Using Artificial Neural Networks to Uncover Real Estate Market Transparency: The Market Value

Laura Gabrielli(✉), Aurora Greta Ruggeri, and Massimiliano Scarpa

University IUAV of Venice, Dorsoduro 2206, 30123 Venice, Italy
laura.gabrielli@iuav.it

Abstract. In real estate property valuation, the availability of **comparables** is crucial. The reliability of the valuation of the **market value** depends on the number and on the accuracy of data that a professional can rely on. International standards suggest using historical prices as comparable since they are real transactions of sale/rent of a property that actually happened in a specific market. However, in the Italian real estate market, historical transaction prices are not available for professionals, and they have to base their valuations, primarily, on the **asking prices** enclosed in the **selling advertisements**. Asking prices can change in the future as they are subject to negotiation. Besides, sell ads always contain incomplete data or even wrong information. In this research, we employ Artificial Neural Networks to estimate how much offer prices and selling advertisements are misleading in property valuation in Italy. We, in a way, assess the opacity of the Italian real estate market, and we designate the major sources of error. The present work is a first step towards developing a model fitted for estimating data accuracy used generally in real estate estimates, namely, asking prices.

Keywords: Market value · Asking prices · Market transparency · Artificial neural networks

1 Introduction and Background

1.1 Comparison

Real estate valuation methods are based on comparison [1]. Comparison is one of the Italian five valuation principles, and it covers all the estimation approaches recognised by international standards [2]. Comparison, nonetheless, is the principal method sustaining the Market Approach aimed at identifying the market value of a property. Therefore, the only basis for estimating market value is a comparison between the property being valued and other similar properties with known price, cost, or income [3]. As far as the real estate market is concerned, the term "comparable" is often used in every language, both colloquially and within professional standards, in order to restrictively refer to a property, which is located in the same area and which is at similar maintenance level than the subject of the valuation, and whose sale/rental transaction has taken place recently [4].

O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12954, pp. 183–192, 2021.
https://doi.org/10.1007/978-3-030-86979-3_14

However, a comparable is not only the evidence of a transaction that has taken place in the past but rather any data that real estate professionals employ to formulate an appraisal judgement. Therefore, comparable can refer to historical transactions and asking prices, surveys, market quotations or land registry data, and evaluations made by other professionals. Not all these data provide the same level of information, and their reliability may strongly vary depending on the data source. In this context, a recent report by TEGOVA [5] aims to identify the role of comparables (and therefore their availability) in the quality of property valuation.

1.2 Market Transparency

Market transparency is a concept that belongs to the scientific literature and the theory of real estate valuation about the quality and the reliability of available data sources in a given market [6–11]. A very transparent market would ensure access to all the necessary comparables to be used in a property valuation. Instead, such data is either not available in an opaque market, not it can be purchased with high expenditures, and professionals must therefore refer to other sources of information [12]. A professional who is operating in an opaque market is, in a way, obliged to rely on information that another professional operating in a transparent market would classify as inadequate quality information. This does not depend on the expertise of the professional himself but rather on the natural constraints and limitations of the market in which the properties are valued [13].

The very same kind of data source that is judged as unsuitable for valuations purposes in a transparent market can be used appropriately in valuations, feasibility assessments or real estate market analyses in an opaque market. Each market is confronted with its data availability, and professionals rely on different data sources depending on the market in which they operate.

1.3 The Italian Real Estate Market

Market transparency is a huge problem in the Italian real estate market. Researchers struggle to collect data to build a large, statistically robust, transparent database. In the Italian real estate market, transactions data about sales/rent of properties are rarely available for professionals. So they must rely mainly on the asking prices included in the selling advertisements. Besides, real estate ads lack information. They are very inaccurate and, usually, they also contain inaccurate data, such as, to name a few, wrong localization, untrue energy class or false maintenance conditions. This may cause significant problems when developing forecasting models to predict the market value as a function of its building and neighborhood characteristics. They rely on data that contain themselves wrong information. However, the development of market value assessment tools is one of the major objectives in real estate appraisal and valuation. Reliable forecasting models should be created for this purpose. Accurate databases should consequently be produced and constantly updated. Several factors influence the market value of a property, and the contribution of each of them should also be precisely taken into account during the market value estimation.

1.4 Aim of the Research

This paper aims to investigate the transparency of offer prices data in the Italian real estate market. The aim is to investigate if the lack of knowledge of the historic transaction price is the only problem or whether the use of offer prices leads to other issues that increase the opacity of these data and the property valuations.

We carried out this study to estimate the error produced over the market value valuation when relying on offer prices only. Besides, we also pointed out the significant sources of error, identifying which variables influence the market value the most while containing misinformation or incorrect data.

2 Method

In the first step of this research, we have developed automated crawling software to automatically download the offer prices and the corresponding characteristics of a set of real estate properties on sale from specific selling websites.

We have defined the web search domain and let the web crawler download the required information from the online sell advertisements. This process allowed us to collect thousands of information about the offer prices and the characteristics of the properties in the chosen real estate market.

This process has led to limited knowledge of the given market since it is based only on selling ads. As stated in Sect. 1.3, the use of selling ads hides many other problems besides the intrinsic inaccuracy of the offer price. Selling ads contain, in fact, wrong information, incomplete data and even false statements.

In order to verify how much this inaccuracy influences the correct estimation of the market value, we collected, in a second step, a smaller number of samples of properties on sale exact in the same market but, this time, manually. “Manually collected data” means collecting data one by one and checking the level of maintenance and their precise location via Google Street View, or Earth, where this was possible. Therefore, we verified the correctness of all the collected information, such as the localisation of the premise or its maintenance conditions. Furthermore, we excluded from the database all the samples whose data could not be verified, and we corrected the wrong information declared in the ads.

Afterwards, we developed an Artificial Neural Network (ANN) based on the database collected by the automatic crawler. The ANN is an algorithm that, in this case, can predict the market value of a building as a function of some chosen building characteristics. The input neurons of the network contain the descriptive data of the premises, while the output neuron is the forecasted offer price.

We then used the same ANN to predict the market value of the database collected manually: the building’s correct characteristics collected manually constituted the input neurons, while the output neurons were the “forecasted market values”. We compared the “forecasted marked values” to the “expected market values”, where the “expected market values” were the prices manually collected.

Comparing a forecast value against its expected value gives a measure of the error produced by the inaccuracy contained in advertisements online. Besides, we could determine which information was having the highest error on the forecast, identifying the significant sources of error due to "opacity" in the Italian real estate market.

3 Creating the Web Crawler

First, we defined the selection criteria to identify a web searching domain. We limited the online search to residential properties on sale (not rent) in Padua. As far as the localisation is concerned, we considered all the fourteen areas the Municipality of Padua is divided into.

We included both new constructions and existing buildings for the building typology, comprising apartments, townhouses, detached and semi-detached houses, lofts and penthouses. This online search has led to 4,167 sale adverts. We have considered the most popular and acknowledged property selling websites in Italy, which we do not specify for privacy reasons.

In order to extract the necessary information with the web crawler from each sale advertisement, it was essential to know their corresponding web address. In fact, each one of the 4,167 results could have been identified through its Uniform Resource Locator (URL) in the form of an "https://…" web address.

All the sale adverts listed on the search-result page have an URL given from the combination of the URL of the search-result page and the serial adverts number.

For this reason, the web crawler we have developed in Python is able to read the URL of the search-result page, which is written in HTML language, extract all the serial numbers of the announces, and consequently build the URL of each data.

Afterwards, we implemented in Python the library "*Beautiful Soup*" to read the HTML pages of every sale advertisement. Beautiful Soup is a Python package explicitly used to parse HTML documents developed by Leonard Richardson. Since it creates a parse tree for all the parsed pages, this library can easily be used to extract data from HTML texts.

After, we have defined a class of objects and functions that produce the set of information extracted from each advertisement. The class is illustrated in Table 1.

Table 1. The class of objects and functions

Class Element	Units	Class Element	Units	Class Element	Units	Class Element	Units
Web URL	text	Construction year	number	Private garden area	sqm	Central heating	yes/no
Id	number	Status	text	Common garden	yes/no	Air Conditioning	yes/no
Zone	text	n. bathrooms	number	Garage	yes/no	Optical Fiber	yes/no
Address	text	n. rooms	number	Garage area	sqm	Building automation	yes/no
Latitude	coordinate	n. floor	number	Car box	yes/no	Photovoltaics	yes/no
Longitude	coordinate	n. of internal floors	number	Car box area	sqm	Solar panels	yes/no
Typology	text	Penthouse	yes/no	Cellar	yes/no	MCV	yes/no
Price	€	Lift	yes/no	Cellar area	sqm	Heat Pump	yes/no
Floor area	sqm	Energy Class	A/B/C/D/E/F/G	Terrace	yes/no	Alarm	yes/no
Price/sqm	€/sqm	Private Garden	yes/no	Terrace area	sqm	Fireplace	yes/no

Finally, we have applied in Python the data analysis library "Pandas" (developed by Wes McKinney) to extract a.xls file from the web crawling and organize data in the form of a table. Each row of the table represents an advertisement, while the columns show the class elements (i.e. property information).

4 Developing a Neural Network

We suggest employing Artificial Neural Networks to elaborate a forecasting tool to predict the market value of a property as a function of its intrinsic and extrinsic characteristics. Neural networks can be considered a computational system that acts out like human brains during learning biological processes. ANNs are basically constituted of artificial neurons, the computational units, and artificial synapsis, the connections between neurons.

ANNs are organized into multiple separated layers of neurons. The input layer contains the input neurons, while the output layer contains the output neurons. Between the input and the output layers, there is (are) one (or more) hidden layer(s). In this study, the input neurons are the intrinsic and extrinsic characteristics of the properties, whereas the output neuron is its corresponding market value.

The set of input neurons is represented as a column vector named [Xr], where $1 \leq r \leq R$, and the set of output neurons can be seen as a column vector called [Yp_forecast], in which $1 \leq p \leq P$. Yp_forecast is a function of vector Xr, so that [Yp_forecast] = f([Xr]).

4.1 Training of the Network

Through the training process, ANNs are able to "learn" how input neurons are related to their corresponding outputs.

Neural networks, in fact, analyse any input-output database and iteratively assess the free parameters of the network, i.e. the weighs (w) and the biases (b), until the best forecasting model is defined.

In order to understand this process, it is necessary to understand how information flow at the single-neuron level. Each z^{th} neuron receives one or more numerical inputs named $x_{z,u}$, $1 \leq u \leq U$, in which U is the total number of inputs/connections entering the z^{th} neuron. The information is combined inside the neuron, and a numerical output is consequently produced. Information is combined through the weight function ($w_{z,u}$) and the bias function (b_z), giving the output Y_z. Specifically, an activation function (φ_z) converts the neuron value into a response value as in Eq. 1:

$$\forall \text{ zth neuron}, Y_z = \varphi z\left(\sum\nolimits_{u=1}^{U}[(w_{z,u} * x_{z,u}) + b_z]\right) \quad (1)$$

During the training process, the weights ($w_{z,u}$) and biases (b_z) of the network are varied until the most reliable forecast is achieved so that vector Yp_forecast becomes the closest as possible to vector Yp_expected. In other words, weights and biases are

iteratively adjusted with the aim of minimizing the error signal (err_p). The error could be assessed as follow:

$$err_p = Y_{p\_expected} - Y_{p\_forecast}. \quad (2)$$

In Eq. 2 err_p is the error, Yp_expected is the target value, while Yp_forecast is the forecast value.

The total error on the forecasts is represented by a cost function as a way to estimate how wrong the forecasts are in comparison to the expected values contained in the dataset. For this reason, training the network means minimizing the cost function.

4.2 ANN as a Forecasting Tool

The database obtained through the web crawling was made of 4167 instances. However, this number had to be decreased by the 31.15% before training the ANN since we had to exclude the incomplete advertisements. In Table 2 we represent the percentage of incomplete announces per each class element.

Table 2. Percentage of incomplete announces per each class elements

Class Element	%	**Class Element**	%
Latitude	0.52%	Status	4.92%
Longitude	0.52%	n. bathrooms	3.39%
Price	3.15%	n. rooms	3.12%
Floor area	5.88%	Lift	0.15%
Price/sqm	8.56%	Energy Class	16.58%
Construction year	31.68%		

The number of training instances had to be further decreased down to 2,840 to eliminate the unlikely values from the dataset. At this stage, in fact, we could already exclude those advertisements that contained obvious outliers (such as 0 € as selling price, or 0 sqm as floor area). Besides, we had to exclude the construction year as a variable from the database since too much data were missing.

In Table 3, the progressive number of excluded outliers present in the corresponding number of advertisement is represented.

Table 3. Number of errors and outliers per number of respective ads.

N. ads	Errors and outliers per number of respective ads
2840	0
733	1
362	2
93	3
27	4
8	5
3	6
0	7

We could now define the training set to train the network by randomly selecting 60% of these 2,840 instances. Another 20% of the instances is randomly taken to define the selection set, and the remaining 20% forms the testing set. The training set is used to build several NN models. These different models are afterwards applied to the selection instances so that the one model that performs best on the selection set is chosen and then tested on the testing instances.

As a result, the ANN trained is based on the collected database shows 6 layers: 1 input layer, 4 hidden layers, 1 output layer. The input layer has 37 input neurons, while the output layer shows only 1 output neuron, i.e. the forecasted market value of the property (€/sqm), which is shown in Table 4. Conversely, the hidden layers present 32 hidden neurons each. The activation function employed is the hyperbolic tangent. The mean squared error function is the training strategy chosen, while the data scaling and unscaling process are based on a mean standard deviation scalarization.

4.3 Testing the Neural Network

The second database we collected manually to test the reliability of the ANN is constituted of 1,065 instances.

Again, we defined the same selection criteria to identify the web searching domain. The online search was limited to residential properties on sale in Padua. We decided to focus on the areas of Duomo, Forcellini, Santa Rita, Prato della Valle, Sacro Cuore, and Chiesanuova. Due to the higher availability of data, it was easier to check the correctness of the information contained in the advertisements. Moreover, those areas represent the Centre (Duomo and Prato della Valle), Semi-centre (Santa Rita and Forcellini), suburbs (Sacro Cuore, Chiesanuova).

As far as the building typology is concerned, we have included detached and semi-detached houses, apartments, townhouses, lofts and penthouses.

However, this time we did not simply transcribe the available data online. Instead, we verified the correctness of all the information. We excluded from the database those properties whose information could not have been verified. We added data when it was

Table 4. ANN input variables and output (target) variable.

n.	Variable	Use	n.	Variable	Use
1	Latitude	Input	20	AirConditioning	Input
2	Longitude	Input	21	OpticalFiber	Input
3	Floor_Area	Input	22	Fireplace	Input
4	Type_Villa unifamiliare	Input	23	Type_Apartment	Input
5	ns_Bathroom	Input	24	Type_Apartment in villa	Input
6	ns_Room	Input	25	Type_Attic	Input
7	Penthouse	Input	26	Type_Farm house	Input
8	Lift	Input	27	Type_Hamlet	Input
9	Energy_Classification	Input	28	Type_Loft	Input
10	Status	Input	29	Type_Portion of attic	Input
11	Garden_Private	Input	30	Type_Unfinished building	Input
12	Garage	Input	31	Type_Building	Input
13	ParkingSpace	Input	32	Type_Building for single family	Input
14	Cellar	Input	33	Type_Office	Input
15	Terrace	Input	34	Type_Single house	Input
16	BuildingAutomation	Input	35	Type_Terraced house	Input
17	CentralHeating	Input	36	Type_Semi-detached house	Input
18	Photovoltaics	Input	37	Type_Multifamily villa	Input
19	MCV	Input	38	Value_m2	Target

possible to find more specific details. For sure, this way of collecting data turned out to be a very long and time-consuming process. Nevertheless, still, it was the only way to produce a sort of litmus test to check the market transparency and the data correctness and the availability of information.

5 Results, Discussion and Conclusion

The ANN developed is now employed on the database collected manually so that the correct characteristics of the properties constitute the input neurons. In contrast, the output neuron forecasts the corresponding marked value. We, therefore, name this prediction as "forecasted market value". Conversely, the "expected market value" is the real asking prices we had collected manually.

Let's compare the "expected market value" against the "forecasted market value". It is possible to notice an average of 32.96% error in the forecasts (43.93% as the maximum error, 10.99% as the minimum error). These errors are enormous, and the problem mainly stands in the wrong information contained in the sell advertisements.

Finally, it is possible to determine which parameter is producing the highest impact on the forecast by analysing the correlation chart of the ANN, which is shown in Fig. 1.

This means that if a piece of wrong information in the ads regards the most impactful data (such as the status or the energy class), a considerable error will be made in the market value forecast.

In conclusion, it is possible to state that using artificial neural networks in combination with a web crawler helped estimate the level of opacity of the Italian real estate

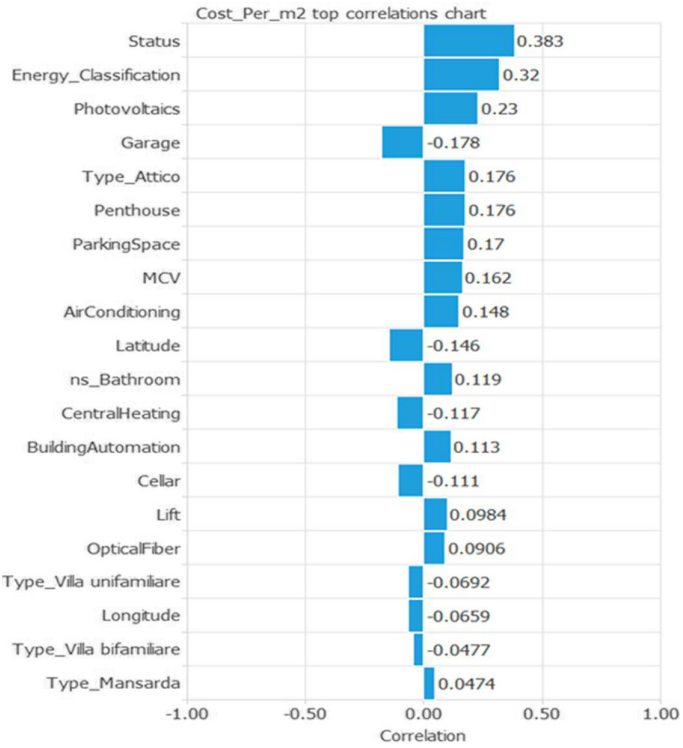

Fig. 1. Input-output correlation chart

market. Among the most significant achievements of this research, the automated web crawler made it possible to rapidly collect a huge amount of data and have a complete overview of all the properties on sale in Padua. Moreover, it is easy to perform this very same kind of analysis on other markets because the web crawler can be immediately applied to different contexts.

The major limitation of this approach is that it is based on offer prices and not on historical transactions. Clearly, the reason is that this is one of the primary sources of opacity in the Italian real estate market, however, as a further development of this research, the authors would like to compare offer prices results against historical transactions to analyse this other significant source of error in market value assessments.

For sure, the authors suggest that selling ads would become more rigorous in the displacement of information, in the correctness of the illustrated data and completeness. Some predefined layout should be slavishly followed by the sellers when composing advertisements, at least to provide complete and accurate information of the property on sale (which would also help potential buyers).

References

1. Forte, C., De Rossi, B.: Principi di economia ed estimo, Etas, Milan (1974)
2. Simonotti, M.: Metodi di stima immobiliare, Dario Flac, Palermo (2006)
3. Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., French, N.: Real estate appraisal: a review of valuation methods. J. Prop. Invest Finan. **21**, 383–401 (2003). https://doi.org/10.1108/14635780310483656
4. Orefice, M.: Estimo civile. UTET, Torino (1995)
5. Loberto, M., Luciani, A., Pangallo, M.: The potential of big housing data: an application to the Italian real-estate market. Banca d'Italia Eurosistema, p. 117 (2018)
6. Eichholtz, P.M.A., Gugler, N., Kok, N.: Transparency, integration, and the cost of international real estate investments. J. Real Estate Finan. Econ. **43**, 152–173 (2011). https://doi.org/10.1007/s11146-010-9244-5
7. Luo, Y., Chau, K.W.: The impact of real estate market transparency on the linkages between indirect and direct real estate. An MPhil Thesis 27 (2013)
8. Schulte, K.-W., Rottke, N., Pitschke, C.: Transparency in the German real estate market. J. Prop. Invest Finan. **23**, 90–108 (2005). https://doi.org/10.1108/14635780510575111
9. Cellmer, R., Trojanek, R.: Towards increasing residential market transparency: mapping local housing prices and dynamics. ISPRS Int. J. Geo-Inf. **9** (2020)
10. Bloomfield, R., O'Hara, M.: Market transparency: who wins and who loses? Rev. Finan. Stud. **12**, 5–35 (1999). https://doi.org/10.1093/rfs/12.1.5
11. Seidel, C.: Valuation of real estates in Germany. Methods, Development and Current Aspects of Research, pp. 213–220 (2006)
12. Newell, G.: The changing real estate market transparency in the European real estate markets. J. Prop. Invest Finan. **34**, 407–420 (2016). https://doi.org/10.1108/JPIF-07-2015-0053
13. Sadayuki, T., Harano, K., Yamazaki, F.: Market transparency and international real estate investment. J. Prop. Invest Finan. **37**, 503–518 (2019). https://doi.org/10.1108/JPIF-04-2019-0043