work were those recommended by the standards of the International Association of Assessing Officers (IAAO, 2010), as well as other metrics already established in the literature, namely: the median assessment level, dispersion coefficient (COD), mean absolute percentage error (MAPE), square root of the mean squared error (RMSE) and coefficient of determination  $R^2$ .

Despite the use of location predictor variables, the residuals may be spatially autocorrelated, a situation in which the estimates generated by the GAM model were not sufficient to capture the entire spatial dependence of observed prices. Thus, this hypothesis [of spatial dependence] was tested by quantifying it using the global Moran’s I index (Moran’s I) described below.

## VERIFICATION OF SPATIAL AUTOCORRELATION OF RESIDUALS

It is known that location is one of the main attributes for explaining prices in the real estate market, giving rise to the famous expression location, location and location are everything in real estate. Therefore, it is common for prices and/or evaluation modeling errors to be spatially autocorrelated. Despite the introduction of location and neighborhood variables in the predictive models, this spatial dependence can be made through the global Moran I index (ARBI, 2014). The global Moran's I index can be understood as the angular coefficient of the linear regression line obtained by OLS of the independent variable of interest (in this case, the residues of the GAM model) with the dependent variable "lagged" by these same residues. The latter ( $Wy$ ) is obtained by the vector product of a weighted neighborhood matrix ( $W$ ) over these errors. In other words,  $Wy$  represents the weighted error of the neighbors on the observation in question.

Therefore, the global Moran I index is the  $\beta_1$  of equation 4 and its statistics are represented in equation 5.

$$W_y = \beta_0 + \beta_1 y + \varepsilon \quad (4)$$

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

where  $n$  represents the number of observations,  $y_i$  is the variable of interest and  $w_{ij}$  is the element of the spatial neighborhood matrix between the observations:  $i$  and  $j$ .

## NEIGHBORHOOD MATRIX

In this work, a neighborhood matrix was chosen as the inverse of the distance between the centroids of the land plots, as recommended in the literature (cf. DANTAS, 2003 and ALMEIDA, 2012)<sup>1</sup>. However, neighborhood relationships did not occur with all data, but were restricted by a "spatial contagion radius" such that all data contained at least 1 neighbor.

## VARIOGRAM AND ORDINARY KRIGING

The semivariogram is a function used to measure the spatial dependence relationships that exist in the data sample (ANDRIOTTI, 2009). It is constructed by measuring the covariances that exist between the data when taken two by two and within a distance  $h$  (ibid). By definition, it is given by the equation below:

$$\gamma(h) = \frac{1}{2} E\{[Z(x_i + h) - Z(x_i)]^2\} \quad (6)$$

, where  $h$  is represents a vector between two points in space. As  $h$  increases,  $\gamma(h)$  also increases until a maximum value at which it stabilizes, which is called sill. The distance at which  $\gamma(h)$  reaches the plateau is called range and represents the limiting distance of spatial dependence. Measurements located at distances greater than the range have a random spatial distribution and are therefore independent of each other. The neighborhood matrix was constructed using the inverse of the distance between the neighbors of each given sample<sup>2</sup>, the global Moran I index was calculated. The null hypothesis of non-spatial dependence of waste was refuted, justifying, this time, a spatial treatment of them by some criterion. We chose to interpolate them

1. Neighborhood matrices based on inverse distance more faithfully reproduce Waldo Tobler's "First Law of Geography": "all things are related to everything else, but things close are more related than things far away" (TOBLER, 1970).

2. It is not always possible to create neighborhood matrices with the radius stipulated by the variogram range value. In situations where the range is smaller than the minimum distance for all data to have at least 1 (one) neighbor, this is an example.

by kriging<sup>3</sup> and then incorporate it into the initial predication, in order to minimize errors arising from possible location variables not used in the initial modeling.

The Ordinary Kriging (KO) technique considers that spatially distributed data can be evaluated at any location from a small sample, and predictions at a given point and position:  $\hat{z}(s_0)$  from an unobserved location:  $s_0$  are obtained from the middleweights:  $\lambda_i$  of certain observations:  $s_i$  (WEBSTER e OLIVER, 2007):

$$\hat{z}(s_0) = \sum_{i=1}^n \lambda_i \times z(s_i) \quad (7)$$

The weights:  $\lambda_i$  (with a sum equal to unity) are chosen so that the error variance ( $s^2$ ) between the observed values:  $s_i$  and those sampled ( $s_0$ ) be minimized (equations 8).

$$\sigma^2 = \sum_{i=1}^n \lambda_i + \gamma(s_i, s_0) + \varphi, \text{ onde } \sum_{i=1}^n \lambda_i = 1 \quad (8)$$

The  $\varphi$  is the Lagrange multiplier that optimizes the kriging variance and depends on the spatial autocorrelation structure of the data:  $\gamma(s_i, s_0)$  obtained in the equation (6).

## RESULTS AND DISCUSSIONS

### GENERALIZED ADDITIVE MODELS (GAM)

As mentioned above, the optimized selection of the smoothing level of the predictors (k value) was obtained from the minimization of the Bayesian Information Criterion – BIC. Figure 1 shows the BIC variation with the spline smoothing exponents.

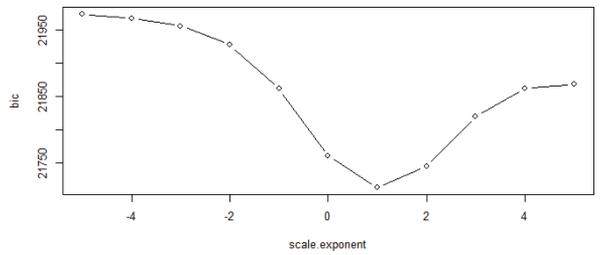


Figure 1 – BIC variation with the smoothing level of the predictors.

Source: own elaboration.

Equation (8) shows the GAM model with 77.30% explanation of the deviance with the fixed factors and smoothed functions that were significant. Table 2 shows the values of the coefficients of the fixed factors and their statistics. The model residuals approximately follow a normal distribution (see Figure 2).

$$\ln(vunit) = 6,12 + 0,44lotend + 0,07agua + 0,09pav + 0,05a2020 + 0,20tr + 0,28of + f_1(xcen,ycen) + f_2(distvp) + f_3(dscom) + f_4(vbt) + f_5(renda) + f_6(taeq) + f_7(are) + f_8(test) \quad (9)$$

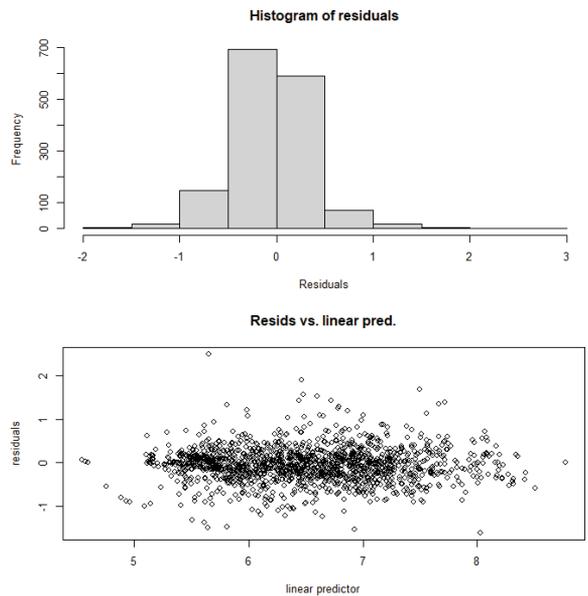


Figure 2 – GAM model residuals.

Source: own elaboration.

3. The kriging technique was formally (mathematically) developed by the Frenchman Georges Matheron, considered the founder of geostatistics, who paid homage to the name of the “kriging” technique in consideration of the South African mining engineer, Daniel Krigie, the first to use the technique. technique in an experimental way.

Parameter	Estimation	Standard Error	p-value
$\beta_0$	6,12	0,04	0,000
$\beta_1$ -lotcnd	0,44	0,09	0,000
$\beta_2$ -water	0,07	0,03	0,040
$\beta_3$ -pav	0,09	0,03	0,002
$\beta_4$ -a2020	0,05	0,02	0,031
$\beta_5$ -tr	0,20	0,05	0,000
$\beta_6$ -of	0,28	0,03	0,000

Table 2 - Estimate, standard error and p-values of fixed factors

Source: own elaboration.

In Table 2, the estimates of the model's fixed factors, represented by the dummy variables (0-without and 1-with), were all positive, as expected, and their marginal effect was calculated by the expression (for logarithmic link function):

$$efeito\ marginal = (e^{\beta_i} - 1) \times 100 \quad (10)$$

Thus it was possible to observe (on average): that there is an increase in the unit value of 55.32% when the lot is within a gated community; when the land has a water supply available, the unit value increases by 7.17% compared to those that do not have it; land located on paved roads is 9.69% more valued; the price values collected in 2020 are 21.66% higher than those in 2019; effective transaction prices are 21.66% higher than ITBI valuations and those referring to offer data are 32.41% higher than these same valuations. All growth presented confirms the initial hypotheses regarding the behavior of real estate market prices.

Figure 3 demonstrates the non-linear relationships of the variables area (area), distances to the main road (distvp), commercialization density (dscom) and the 2014 PGV land base value (vbt) and their respective smoothing functions, allowing some relevant considerations.

It is observed that the unit value of land decreases with area (principle of diminishing

marginal utility); the unit value decreases rapidly up to a distance of approximately 400m from the main road and then begins to decline more smoothly; the unit value grows up to 80% of commercial properties in the public area, starting to decline from there until 100% and the base value of PGV land significantly influences the unit values of land up to R\$ 400/m<sup>2</sup>, stabilizing from there.

The final predictions of the GAM model are found in Figure 4, where a greater error is observed for observed unit values greater than R\$2,500/m<sup>2</sup>.

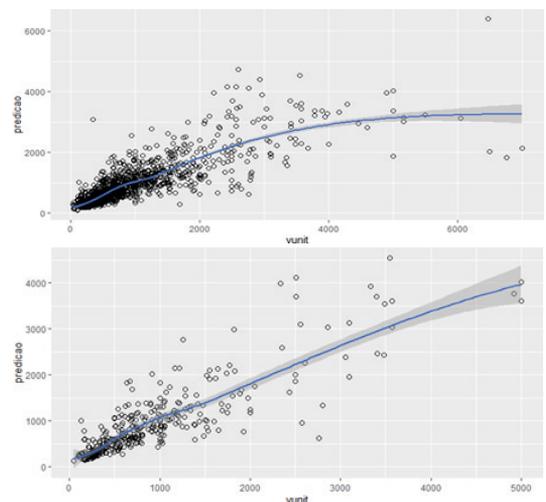


Figure 4 – Dispersion of what was observed (v unit) vs. prediction of the GAM model on the training and testing samples, respectively.

Source: own elaboration.

Figure 5 presents a three-dimensional map of the GAM model predictions, keeping all other variables fixed, with the exception of the centralized coordinates, xcen and ycen. Through it, the behavior of predictions over the surface of the municipality is verified, indicating an increase in prices with high values of ycen (approaching the north coast of Av. Beira Mar) and their decline in the east and west directions, being, in this last [west], a greater decrease was observed (the east coast with “Praia do Futuro” has higher observed unit prices).

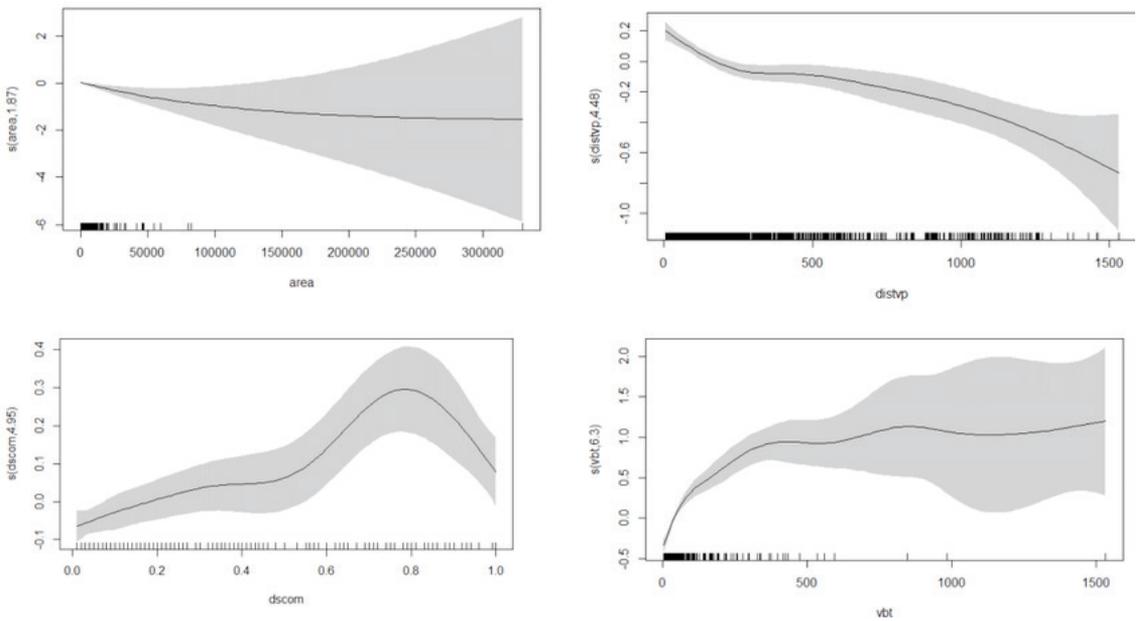


Figure 3 – Demonstration of the non-linear relationships of the variables area, distvp, dscom and vbt.

Source: own elaboration.

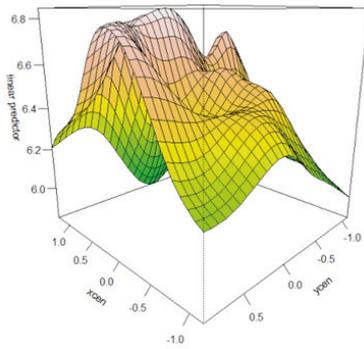


Figure 5 – Contour map of the GAM model predictions with variation only in the xcen and ycen variables.

Source: own elaboration.

### MORAN'S I INDEX OF RESIDUALS FROM GAM REGRESSION AND ORDINARY KRIGING

As previously stated, a neighborhood matrix was constructed with a distance of 1.542m, which is the minimum distance so that all data can have at least one neighbor. The null hypothesis of no spatial autocorrelation of residuals at 1% (p value ~ 0) was rejected, with the Moran's I statistic equal to 0.2906 (positive autocorrelation). The Moran scatter diagram

can be found in Figure 6 (left). As a result, there is a need to treat these wastes to improve the performance of the final prediction. To this end, a theoretical spherical variogram was adjusted with nugget=0, level of 233464 m and range of 870m, as shown in Figure 6 (right) and the interpolation of those residues was carried out using ordinary kriging.

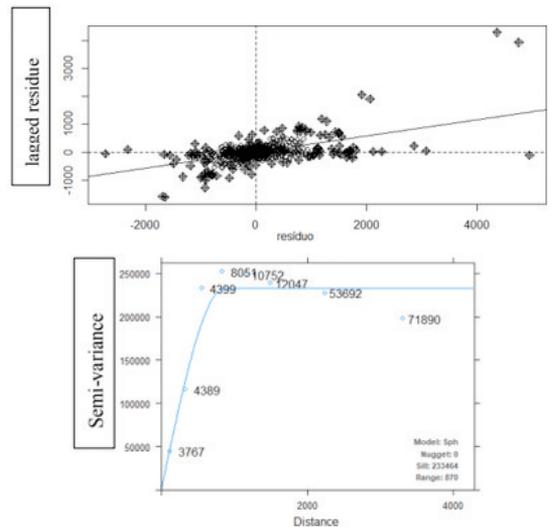


Figure 6 – Moran's I scatter diagram and theoretical variogram.

Source: own elaboration.

Figure 7 shows the residual map of the GAM model (training sample) interpolated by ordinary kriging. From this interpolation, the final prediction was calculated by summing its initial prediction value with the interpolated residue, both for the training sample and the test sample.

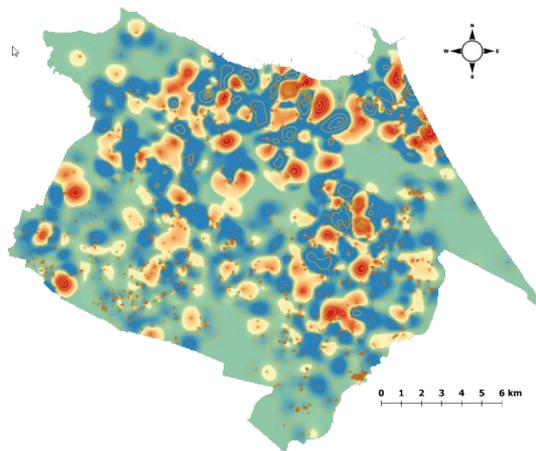


Figure 7 - Ordinary kriging map of GAM model residuals on training data.

Source: own elaboration.

In turn, the performance indicators for these predictions were calculated for each of the models and are presented in Table 3.

Performance Metric	GAM		GAM KO	
	Training	Test	Training	Test
Assessment level	1,05	1,06	1,00	1,00
COD	35,76%	34,40%	7,18%	13,06%
RMSE	474,35	413,14	112,04	204,48
MAE	262,08	253,93	50,43	90,48
MAPE	37,70%	36,71%	7,19%	13,10%
R <sup>2</sup>	0,68	0,73	0,98	0,93

Table 3 – Comparison of performance of GAM models with and without interpolation of residues by ordinary kriging (KO).

Source: own elaboration.

Table 3 shows a significant improvement in performance indicators over test predictions when added to their respective interpolated residuals. The dispersion coefficient (COD)

increased to 13.06%, which meets IAAO recommendations (maximum 30%), with 93% of the variability in observed prices explained by the model variables.

## CONCLUSION

In this work, we used a methodology applied to a sample of 1,924 data on urban land in Fortaleza obtained from the real estate market, from assessments and declarations for ITBI launch.

Initially, a Generalized Additive Model (GAM) was conducted, which confirmed some hypotheses regarding the behavior of real estate market prices, such as an average increase in the unit value of 55.32% when the lot is within a gated condominium; 7.17% when there is a water supply available and 9.69% when located on a paved road. It was observed that the tested area, utilization index, distance to main roads, commercialization density, income and base IPTU (urban property and territorial tax) value of the current PGV do not have a linear relationship with unit prices, each having a certain marginal effect curve.

Although an interaction term for terrain coordinates was incorporated into the GAM model, this was not sufficient to eliminate the spatial autocorrelation of the residuals. Thus, a spherical variogram was adjusted over these residues and interpolated using ordinary kriging, obtaining a surface of residues to be added to the final prediction.

With this procedure, there was a substantial improvement in all performance measures used in this work, such as COD, which went from 34.40% to 13.06%, which meets IAAO recommendations (maximum 30%). The MAE, from 253.93, became 90.48 and R<sup>2</sup> from 73% became 93%.

The proposed methodology using Generalized Additive Models and geostatistics for data from urban land in the municipality is very promising for use in mass assessment

of this nature, but must be used with great accuracy, as measurement errors in data collection can bias the prediction. Final.

## REFERENCES

- ANDRIOTTI, José Leonado Silva. **Fundamentos de Estatística e Geoestatística**. São Leopoldo: Editora Unisinos, 2009. 165 p.
- ANSELIN, Luc; REY, Sergio J. **Modern Spatial Econometrics in Practice: a guide to GeoDa, GeoDaSpace and PySAL**. Geoda Press LLC. Chicago, IL, 2014. 368 p.
- ARBIA, Giuseppe. 2014. **A Primer for Spatial Econometrics With Applications in R**. London : Palgrave Macmillan, 2014. 230 p.
- DANTAS, Rubens A. **Modelos Espaciais aplicados ao mercado habitacional – um estudo de caso para a cidade de Recife**. Recife, 2003 (Tese – Universidade Federal de Pernambuco – UFPE).
- GOLGHER, André B. **Introdução à econometria espacial**. Jundiaí: Paco Editorial, 2015, 384 p.
- HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: data mining, inference, and prediction**. 2. ed. Standford: Springer, 2008. 764 p.
- INTERNATIONAL ASSOCIATION OF ASSESSING OFFICERS (IAAO). 2010. **Standards on Ratio Studies**. International Association of Assessing Officers: Kansas City, Missouri, US.
- TOBLER, W. R.. **A Computer Movie Simulating Urban Growth in the Detroit Region**. Economic Geography, [S.L.], v. 46, p. 234-234, jun. 1970. JSTOR. <http://dx.doi.org/10.2307/143141>.
- TEAM, R. C.: 2012, **R: A Language and Environment for Statistical Computing**, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org/>.
- VEIE, Kathrine Lausted, PANDURO, Toke Emil. **An alternative to the standard spatial econometric approaches in hedonic house price models**, 2013. Disponível em: [www.ifro.ku.dk/english/publications/foi\\_series/working\\_papers/](http://www.ifro.ku.dk/english/publications/foi_series/working_papers/). Acesso:15 de agosto de 2020.
- WOOD, Simon N.. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. **Journal Of The Royal Statistical Society: Series B (Statistical Methodology)**, [S.L.], v. 73, n. 1, p. 3-36, 14 set. 2010. Wiley. <http://dx.doi.org/10.1111/j.1467-9868.2010.00749.x>.
- WOOD, Simon N.. **Generalized Additive Models: an introduction with r**. London: CRC Press, 2006. 397 p.
- WOOLRIDGE, Jeffrey M. **Introdução à econometria: uma abordagem moderna**. 6. ed. São Paulo: Cengage Learning, 2016. 848 p.