v. 4, n. 17, 2024

Scientific Journal of Applied Social and Clinical Science

REAL ESTATE PRICE PREDICTION USING SPATIAL ARTIFICIAL NEURAL NETWORKS: AN ALTERNATIVE TO MARKET VALUE FOR TAXATION OF URBAN REAL ESTATE

Arno Paulo Schmitz
Professional and Technological Education Sector, ``Universidade Federal do Paraná`` (UFPR)

André Klingenfus Antunes
``Universidade Federal do Paraná`` (UFPR)

Abstract: The spatial issue is important for many empirical problems. The location of properties in relation to their price (or market value) for property taxation is an example and requires efficient methodologies for its prediction. This study was dedicated to comparing two neural network methodologies for predicting the market value of properties that incorporate the location of the properties in different ways. To this end, a sample of property offers for the city of Curitiba/Brazil was used. In one model, using GWANN, the location was attributed through the latitude and longitude of the properties, intrinsic characteristics of the property and other locational characteristics such as distances to points of interest (POIs). The other model (RNAD), a traditional ANN, differed from the first by considering the neighborhood where the properties are located.

Keywords: *GWANN, ANN, Accuracy, Spatial Data, Non-Spatial Data.*

INTRODUCTION

Currently, methodologies linked to the area of machine learning have been used to predict property prices. The use of these methods aims especially at reducing prediction errors. Therefore, accuracy depends on the technique used and the problem in question (Binoy, Naseer, Kumar, & Lazar, 2021).

ANNs (Artificial Neural Networks) imitate the behavior of the human nervous system/brain with the aim of acquiring knowledge. For property pricing, this means learning from a sample that contains properties with their characteristics and assigning a price to a particular property.

In certain types of analyses, the location issue is important, that is, considering the location is essential to better understand what is being studied (Fotheringham, Brunsdon, & Charlton, 2002). For example, the location of the property is important for defining the price, since the property may have a certain increase in price (compared to other similar properties) simply because it is in a certain location. There are several experiments in modeling ANNs with different ways of imputing location. An extensive review of publications using ANNs and other methodologies for predicting real estate prices was published by Binoy et al. (2021). Few studies use ANN methodologies that incorporate location as an explicit characteristic of the property. One example is the method by Hagenauer & Helbich (2021), whose ANN called GWANN ("Geographically Weighted Artificial Neural Network") uses weights (like those used in the GWR model – "Geographically Weighted Regression") calculated from geographic coordinates to obtain a distance matrix that weights the model variables.

Three other examples are the methodologies by Rao, Wang, Wang, Khorasgani, & Gupta (2020), which are ANNs for longitudinal data (space/time). The first one is called GWFLM ("Geographically Weighted Functional Linear Model"), the second one is called SARFLM ("Spatial Autoregressive Functional Linear Model"). The third ANN methodology of this type is the graphical convolutional ANN (semi-supervised) (Zhu, Liu, Yao, & Fischer (2021).

In addition to the spatial issue, it is also possible to use property price predictions as a proxy for the property's market value. This value serves as the basis for the taxation of urban properties. In Brazil, the IPTU (Urban Property and Land Tax) is charged as a percentage of the property's market value, which is the result of the built area and the total area of the property, and other market conditions that determine the price (Afonso, Araújo, & Nóbrega, 2013).

Therefore, the general objective of this study is to compare the accuracy of the GWANN model (which incorporates the geographic location of the properties) with another basic ANN model that incorporates the spatial issue through binary location variables (neighborhoods). Both models were estimated with the same parameterization and sample of properties for the city of Curitiba/PR/Brazil obtained in Aug. 2021. To achieve this goal, the following specific objectives are imposed:

1. Obtain data on Points of Interest (POIs) - (shape files with information on POIs such as: schools; squares and gardens; hospitals, etc.); estimate the distances between properties and the nearest POIs; and build variables with this data.
2. Estimate two ANNs: the first (non-spatial) with variables that have exclusive data on properties, the neighborhoods where they are located and other proximity variables (distances) to the POIs - which is called the ANND model; and the second (spatial) with variables that have exclusive data on properties (including location - latitude and longitude) and other distance variables to the POIs - GWANN model;
3. Calculate accuracy indicators for the 2 models and compare them with each other.

METHODOLOGY AND DATA

RNAD MODEL

The ANN methodology requires a large number of parameters to be defined. In general, the specification of a neural network depends on: Topology ("feedforward/backforward propagation"); Number of neurons, nodes and layers; and learning rule. To predict real estate prices, it is necessary to use a supervised ANN method, since the aim is to predict a specific property price. To this end, the ANN "inputs" are the variables that aim to explain the behavior of real estate prices. On the other hand, the "output" is a vector with the prices of these same properties, as explained generically by Reed & Marks II (1999). The "inputs" weighted by the weights are subjected to an activation function, which can also be called a transfer function (Dangeti, 2017 | Hastie, Tibshirani, & Friedman, 2009). In practice, the use of "feedforward" and "backward propagation" networks are common. Formally, an ANN with n "inputs" can be represented as:

$$y_x = f\left(\sum_{i=1}^{n} w_i x_i\right)$$

Where: y_x are the "outputs" obtained by the "inputs" (x) received by RNA; w_i (x_i) "inputs"; x_i are the (i) "inputs"; $f(.)$ is the activation function.

GWANN MODEL

Developed by Hagenauer & Helbich (2021), the geographically weighted ANN (GWANN) couples' geographic weights to the connection weights of a traditional ANN. Therefore, its architecture is similar to that of a traditional ANN. The difference is that each of the GWANN output nodes is assigned a location in geographic space. This allows estimating the spatial distances between the "outputs" and the actual locations of the output nodes.

In addition, GWANN uses the geographically weighted error function to calculate an error signal. In the case of regression ANN, the geographically weighted error function is defined by the Equation:

$$E = \frac{1}{2}\sum_{i=1}^{n} v_i(t_i - o_i)^2$$

Where: E is the geographically weighted error; v_i is the geographically weighted distance between the observation (i) and the location of another node observation (i); t_i is the target value (the number of target values is equal to the number of nodes); o_i is the "output" of the node (i).

In the geographically weighted error function, the calculation of the "back propagation" error signal (for model training) is given by the expression:

$$\delta_j = \begin{cases} \varphi'(NET_j)\, V_j(o_j - t_j) & if\ j\ is\ an\ exit\ node \\ \varphi'(NET_j) \sum_{k=1}^{n} \delta_k W_{jk} & Otherwise \end{cases}$$

Where: δ is the sign of error; o_j is the knot (j) from the output layer; t_j is the target value of the node (j); W_{jk} are the weights of the connections between the node (j) and (k); δ_k is the sign of the node error (k); NET_j is the network input to the node (j); φ' is the partial derivative of the activation function; V_j is the GW distance between the observation at the middle layer node and the location of the output node (j). A graphical example of GWANN is shown in figure 1.

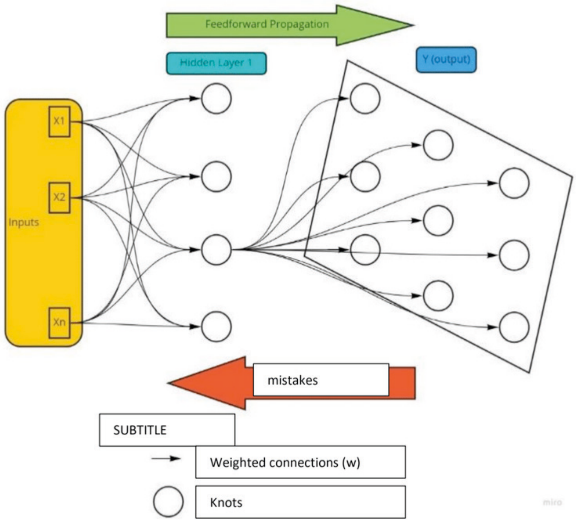

Figure 1: GWANN with two layers (one hidden layer and one output layer)

Source: Prepared by the author, based on Hagenauer & Helbich (2021)

The rectangle shows that the output nodes are distributed at locations on a plane. Although each node in the hidden layer has connections to all output nodes, for the sake of illustration, the output connections are shown for only a single node in the hidden layer. Then, the difference between the output layer values and the target values is weighted by the spatial distance between the location of the output nodes and the observation of the hidden layer node.

Geographic weighting is used only in estimating the error signal of the output layer nodes. On the other hand, the other nodes propagate the error signal of the later layer nodes backwards in the network.

DATABASE

Some data that would be useful for the present study are not accessible or available. However, Table 1 presents the variables with available data that were used in the present study.

The first 16 variables in the table refer to the sample of apartments and houses (ground floors and townhouses in condominiums or not) for the city of Curitiba/PR/Brazil obtained from ``*Imóvel web*`` (2021). The other variables were constructed based on the location of the properties.

PROGRAMMING, SETTINGS AND HYPER PARAMETERS

The "gwann", "MLmetrics" and "Metrics" packages (for the GWANN model) were used to estimate the accuracy indicators, which are executable in the "R" software, and the "scikit learn" library (for the RNAD model) in Python (Pedregosa et al., 2011).

All quantitative variables (that are not binary) in the database were normalized to reduce variance and, therefore, have zero mean and unitary variance. In addition to estimating the neural networks with the complete database, the sample was subdivided

Acronym	Variable Description	Model	Source
Lon	Longitude of the property location	GWANN	Real estate sales and offers data. ``*Imóvel web*`` (2021) – collected between Aug 1, 2021 and Aug 31, 2021
Lat	Latitude of the property location	GWANN	
Price	Property price – Response variable	RNAD AND GWANN	
Parea	Private area of the property in m2	RNAD AND GWANN	
Age	Age of the property	RNAD AND GWANN	
Tarea	Total area of the property in m2	RNAD AND GWANN	
Bed	Number of bedrooms	RNAD AND GWANN	
Bath	Number of bathrooms and lavatories	RNAD AND GWANN	
Ensuit	Number of suites	RNAD AND GWANN	
Garag	Number of parking spaces	RNAD AND GWANN	
Barb	“Dummy” for barbecue (1 = it has; 0 = otherwise)	RNAD AND GWANN	
Balc	“Dummy” to balcony (1 = it has; 0 = otherwise)	RNAD AND GWANN	
Elev	“Dummy” for elevator (1 = it has; 0 = otherwise)	RNAD AND GWANN	
Fitg	“Dummy” for gym and/or pool and/or games room (1 = it has; 0 = otherwise)	RNAD AND GWANN	
Party	“Dummy” for party area and/or barbecue (1 = it has; 0 = otherwise)	RNAD AND GWANN	
Categ	“Dummy” for property category (1 = apartment; 0 = otherwise)	RNAD AND GWANN	
Plaz	Distance from the property to the nearest square or garden	RNAD AND GWANN	Curitiba Urban Planning and Research Institute (2021) - (“shapefiles” with data on the location of POIs)
Park	Distance from the property to the nearest park or forest	RNAD AND GWANN	
Trans	Distance from the property to the nearest bus station	RNAD AND GWANN	
Kidca	Distance from the property to the nearest daycare center	RNAD AND GWANN	
School	Distance from the property to the nearest municipal primary school	RNAD AND GWANN	
Health	Distance from the property to the nearest hospital/ health unit	RNAD AND GWANN	
Bike	Distance from the property to the nearest bike path	RNAD AND GWANN	
Neig	Neighborhood where the property is located (binary categorical variables – 72 variables)	RNAD	
Crime	Number of police occurrence records in the neighborhood where the property is located (crimes against the person (assaults and violence) or property (residences and vehicles) –year 2020	RNAD AND GWANN	Secretariat of Public Security of the State of Paraná (2021)

Table 1: Variables used in the estimated models, according to source

into markets, by property type: 1 bedroom, 2 bedrooms, 3 bedrooms and 4 bedrooms and more; with the aim of verifying the results of the models in these specific real estate markets. For all models, the database was randomly divided into 80% for training and 20% for testing. A 5-Fold cross-validation was used, alternating the divisions of the database between training and testing. The "Tanh" activation function was used (the only one available for the "gwann" package). The maximum number of 2,000 iterations was also used and the processing was carried out in batches ("batch-size") of 50 sample elements. Additionally, the weight adjustment was performed using the SGD (stochastic gradient descent) optimizer with momentum equal to 0.9 and a Nesterov accelerator (also the only one available in the "gwann" package). Furthermore, a learning rate of 1% (0.01) was used.

For the number of hidden layers, 1 (one) hidden layer was used for all models, since the "gwann" package, to date, only allows this possibility. The number of neurons in the hidden layer was obtained by experimentation in the case of the GWANN model.

For the RNAD model, the iterative process via cross-validation and "grid search" in Python with the "Scikit Learn" library (search for the best number of neurons) were used. In the GWANN model, the Gaussian kernel function was used to estimate the spatial weight matrix and the optimal bandwidth was calculated by cross-validation.

The hardware computing structure used to estimate the models was Google Colaboratory (2022) - accessed at: <https://colab.research.google.com/notebooks> -, specifically Google "Colab Pro+". "Colab Pro+" provides access to virtual machines with T4 or P100 GPU. The amount of dedicated RAM in the "Colab Pro+" processing configuration was 51Gb High-RAM.

ACCURACY INDICATORS

The accuracy indicators that were used (because they are the most commonly used in published studies) are: MAPE - Mean Absolute Percentage Error; MAE - Mean Absolute Error; RMSE - Root Mean Squared Error; R2 - Coefficient of Determination. These indicators were calculated using the following expressions:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{p_i - \widehat{p_i}}{p_i}\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - \widehat{p_i})^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - \widehat{p_i}|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - \widehat{p_i})^2}{\sum_{i=1}^{n}(p_i - \bar{p})^2}$$

Where: p_i is the i-th observed (real) price of the sample; $\widehat{p_i}$ is the i-th predicted price; $\bar{p}$ is the average observed (real) price of the sample.

RESULTS AND DISCUSSION

The sample collected and used in the estimates contains properties in all neighborhoods of the city of Curitiba, a total of 9,788 properties, of which 6,640 are apartments and 3,148 are single-story houses or townhouses (in condominiums or not). Regarding the estimated models and their samples, Table 2 shows that the GWANN models are more computationally demanding, since the processing times are higher in all models. This occurs despite the greater number of variables in the RNAD models, which incorporate the binary variables of the neighborhoods.

When analyzing the information from the estimated models (Table 3), it can be clearly identified that the RNAD model, which is

Indicators by model/ market			Processing Indicators			
			Size of sample	Number of variables	Band size (GWANN model)	Processing Time (seconds)
Models	RNAD	1Q	523	93		1,7
		2Q	1.631	93		8,1
		3Q	4.618	93		44,0
		4Q+	3.016	94		46,8
		Geral	9.788	94		57,4
	GWANN	1Q	523	23	191	5.2
		2Q	1.631	23	25	63.4
		3Q	4.618	23	778	206.5
		4Q+	3.016	24	138	132.9
		Geral	9.788	24	1574	1066.3

Table 2: Sample information and processing indicators by model and real estate market

Source: Original research results

Indicators and Models by Market			Performance Indicators				
			RMSE	MAE	MAPE	R^2	Number of Neurons
Models	RNAD	1Q	0,10311	0,07022	0,10941	0,21	84
		2Q	0,15706	0,09291	0,51148	0,68	98
		3Q	0,22092	0,15198	0,87632	0,81	49
		4Q+	0,53561	0,37643	1,27597	0,78	54
		Geral	0,33816	0,21527	0,87425	0,88	51
	GWANN	1Q	0,07285	0,05380	0,06054	0,38	20
		2Q	0,24927	0,10071	1,39416	0,46	12
		3Q	0,30373	0,19681	1,12602	0,67	18
		4Q+	0,63194	0,45373	1,09883	0,70	17
		Geral	0,48412	0,31773	1,07066	0,76	21

Table 3: Accuracy indicators of the implemented models, by real estate market

Source: Original research results

the model with non-georeferenced spatial characteristics, obtained better predictive performance in almost all estimates, except for the 1-bedroom real estate market. The GWANN model, which considered geographic location, performed better in the 1-bedroom market, as this type of property is more concentrated in the downtown neighborhood of Curitiba. However, in general for predictive models, an R2 coefficient of 0.38 does not mean good predictive capacity.

On the other hand, the RNAD model, which included spatial variables based on distances to POIs and binary variables indicating neighborhoods, outperformed most markets. This can be seen when comparing the RNAD and GWANN models for the 2-bedroom and 3-bedroom markets, whose MAE, MAPE, RMSE, and R2 indicators were significantly superior (22 percentage points and 14 percentage points, respectively). It must be noted that the RNAD model also outperformed the total sample (with all markets and properties).

It is worth noting that the RNAD model achieved this superior performance despite the large number of variables and neurons in the intermediate layer. Despite the better indicators of the RNAD model, the GWANN model obtained good predictive capacity for the 3-bedroom and 4-bedroom markets and also for the sample in general (although inferior to the RNAD), with lower prediction errors (RMSE, MAPE and MAE) and with higher R2 indicators (0.67, 0.70 and 0.76 respectively).

FINAL CONSIDERATIONS

The conclusion from the results obtained is that, in general, the RNAD model performed better, except for the 1-bedroom market. This market has a smaller sample size and the properties are more concentrated in a neighborhood (city center), and are therefore closer together. Therefore, this model may be more accurate when the observations are closer together, given the effects of bandwidth.

Another issue that may be the origin of the result, but which deserves further study, is that perhaps the exact geographic location of the property (latitude and longitude) is not as important in defining the price of the property. In this regard, the neighborhood where the property is located may be more important. This would justify, in a way, some superiority of the RNAD model.

However, in predictive terms, the GWANN model provides an interesting answer, which is the possibility of pricing a property with geographic precision. In other words, the answer to the following question: What is the price of the property with such characteristics at the address "such"? Therefore, this model may be interesting when it comes to its practical utility. It must also be noted that the GWANN model obtained good predictive results for the 3-bedroom and 4-bedroom markets and for the sample in general.

For the GWANN model, the ANN was estimated with binary variables. However, in the partition of the test sample, some binary variables with only zeros were selected, which results in zero variance. This made it impossible to estimate the GWANN in this configuration.

Perhaps, making the estimates by neighborhoods (eliminating the need for binary variables) would be a way to improve the performance of the GWANN model. This expectation remains the subject of research for other and future studies.

Following up on this study, other studies can be developed to verify why in certain markets, despite the same parameterization of the ANNs, the models presented very different performance, that is, with very different quantitative indicators. This was observed in the 1-bedroom and 2-bedroom markets. Another issue that demands further studies is the comparison of the models used in this study with other spatial neural network models, which can contribute to a better understanding of this type of empirical problem (real estate prices).

REFERENCES

Afonso, J. R. R., Araújo, E. A., & Nóbrega, M. A. R. D. (2013). O IPTU no Brasil: um diagnóstico abrangente [PDF]. Retrieved from: https://repositorio.idp.edu.br/bitstream/123456789/1541/1/IPTU%20no%20Brasil%20Um%20Diagn%C3%B3stico%20Abrangente.pdf

Binoy, B. V., Naseer, M. A., Kumar, P.P. A., & Lazar, N. (2021). A bibliometric analysis of property valuation research. International Journal of Housing Markets and Analysis. 62-14. doi: https://doi.org/10.1108/IJHMA-09-2020-0115.

Curitiba. Instituto de Pesquisa e Planejamento Urbano de Curitiba [IPPUC]. (2022). Dados geográficos. Retrieved from: < https://ippuc.org.br/geodownloads/geo.htm>. Acesso em: 10 out. 2021.

Dangeti, P. (2017). Statistics for machine learning. Birmingham, England: Packt Publishing Ltd.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & sons Ltd, Chichester, England: John Wiley & sons Ltd,

Google Colaboratory. (2022). Processing suite. Retrieved from: <https://colab.research.google.com/notebooks>. Acesso entre: 02 fev 2022 a 03 ago 2022.

Hagenauer, J., & Helbich, M. (2021). A geographically weighted artificial neural network, International Journal of Geographical Information Science. 1-21. doi: DOI: 10.1080/13658816.2021.1871618.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. New York, USA: Springer.

Imovelweb. (2021). Portal de negócios de imóveis. Retrieved from: www.imovelweb.com.br

Pedregosa, F, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B, Grisel, O., … Duchesnay, E. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Retrieved from: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Rao, A. R., Wang, Q., Wang, H., Khorasgani, H., & Gupta, C. (2020). Spatio-Temporal functional neural networks. In: IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), IEEE, 2020, Sydney, NSW, Australia. Anais..., 81-89. doi: 10.1109/DSAA49011.2020.00020.

Reed, R., & Marks II, R. J. (1999). Neural smithing: supervised learning in feedforward artificial neural networks. Cambridge, USA: Mit Press.

Secretaria de Segurança Pública do Estado do Paraná [SSP-PR]. (2021). Base de dados de criminalidade. Restricted document.

Zhu, D., Liu, Y., Yao, X., & Fischer, M. M. (2021). Spatial regression graph convolutional neural networks: A deep learning paradigm for spatial multivariate distributions. GeoInformatica, 1-32. doi: https://doi.org/10.1007/s10707-021-00454-x.