Detecting information transparency in the Italian real estate market: a machine learning approach

Laura Gabrielli, Aurora Greta Ruggeri**, Massimiliano Scarpa****

Parole chiave: Real Estate Market analysis; Market value; Asking price; Market transparency; Machine learning; Artificial Neural Networks

Abstract

This research aims to understand how market transparency and data reliability can influence valuation procedures and decision-making processes in the Italian real estate market.

Through the analysis of three different real estate markets and the validation of the information collected, this paper's goal is to understand whether and to what extent the use of asking prices instead of actual purchase and sale prices can lead to valuation errors, increase the uncertainty of valuation, and undermine investment decision-making processes.

The research results highlight the primary sources of information opacity in the Italian real estate market, classifying them according to their impact on real estate value.

The novelty of this research lies in the integrated use of machine learning techniques, computer programming and multi-parametric valuation procedures to understand and manage information opacity in the Italian real estate market, particularly regarding the estimation of the market value of properties belonging to the residential segment.

1. INTRODUCTION

Market transparency in a given market is related to accessibility and availability and the quality and reliability of data and information. Therefore, transparency is one of the prerequisites of a perfectly competitive market and the equilibrium condition. It translates into the perfect information for market agents (in particular consumers) on the prices and characteristics of goods or services.

In a **transparent market**, good quality data are available and accessible to both demand and offer. In such a case, agents can make purchase or production decisions based on the available data without requiring a high expenditure of resources for their collection. This assumption is crucial for the proper functioning of the market and the achievement of its efficiency. In contrast, if such data is either unavailable or very poor, the market is defined as **opaque.**

According to Schulte et al. (2005): "Real estate markets can be described as transparent when it becomes clear how the market mechanisms and the variables behind these mechanisms work, i.e., when there is as much information as possible available at any point in time."

John Lang Lasalle (2022) (hereafter JLL) defines a transparent real estate market as an open and clearly organised market that operates within a legal and

regulatory framework characterised by a consistent approach to applying rules and regulations.

The JLL's Global Real Estate Transparency Index (GRETI) 2022 ranks 94 Countries and 156 Cities according to market transparency. Since 1999, this index has considered several parameters, including investment performance measurement, market fundamentals, governance of listed vehicles, regulatory and legal environment, transaction processes, and sustainability transparency. In addition, the investment performance measurements include property valuations. Highly transparent markets are advancing thanks to technology, climate action, capital markets diversification and regulatory changes. It also marks a growing divergence between the leading and other opaquer markets, with many stalling or falling back in transparency growth. Sustainability was the main driver of transparency growth in the GRETI 2022 index, with many countries adopting mandatory energy efficiency and emission standards for buildings. From the digitalization perspective, the availability of new high-frequency and specific data is increasing the transparency of the real estate sector, leading to a greater understanding of how markets and buildings work.

In addition to JLL's reports, in recent years, the relationships between transparency and property market activities, such as investment, property development and valuation, have been examined in several papers (e.g. Farzanegan and Gholipour, 2014; Gholipour and Masron, 2013; Newell, 2016). The bibliography agrees that the problem of transparency in different countries can make real estate valuation, the ability to identify suitable investments, and dialogue between different market players difficult. Moreover, high market transparency is the first line of defence against uncertainty. Nevertheless, real estate transparency is not an unambiguous concept.

This paper aims to discuss how **information opacity** still represents a crucial problem in the Italian real estate market, specifically focusing on access to information in the valuation of the property market value.

For a deep understanding of the transparency level in the Italian real estate market and the possible error that would be made in real estate valuations due to lack of transparency, a model for defining and "measuring" the opacity of the market is proposed below.

To this extent, market transparency is here investigated with the help of machine learning techniques, and a set of Artificial Neural Networks (ANNs) is therefore developed with the aim of identifying the major sources of opacity in the market. In particular, the ANNs are intended as a multi-parametric forecasting tool able to estimate the market value of a property as a function of several building characteristics. The ANNs are trained on databases downloaded from specific selling websites with the help of an automated downloading procedure developed ad-hoc with the Python® computer language. These forecasting models reflect the "opacity of the market" because they are trained on raw data (opaque information), and selling ads, as will be further discussed, always contain incomplete, misleading, or even wrong information. In a second moment, the databases are re-downloaded by hand. This procedure allows for checking up on the correctness of all the information claimed in the selling ads, producing more transparent databases. The ANNs are implemented on these corrected databases, allowing us to understand how much information opacity influences the market value assessment. Besides, the primary sources of opacity are isolated and analysed.

As such, next **Section 2** will introduce and clarify some concepts that will be useful to the purpose of this analysis. Then **Section 3** will present the method adopted to discuss information opacity in the Italian real estate market, while **Section 4** will introduce the three practical case-studies object of this research. **Section 5** will present the ANNs training and testing procedures, and **Section 6** will illustrate the impact of the variables on the market value forecast. Finally, **Section 7** will discuss the conclusions of this work.

2. BACKGROUND

2.1 Transparency in real estate market

According to transparency theory, there is strict transparency in the market when information sharing occurs with a high degree of equivalence between market participants (supply and demand), while a lack of transparency causes information asymmetry in the market (Yun and Chau, 2013). Information asymmetry is when some market participants are more informed than others concerning the transaction information (Akerloff, 1970), thus leading to distorted outcomes, increased transaction costs, and increased risk.

The concept of transparency is conceived differently in the real estate market and is linked to several aspects. Schulte et al. (2005) link it to the possibility of obtaining information on the real estate market and the sub-markets into which it is divided. For Linquivitz (2012), transparency mainly concerns real estate transactions and five related aspects: legal information, financing taxation, transaction costs and data. Other authors directly connect the level of transparency to the technological development pervading the real estate market. Transparency requires open access to new and granular information with extensive geographical coverage but, at the same time, entails complete integrity and accuracy of sources (Ionaşcu and Anghel, 2020).

In real estate, a more transparent market is believed to attract more investments and investors (Razali and Adnan, 2012). More specifically, the higher the level of real estate transparency (RET), the higher the

participation and number of foreign investors in real estate (FREI). RET has been investigated in relation with also the default on mortgages – DOM (Gholipour et al., 2020).

Due to the globalisation of property transactions and the presence of foreign investors, the demand for (better) information on real estate data has increased significantly (Farzanegan and Fereidouni, 2014), but this could not be applied in every market, as stressed by the last report of JLL (2022).

Each real estate market shows its specific level of transparency, and the professionals in the field of property valuation must cope with the different quality and availability of the data sources they can use. For example, suppose a professional operates in an opaque market. In that case, she/he should base her/his appraisal judgements on poor-quality data, leading to less reliable valuations. If the same professional operates in a transparent market, she/he can dispose of numerous and detailed information, producing a more robust appraisal. Transparency in real estate has to be analysed concerning the peculiarities of the sector, which determine the different functioning of the real estate market compared to any other market (Arnott, 1987).

Besides, in the field of property investment, market opacity caused by a lack of detailed information on asset characteristics and prices leads to allocating resources inefficiently and will also increase investment risk. This is related to the problem of information asymmetry, which involves both the consumers, who cannot know all the legal, technical, and economic aspects of the asset they are willing to buy and the investors, who approach the investment with higher risk. In addition, market opacity makes property valuations more uncertain and prevents buyers from exercising control over the price and quality ratio of the properties they intend to purchase (Guerrieri, 2011).

A significant problem regarding market opacity, market asymmetry, and the availability of information concerns real estate valuation in Italy.

Italian valuation has faced international standards since the introduction of the first property investment funds and international listed companies in the late 1990s. Until then, an alignment with the other countries had not been necessary, but the arrival of international investors required further development in the Italian real estate and valuation sectors. Indeed, reform was needed in some respects, especially to codify the standard contents of a valuation report and clarify the steps to the estimation of the property value. The International Standards translated and introduced in Italy since the 2000s have sought to identify shared languages and procedures, best practices and market information analysis, contents and processes of a real estate appraisal. However, valuation methods and approaches were less concerning, as they were already used in the Italian appraisal sector. What required more effort was the access and the use of data and information valuations were based on. Adopting International Standards was impossible without first establishing the rules for collecting and analysing market data.

Undoubtedly, during the 2000s, in connection with the development of real estate funds and solid market growth, there has been a marked improvement in the quality and quantity of economic information.

The construction of historical price series and the development of numerous real estate market databases that provided real estate prices in different geographical areas have helped to spread good practices.

In the last decade, a series of publications, norms and regulations have been divulgated in the real estate market about collecting and analysing data (UNI/PdR 53:2019, 2019).

The Uni Standard, in particular, aims to identify guidelines on the methodology for collecting economic and real estate data to achieve the objectives of accuracy and comprehensiveness of the data collected and to achieve transparency and objectivity in real estate valuation.

In fact, it is possible to notice that the market transparency level in Italy has been improving according to JLL's Global Real Estate Transparency Index. This Index is expressed as a number so that the lower the number, the lower the market opacity. The Index attributed to Italy in 2001 was 2.82 (23rd out of 41 countries), while in 2022 it is 2.12. Italy is now in 19th place out of the 94 Countries analysed by the company. Italy has lost two positions since 2020, and it is 12th out of 29 European countries.

The partial scores (Fig. 1) for the individual categories of which the Index is composed cover Investment Performance, Market Fundamentals, Listed Vehicles, Regulatory and Legal, Transaction process and Sustainability. The best scores for Italy concern the transactions processes (1,33) related to pre-sale information, bidding processes, professional standards of agents, and anti-money laundering regulations. The worst performance is shown in the sustainability index (2,80), which involves green building certifications, energy reporting, energy benchmarking and efficiency standards, emissions reporting and standards, green leases, etc. The indices with the lowest value are those indicating greater transparency.

However, even though improvements have been made in the field of market transparency in Italy, several problems with sharing data and information still affect and slow down the fields of property valuation and investment in this Country. Without a doubt, the real estate market still needs to continue the process of evolving transparency in its data.

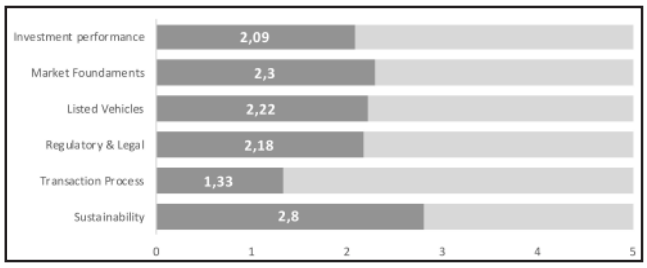

Figure 1 - *Italian Market Sub-Indices Performance* - Source: *John Lang Lasalle*.

2.2 Price versus value

For the purposes of this research, i.e. discussing information transparency in the Italian real estate market, it is crucial to distinguish between two key concepts that belong to the field of property valuation: price and value.

As thoroughly explained by (French et al., 2021; Forte e De Rossi, 1974; International Valuation Standards, 2020), **price** is the effective figure at which a property has been sold in an open market, and it is, therefore, historical data that can only be observed once the transaction has occurred. Conversely, **value** is a prior estimate of a price or, in other words, an estimate of the most likely figure that would be paid if a property were sold in an open market on the date of the valuation. Prices are historical facts, while values are hypothetical estimates of prices. Therefore, prices and values are profoundly different concepts, occurring at different negotiation times and embodying different roles.

The difference between prices and values is called **valuation accuracy**. As, again, well described in (French et al., 2021), valuation accuracy is where the market valuation (the estimate) differs from the actual price achieved. Valuation accuracy represents, therefore, how reliable the value assessment has been. This leads to the introduction of another important concept, namely valuation variability. The difference between two (or more) valuations performed by different professionals about the same property, at the same time, in the same market, is called **valuation variability**. Thus, we are talking about valuation accuracy, where market valuations differ from the actual sale prices (adjusted for time). However, when we look at how one valuation of the same asset differs from another, this is valuation variability.

Valuation accuracy and valuation variability, to a certain extent, both indicate the "error" that is made in the process of estimating the value of a property if compared to the actual sale price. A value assessment only is an estimate of a price based on the available information when the valuation is done, it is not an exact calculation of a figure. No mathematical formula or sophisticated econometric algorithm indicates a property precise value. It is challenging to obtain highly reliable assessments of market values since valuations are always characterised by uncertainty, which the valuer's skill, intuition, competence and experience cannot eliminate.

The reasons may reside on several grounds, some may certainly be under the valuer's control, but others are not. In fact, the value of an asset may be overestimated/underestimated, for example, due to insufficient analysis, misleading personal perceptions, prior assumptions, or data misinterpretation, which is (in a way) indeed the valuer's responsibility. However, most of the causes of low valuation accuracy might be out of the professional's control.

The paper illustrates that the first cause of low valuation accuracy is when a professional operates in a market with high **information opacity**.

Since information transparency/opacity is related to data availability and data correctness in a specific market, valuation accuracy and variability highly depend on the reliability of the information provided.

2.3 Comparison and comparables

Market transparency is directly related to both availability and correctness of the comparables, in all forms, that a professional can rely upon. In the Italian valuation discipline, one of the principles of real estate valuation states that the method is unique and is founded on **comparison** (Forte and De Rossi, 1974). Even in the International scenario, the comparison constitutes, in fact, the basis of all the valuation approaches recognised by international standards, i.e. the Market (or Sale) Comparison Approach, the Income Approach and the Cost Approach (Simonotti, 2006).

Value assessment procedures build up a comparison between the property being valued against other **comparables**, where a comparable can be defined as a property similar to the object of valuation but whose price/income/cost is known and belong to the same submarket. In addition, the comparable characteristics should be similar to those of the object of valuation in terms of intrinsic and extrinsic features, such as, to name a few, location, maintenance level, size, area or building typology.

For this reason, the property data of the properties used in the comparison are a crucial element for the process of determining the market value, whether economic facts (prices, costs) or technical data (quantitative and qualitative characteristics). Based on the quality and quantity of the available data, the valuer will choose the most appropriate approach and method to determine the value (French and Gabrielli, 2018).

That being said, several kinds of data sources could represent a comparable that real estate professionals use to estimate the price. In relation to the property market value calculation and the method applied, a

comparable could be an estate whose sale/rental transaction has taken place recently (historical price), it could be a property currently on sale (asking price), as well as any information contained in land registry data, market quotations, market surveys, or property appraisals made by other professionals.

Clearly, different comparables provide different kinds of information, leading to stronger or weaker appraisal judgements.

To this extent, a recent report by TEGOVA (French, 2020) points out how comparables play a crucial role in the quality and reliability of property appraisals, discussing how European Countries are adopting opposite legislations, tools and approaches to valuation. The TEGOVA report distinguishes two levels of information from which professionals can extract data, i.e. «hard» and «soft» information. The term "**hard information**" refers to historical data on the sale/rent of a property when all economic (price) and technical aspects (size, state of maintenance, number of rooms, floors, etc.) are known. The retrieval of this information is related to a transaction document (deed of sale), usually filed at notaries' offices, notary archives or Land Registry.

Conversely, "**soft information**" indicates indirect sources, reports from research companies, agencies, public databases or journals. In this case, the valuer needs help understanding how data have been elaborated or collected, and only a few general details are available. Usually, soft information is also presented in the form of aggregated data.

In the Italian real estate valuation system, this dual classification could be identified with the terms "**direct sources**" and "**indirect sources**", even if there is no total adherence between these two classifications. Direct sources also include the asking prices contained in sell advertisements and listings. On the contrary, in the TEGOVA report, asking prices are classified as soft information, together with indirect sources.

This difference between data sources classification in each country is related to the **availability of information**, local legislation, and the **degree of market transparency**. The adequacy of comparables is not universal (as JLL's Index shows), nor can it be used in the same way even within the same country (Ionașcu et al., 2021), as in the case of different data available in large cities versus small cities. This implies that international valuation standards should not be overly prescriptive in codifying the appropriate use of comparables, as each data source may play different roles depending on the transparency of the market (Sadayuki et al., 2019).

2.4 Relying on asking prices

Sellers indicate the **asking price** in the hope of attracting potential buyers willing to pay it. This price is identified through valuations, market analysis, or real estate agents. This price is placed at the beginning of the sale or sales negotiation and is not always a definitive price, being subject to a **margin of negotiation** in most cases. Only the sale price is placed at the end of the negotiation and it is a historical, unchangeable figure. The two prices are, therefore, diachronic.

Scholars repeatedly use asking prices for research related to property market analysis and property price modelling (Pozo, 2009; Hayunga and Pace, 2016; Gordon and Winkler; 2016). Moreover, some studies have related and compared the asking prices with the sale prices in real estate markets (Anglin et al., 2003; Beracha and Seiler, 2014; Knight, 2002). Curto et al. (2015) illustrated how the use of asking prices for property valuation is typical in Italy due to the lack of data on sales prices. The same study notes that there could be problems in using such values if they were used in property valuation. An animated debate took place among scholars and real estate market professionals in the summer of 2017 on the online Italian magazine "Monitor Immobiliare"[1], debating whether asking prices should be used and when. Opinions are not convergent: many professionals assert that the use is necessary when asking prices are unavailable. Others reject the employment of asking prices because they cannot reflect the market value of the real estate, while others accept a form of mitigation. There is much distance between what is written in the manuals, rules and regulations, and the real estate market, so operators must find adaptation forms necessary for their professional "survival". In Italy, despite greater market transparency and the construction of historical databases, access to data on buying and selling is still limited, meagre or hugely expensive. Also, there is a common misconception that historic price information suggests that investors are not pricing the future. This is not the case; an investor pays an agreed figure for the future expected returns, and thus price captures this view of the future. Valuations are proxies for price and thus also do the same, they reflect current market expectations of the future. The valuation is, therefore, the best estimate of the (future) sale price at the date of the valuation. This value must consider the particular real estate cycle, demand trends and the sustainability of this value in the extended period. Furthermore, in order to do this, adequate information is required.

Italian professionals often use **asking prices** as comparables. Asking prices are easily available from different online real estate marketplaces and they are cost – free. Following the suggestions of international standards, norms and regulations have recommended which data should be used in property valuations. The Italian standards (UNI 11612:2015, 2015) advocate that in

[1] *https://www.monitorimmobiliare.it/*

the case of valuations in which insufficient, undetectable and/or unreliable transactions have taken place in a recent period, asking prices may be taken into account only on a residual basis. However, the relevance of this information must be clearly defined, critically analysed and justified in the valuation report. This practice, however, extends beyond the exceptionality indicated by the Standard.

However, it is crucial to emphasise that even online real estate sales sites reflect the opacity of the Italian market. For example, it is possible to compare the data available online in Italy with those on i.e. American real estate websites. In the latter, it is displayed how often and for how many times a property has been bought and sold, the different selling prices, the amount of property taxes, and running costs. All this is possible thanks to the information in the real estate registers, which, in this case, is totally available online.

On the contrary, deeds of purchase are not accessible online on a large scale by real estate Italian players, and often the information is **scarce**. This is also the case when querying the data of the Agenzia Delle Entrate (Revenue Agency). For example, in the section 'Consultation of real estate price', economic information is available. Still, the information about the real estate unit sold needs to be improved (only the size, cadastral category, and date of the deed of purchase can be known).

Besides, even if historical transactions of proper comparables are found, still, it is not possible to ensure the **veracity** of the information declared in a deed. In fact, the search for comparables in Land Registries may lead to **incomplete or wrong information** about the transaction, the selling price or the property's characteristics.

Quantitative and qualitative characteristics, considered explanatory price variables, are an information problem when collecting sale prices.

Should asking prices, therefore, be considered unsuitable data for real estate valuations? What is certain is that a valuer should be conscious of the level of opacity they bring.

In the present research, the Authors base the entire analysis on asking prices. In particular, the comparables used to train the ANNs are represented by the online ads collected on specific selling websites in Italy.

Nevertheless, before debating whether the asking price may be an appropriate value to represent the (future) purchase price, it is necessary to understand how truthful the characteristics described in the advertisements are and how these, if wrongly described for various reasons, may influence the process of estimating market value.

How can properties summarily described in an advertisement be considered valid comparable data?

Can the consistencies, the state of maintenance, and the property characteristics be inferred with an adequate degree of reliability from commercial advertisements?

What mistake do valuers make if they use asking prices instead of historical transactions?

What are the most significant **sources of opacity** in the market?

3. A METHODOLOGICAL APPROACH

The methodology of analysis adopted in this research paper integrates computer programming procedures, machine learning techniques and multi-parametric real estate analysis, intending to understand how and how much the opacity in the Italian market affects the forecasts of market values. For a given market (case study), the following steps are defined.

- In the first step, an **automated crawling software** is developed in Python® language. A web crawler is a software that is able to browse and analyse the contents of web pages in a methodical and automated way. A web crawler is created for this research in order to parse specific Italian selling websites and automatically download the asking prices of the real estate properties currently on sale. Several building characteristics are also downloaded via the same web crawler alongside the asking prices. This procedure allows one to download thousands of observations easily. Each observation represents a property on sale whose asking price and characteristics are known. The database, collected via the web crawler for a specific market, is generally named $DB_{crawler}$.
- In the second step, an **Artificial Neural Network** is developed based on the $DB_{crawler}$, therefore being called $ANN_{crawler}$, in order to forecast the value of a premise depending on its characteristics in the given market.
  To this extent, in the field of machine learning, ANNs are a computational system that is able to learn patterns and procedures (Ćetković et al., 2018). Neural Networks are made of artificial neurons and artificial synapsis: the neurons are the computational units, while the synapsis connect the neurons between each other. Artificial neurons are organised into multiple separated layers so that the input neurons are displayed in the input layer, and the output neurons are contained in the output layer. At the same time, in between, there are multiple hidden layers of neurons. As in machine learning procedures, a network can "**learn**" how specific input information flow from the input neurons and generates an output. A network can be trained on any dataset of input-output information, consequently building a predictive model (Pittarello et al., 2021). In this paper, the input neurons of the $ANN_{crawler}$ are some selected characteristics of the premise, while the output neuron is its correspondent market value.

Specifically, ANNs here play the role of a **multi-parametric market value assessment technique**. ANNs are, in fact, used to analyse several building characteristics and forecast their market value. In the forecasting function, each independent variable (the building features) contributes differently to the estimate of the price (the dependent variable) (Simonotti, 2006). This way, it will be possible to isolate the impact that every variable brings to the price and understand what kind of error could produce a different level of opacity in information for each building feature.

The column vector of the input neurons is called $[X_r]_{crawler}$, with $1 \leq r \leq R$ and R being the total number of observations in the database. The output neuron is a one-column and one-row vector indicated as $[Y\_forecast]_{crawler}$. $[Y\_forecast]_{crawler}$, through the hidden layers, can also be defined as $[Y\_forecast]_{crawler} = f([X_r]_{crawler})$, namely a function of vector $[X_r]_{crawler}$.

When a network is trained on an input-output database, the network will "learn" how the input data are related to the corresponding output. A forecasting model is consequently assessed. During the training on a database, in fact, the network iteratively changes its free parameters, namely the weights (w) and the biases (b) of the network, until the best forecasting model is found.

At the single-neuron scale, during the training, every z^{th} neuron gets one or more numerical inputs, here named as $x_{z,u}$, with $1 \leq u \leq U$, and U is the number of artificial synapsis entering the neuron. Then the received information is all combined inside the z^{th} neuron using a weight function ($w_{z,u}$) and a bias function (b_z), so that a numerical output (Y_z) is produced. Then, an activation function (φ_z) transforms the neuron value into a response value, as defined through **Equation 1**, in the form of a weighted sum:

$$\forall \text{ neuron } z^{esimo}, \quad Y_z = \varphi z[(\sum_{u=1}^{U}[w_{z,u} * x_{z,u}) + b_z]) \quad (1)$$

During the training of the network, the weights and biases are therefore iteratively changed until the best forecasting model is achieved, or, in other words, until $[Y\_forecast]_{crawler}$ becomes the closest to $[Y\_expected]_{crawler}$. In this research $[Y\_expected]_{crawler}$ is the target, i.e the market value contained in the training database ($DB_{crawler}$), and $[Y\_forecast]_{crawler}$, as defined before, is the value produced by $ANN_{crawler}$. The difference between the target value minus the forecast produces an error signal named $[err]_{crawler}$, so as in **Equation 2**:

$$[err]_{crawler} = [Y\_expected]_{crawler} - [Y\_forecast]_{crawler}. \quad (2)$$

When training the previously defined $ANN_{crawler}$ on the information contained in $DB_{crawler}$, the aim is to minimize the sum of the error signals produced.

- In the third step, the Authors intend to verify the level of transparency of the information automatically collected via the web crawler. In fact, besides the inherent inaccuracy of the use of asking prices in property valuations, it is possible to verify that selling ads also contain false statements, wrong information or incomplete data.
  Since the paper aims to assess how and how much this misleading information affects the correct estimation of the property market value, in the third step of the research development, a **control set** of observations are re-collected **by hand**. However time-consuming this procedure is, it represents the only way to verify the correctness of all the information contained in the selling ads so that incomplete advertisements whose data could not be verified are excluded, incomplete information is found, and incorrect data are adjusted. This second set of information collected by hand constitutes the control database named DB_{hand}.
- In the fourth step, the previously developed $ANN_{crawler}$ is tested on the DB_{hand}. This time, the input neurons are the characteristics of the premises collected and corrected by hand, therefore named $[X_r]_{hand}$, while the output neuron is the consequent forecast of a market value $[Y\_forecast]_{hand}$. The expected market value, or the target value, is the price given in the selling ads, named $[Y\_expected]_{hand}$. Clearly, $[Y\_expected]_{hand} = [Y\_expected]_{crawler}$. The difference between the expected market value and the forecast, as in **Equation 3**, represents an estimate of the error produced due to the inaccuracies contained in the online ads (market opacity):

$$[err]_{hand} = [Y\_expected]_{hand/crawler} - [Y\_forecast]_{hand} \quad (3)$$

4. THREE PRACTICAL CASE-STUDIES IN NORTH ITALY

4.1 Databases downloading: the web crawler

For this research, it was chosen to analyse information opacity in three different Italian markets, specifically the real estate markets in:

- Bologna,
- Padova,
- Treviso.

In the Authors' opinion, these three cities could be good representative case-studies because they embody three different market sizes, where the market size refers here both to the size of the city, the number of elements

constituting supply and demand and the volume of transactions. Besides, although it is clear that every real estate market represents a specific case, none of these three cities is a strongly unique case as, for example, the markets in Venice, Rome or Milan might be. The results can therefore be significant also for other similar markets in Italy.

The first problem was understanding how to realistically collect the data and information required to develop the ANNs. As introduced in **Section 3**, an **automated crawling software** was developed in Python language to parse the real estate properties on sale in Bologna, Padova and Treviso from specific selling websites and automatically download their asking prices together with several characteristics of the buildings.

The web crawler online search must be targeted by the definition of three different **web searching domains**, each domain specific to one city. The three domains all comprise residential properties on sale in Bologna, Padova and Treviso. Non-residential properties are excluded from this study, such as commercials and directional, as well as premises on rental. As far as the building typology is concerned, new constructions and existing buildings are both included, such as apartments, attics, terraced houses, and multi/two/single-family villas.

For the first domain, all the 10 areas in Bologna are included in the downloading procedure, while the second domain contains all the 14 areas the Municipality of Padova is divided into. Finally, the 7 areas in Treviso are all comprised in the third domain.

In order to allow the web crawler to extract the information from each selling ad, it is necessary to insert inside the Python code each advertisement's own Uniform Resource Locator (URL) in the form of a web address, like "https://...".

Since it clearly would have been an unfeasible operation to manually open all the selling ads to copy and paste their web URLs inside the Python code, the goal was to understand how to produce them automatically.

It was possible to notice that each advertisement showed an URL given by the combination of the search-result page URL plus a serial number, where the search-result page can be defined as the list of all the selling ads resulting from the online search in the given domains.

Therefore, the web crawler has been programmed to extract the URL of the search-result page first, identify the serial numbers of the advertisment, and automatically build the corresponding URLs.

After that, the Python library "Beautiful Soup" was implemented to parse all the HTML pages of the sale advertisement. "***Beautiful Soup***" is a package developed by Leonard Richardson to analyse HTML documents. With the help of this library, it was possible to extract data and information from the HTML texts creating a parse tree for all the parsed pages. A **class** of objects and functions was built in Python to define the set of information to be extracted from each advertisement. The class used in each domain is illustrated in **Table 1**.

Table 1 - *Class of objects and functions*

Class element or Function	u.o.m.	Class element or Function	u.o.m.	Class element or Function	u.o.m.
IDENTIFICATION		PRICE		Lift	binary
Web URL	text	Price	€	Private Garden	binary
Ad number	number	Floor Area	mq	Private Garage	binary
Ad identification code	number	Unitary Price	€/mq	Shared garden	binary
LOCATION		FEATURES		Parking Space	binary
Zone	text	n. bathrooms	number	Basement	binary
Latitude	coordinate	n. bedrooms	number	Terrace	binary
Longitude	coordinate	Floor Min	number	Building Automation	binary
BUILDING TYPOLOGY		Floor Max	number	Central Heating	binary
Apartment	binary	Last Floor	binary	Photovoltaic System	binary
Attic	binary	MAINTENANCE		Mechanical Ventilation	binary
Terraced house	binary	Energy Class	number (1 to 10)	Air Conditioning	binary
Multi-family villa	binary	Mantenance Level	number (1 to 4)	Optical Fiber	binary
Two-family villa	binary	Construction Year	number (year)	Fireplace	binary
Single-family villa	binary	CHARACTERISTICS		Alarm	binary

The **class** of objects and functions defined above has been determined according to a specific analysis of the available information contained in the property selling websites and according to the most common characteristics of the buildings used in **multi-parametric** market value assessment procedures (Feng and Zhu, 2017; Wang and Xu, 2018). The attributes include structural/physical characteristics, neighbourhood, and location.

After the HTML pages of the selling ads had been parsed, the data analysis library "***Pandas*** "was applied in Python. Developed by Wes McKinney, the "Pandas" library is used to extract a .xls file from the web crawling procedure so as to display data and information in the form of a table. Each table row displays one observation, while the columns show the class elements downloaded per each ad. The database in Bologna, called $DB(B)_{crawler}$, presents 2,455 observations, in the downloaded database for Padova, named $DB(P)_{crawler}$, there are 2,884 observations, while the database in Treviso, defined as $DB(T)_{crawler}$, shows 1,473 observations.

4.2 Databases cleaning: removal of incomplete records and outliers

After the downloading procedure, it was necessary to clean the three databases from missing or misleading observations. First of all, incomplete ads were excluded from the training databases. Then, observations containing obvious errors or outliers were also excluded from the databases, for example, when showing a null selling price or a null floor area. Specifically, the percentages of incomplete/misleading observations are illustrated in **Table 2**, divided by each city and class element. Therefore, the percentages in **Table 2** represent the observations that were eliminated from the sampling because the data was not complete (or correct) in all its classes of objects. For example, the "energy class" and the "construction year" represent the highest percentage of lost observations due to missing information.

As such, the training databases downloaded via the web crawler had to be decreased, respectively, down to 1,665, 2,122 and 867 observations.

5. THE THREE NEURAL NETWORKS

5.1 Training with the Cuckoo optimisation algorithm

Once the three databases ($DBs_{crawler}$) had been cleaned from unusable observations, it was possible to perform the training of the three $ANNs_{crawler}$. The **training sets** employed to train the networks are formed by randomly selecting 60% of the observations of the $DBs_{crawler}$ per each city. Then, the **selection sets** are defined by randomly

Table 2 - *Percentage of lost advertisment per class*

Class	Percentage of lost observations		
	Bologna	Padova	Treviso
Zone	4,56%	0,43%	5,06%
Latitude	19,85%	2,77%	5,50%
Longitude	1,18%	0,73%	5,57%
Apartment	1,18%	0,00%	0,00%
Attic	1,18%	0,00%	0,00%
Multi-family villa	1,18%	0,00%	0,00%
Single-family villa	1,18%	0,00%	0,00%
Terraced house	1,18%	0,00%	0,00%
Two-family villa	1,18%	0,00%	0,00%
Price	1,18%	1,80%	2,65%
Floor Area	1,18%	1,32%	0,41%
n. bathrooms	1,18%	3,16%	3,73%
n. bedrooms	1,18%	3,47%	3,39%
Floor Min	1,18%	2,70%	5,84%
Floor Max	1,18%	2,70%	5,84%
Last Floor	1,18%	0,45%	0,20%
Energy Class	1,18%	22,33%	28,45%
Maintenance Level	4,56%	6,14%	7,06%
Construction Year	19,85%	28,54%	21,90%
Lift	1,18%	0,45%	0,20%
Private Garden	1,18%	0,45%	0,20%
Shared Garden	1,18%	0,45%	0,20%
Private Garage	1,18%	0,45%	0,20%
Parking Space	1,18%	0,45%	0,20%
Basement	1,18%	0,45%	0,20%
Terrace	1,18%	0,45%	0,20%
Building Automation	1,18%	0,45%	0,20%
Central Heating	1,18%	0,45%	0,20%
Photovoltaic System	1,18%	0,45%	0,20%
Mechanical Ventilation	1,18%	0,45%	0,20%
Air Conditioning	1,18%	0,45%	0,20%
Optical Fiber	1,18%	0,45%	0,20%
Fireplace	1,18%	0,45%	0,20%
Alarm	1,18%	0,45%	0,20%

taking another 20% of the remaining instances per city, whereas the left-over data create the respective **testing sets**. Three dataset split ratios have been considered: 80%-10%-10%, 70%-15%-15% or 60%20%20%.

The definitive split ratio has been chosen depending on the number of samples and inputs present in the dataset and the model. In this case, the databases are pretty small. However, numerous input neurons make the model more complex, so keeping an adequate number of observations in the selection and testing sets was crucial.

The training procedure was developed in Python code and implemented separately for each city. The training sets are firstly used to generate different ANN models; then, these models run on the selection sets to identify the one ANN that performs best also on the selection set. The testing set is finally used to calculate the error on the forecasts.

This training process is performed inside an optimisation procedure that allows to test the different ANN models to minimise the error on the forecasts. The optimisation algorithm employed during the training of the networks is the **Cuckoo optimization algorithm** (Chiroma et al., 2017; Mareli and Twala, 2018). It is a nature-inspired optimisation algorithm that identifies and compares all the ANN architectures showing a relative optimum (i.e. the minimum error) and chooses the global optimum among them. The Cuckoo search is, in fact, tailored to solve global optimization problems since it employs switching parameters and balancing local and global random walks.

For the sake of simplicity, the neural network developed for Bologna will be indicated as $ANN(B)_{crawler}$, for Padova as $ANN(P)_{crawler}$, and for Treviso as $ANN(T)_{crawler}$. The results of the ANNs developed are presented in **Table 3.** The mean errors produced on the testing set are 9,50% for Bologna, 11,75% for Padova and 13,55% for Treviso.

Errors are around the same order of magnitude. However, we can see an inverse correlation with the market dimensions: the more significant the market, the lower the error, and vice-versa.

Table 3 - *ANNs characteristics*

NNs	N. of inputs	Output	Number of hidden layers	Number of neurons per hidden layer	Activation functions	Scaling method	Training strategy
ANN(B)crawler	32	Market Value	7	184	ReLu	minimum-maximum	mean squared error
ANN(P)crawler	32	Market Value	7	144	ReLu	minimum-maximum	mean squared error
ANN(T)crawler	32	Market Value	7	144	ReLu	minimum-maximum	mean squared error

5.2 Testing information transparency

This section is dedicated to the use of the previously developed $ANNs_{crawler}$ to discuss how market transparency (or rather its opacity) affects the reliability of the forecasts in Italian markets.

For this purpose, the three databases have been re-collected by-hand to test the validity of the information that previously has been automatically downloaded via the web-crawler and used to build the networks. The databases re-collected by hand are called $DB(B)_{hand}$ for Bologna, $DB(P)_{hand}$ for Padova, and $DB(T)_{hand}$ for Treviso.

Obviously, the same selection criteria have been employed to identify the web searching domain as for the previous online search. The domains are again limited to residential properties on sale (excluding rental or other types of contracts), comprising existing buildings and new constructions.

During this second database collection, the correctness of all the information was verified, and those properties whose data could not have been verified were directly excluded from DBs_{hand}.

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

The three databases DBs_{hand} have been implemented on the $ANNs_{crawler}$, so that the **correct characteristics** of the properties now constitute the input neurons to produce a new market value forecast. As the Authors expected, the error produced on the forecasts is higher than before. The mean error for Bologna is 19,45%, for Padova is 25,57%, and for Treviso is 26,97%. This increase in the error on the forecasts is due to market opacity or, in other words, it is due to the wrong information stated in the ads. The larger the real estate market, the lower the average error rate: smaller markets show less transparency in the data described by property advertisements.

The **location** turned out to be the major source of opacity. Only a few checked advertisements showed the exact location of the building, whereas most were placed in another wrong street/district.

As a general trend, this problem was more accentuated

in Padova and Treviso than in Bologna. Besides, several ads had to be even excluded from this analysis because they did not provide enough information (pictures and descriptions) to allow for the exact identification of the position of the house. Certainly, the reasons behind this high opaqueness reside in the privacy requested by the owners and in the specific commercial strategy adopted by the real estate agency. In fact, if a potential buyer is not able to find the position of the house, a brokerage will be necessary. However, the lack of information about the location is excessive and misleading. Several ads indicated a location for the premises, which turned out to be completely wrong, even confusing central, semi-central and suburban areas.

Another huge source of opacity was the **maintenance level** since sometimes it did not match the other information reported in the ads, such as the energy class, the construction year, or the available installations. Generally, the maintenance conditions were optimistically assessed. In some cases, the pictures clearly showed maintenance levels in much worse conditions than the ones publicised in the advertisements. However, this assessment is subjective, and the authors modified the declared maintenance conditions only when undoubtedly different from the pictures provided. Other errors often contained in the ads regarded the identification and definition of the **building typology**, the availability of a **basement**, the **floor level**, the presence of a **garage** and a private/shared **garden**. In this regard, the online ads lacked some of this information, or they were not providing consistent data. The advertisements, in fact, contained both a description and a table, and some information presented in the description was completely different from those in the table.

Again, the definition of **penthouses** was usually vague and lacked transparency. The Italian word used in the advertisment to indicate a penthouse, i.e. "attico", should be used for luxury apartments placed at the last floor of a building. However, in some advertisements, the term "attico" was also used for regular or cheap apartments and attics. Due to the uncertainty given by this parameter, the Authors decided to flag just the condition of the "last floor", with no reference to the luxury level.

Other errors regarded the **installations**: the air conditioning or the mechanical ventilation systems were declared to be present in the house, but just the overall structure of pipes and ducts was prearranged for a hypothetical future installation.

There was also another branch of issues highlighting information opacity that was not directly related to the characteristics of the buildings. The online ads suffered a high **modification/expiration rate**. As a consequence, many advertisment were removed, reintegrated or modified in a time span of a few weeks only. This has led to require frequent refreshes and re-downloading of the databases, and the automated web-crawler developed in the frame of this research turned out to be extremely useful. However, also constant manual checks had to be performed multiple times, significantly increasing the total workload.

Another issue regarded an **asymmetric distribution** of information. For example, the ads for new buildings were usually richer in data than in the case of old buildings. As a consequence, more advertisements for old buildings had to be discarded during the manual check due to the lack of essential information. For this reason, the three databases had comparatively more advertisment about new buildings.

6. DEFINING VARIABLES IMPORTANCE

In the paragraph above, the primary opacity sources have been discussed. However, not all the errors in the ads produce the same impact on the market value prediction. The opacity in certain information is much more significant than in other. For this reason, it may be helpful to determine which input parameter shows the highest impact on the market value through a **feature importance** analysis.

Among the approaches that help calculating the variables' impact on the output are the Filter-based, the Wrapper, and the Embedded methods (Tatwani and Kumar, 2019).

Filter methods are based on univariate statistics, such as Pearson's correlation coefficient, the chi-square test, Fisher's Score, the Variance Threshold, the Dispersion ratio or the Mean Absolute Difference. The Wrapper based approaches consider the selection of a set of features as a search problem (Ghosh et al., 2020; Suresh and Narayanan, 2019; Yassi and Moattar, 2014), such as the Forward Feature Selection, the Backward Feature Elimination, the Exhaustive Feature Selection, or the Recursive Feature Elimination. Finally, Embedded methods combine the qualities of both Filter and Wrapper methods, such as the LASSO Regularization and the Random Forest.

Feature importance is assessed here using the **Random Forest** (RF) methodology. This approach was chosen because Embedded methods are highly accurate and show excellent generalisation properties (Siham et al., 2021).

A Random Forest is a particular **classifier** formed by a set of decision trees (simple classifiers) (Ugolini, 2014), where a decision tree, in the field of computer science, is a data structure made from nodes and arcs. A decision tree is read from top to bottom. The tree's nodes are the elements that contain the information, while the arcs are the connections between the nodes. The starting node is the root and does not have any input arcs, whereas the terminal nodes are named the leaves and do not show any outgoing arcs.

Each decision tree in a Random Forest procedure is built (i.e. trained) from a random subset of the training set. In this case, the three training sets are the buildings' information databases respectively collected for Bologna, Padova and Treviso. Besides, each decision tree is also built over a random extraction of the features analysed (i.e. the buildings information). This randomness in selecting features and observations is a key part of constructing the classifiers, and it is meant to increase their diversity to decrease their correlation.

In order to define the importance of each feature, it is necessary to measure how much each feature decreases its impurity during the training. In fact, the more a variable diminishes its impurity, the more significant that variable turns out to be. In classification (discrete variables), the impurity is given by the Gini impurity or by the information gain/reduction in entropy. In regressions (continuous variables), instead, the impurity is given by the variance.

A column matrix was defined, where the x-axes represent the features (the variables), and the y-axis shows the target (market value). The "**Numpy**" library was used to perform the RF-Regressor, and the analysis was conducted in Python. The RF Regressor is able to calculate the importance coefficients for each feature. The 70% of the observations were employed as the training set, whereas the leftover 30% was used as testing set. During the RF procedure, 2,000 trees were built, and the threshold set is 0.75 of the mean value of the importance coefficients calculated. In this case, the decrease in the impurity of each feature is assessed as the average of the decreases given by each tree constituting the forest. This way, the final importance of each variable is estimated. The importance coefficients calculated by the regressor are shown in **Table 4**.

Suppose a piece of wrong information in the ads regards the most impactful data, such as latitude and longitude (so location), maintenance level, or floor area. In that case, a considerable error will be made in the market value forecast. On the contrary, the less impactful variables are the installations and technologies such as air conditioning, mechanical ventilation, alarm, lift, or building automation. Among the less impactful variables are also the basement, the shared garden and the floor.

Table 4 - *RF importance coefficients*

Class element or Function	coefficienti RF		
	Bologna	Padova	Treviso
Variable	%	%	%
Latitude	22,43%	19,69%	15,38%
Longitude	29,30%	19,52%	13,38%
Typology	0,75%	0,72%	1,05%
Floor Area	8,17%	7,35%	7,71%
N. Bathrooms	0,87%	0,93%	1,20%
N. Bedrooms	1,50%	1,12%	0,97%
Floor Min	2,03%	2,07%	4,02%
Floor Max	1,53%	1,63%	1,43%
Last Floor	0,61%	0,40%	0,37%
Energy Class	3,19%	3,31%	6,99%
Maintenance Level	14,76%	20,19%	25,80%
Construction Year	7,55%	15,46%	10,03%
Lift	0,62%	1,00%	0,49%
Private Garden	0,65%	0,35%	0,34%
Shared Garden	0,50%	0,53%	0,99%
Private Garage	0,53%	1,01%	4,11%
Parking Space	0,61%	0,59%	0,51%
Basement	0,81%	0,78%	0,52%
Terrace	0,60%	0,53%	0,78%
Building Automation	0,15%	0,09%	0,12%
Central Heating	0,62%	0,40%	1,34%
Photovoltaic System	0,08%	0,22%	0,29%
Mechanical Ventilation	0,00%	0,16%	0,23%
Air Conditioning	0,46%	0,29%	0,23%
Optical Fiber	0,59%	0,54%	0,50%
Fireplace	0,50%	0,45%	0,59%
Alarm	0,58%	0,68%	0,64%

7. DISCUSSION AND CONCLUSIONS

This work has integrated real estate market analysis with multi-parametric market value assessment techniques, computer programming and machine learning procedures to **investigate market opacity in Italy**.

First, an automated web crawler developed in Python language helped to rapidly collect a considerable amount of observations describing the properties on sale in Bologna, Padova and Treviso. Based on these three databases, three corresponding Artificial Neural Networks have been trained in Python in order to forecast the market value of a property as a function of 32 input characteristics, including, among the others, location, maintenance level, installations and technologies, bounding typology, terrace, garage, or garden. Then, the three databases have been re-collected a second time by hand. This procedure allowed to control every piece of information stated in the ads. Wrong information was corrected, and non-verifiable observations were excluded. These three

"cleaned" databases have been implemented on the Artificial Neural Networks developed before: the error produced on the forecasts represents the error in the estimate of the market value due to market opacity (or, in other words, due to the wrong information contained in the ads). Then, a feature importance analysis was performed based on the Random Forest methodology.

As a result, this research can help understand how and how much market opacity in Italy affects the reliability of property valuations. Artificial Neural Networks are adequate forecasting statistical procedures, and neural network models accurately describe any input-output relationship. Using a neural network model to compare the results of an opaque database versus a "clean" database has led to determining how much a valuer misses the best estimate due to a lack of market transparency.

Besides, this multi-parametric analysis has also allowed to identify the most impactful variables on the estimate. This is crucial as those building characteristics/ information must be checked carefully to assess the market value properly. Therefore, a higher opacity in the most impactful variables will lead to a higher error in the forecast. In contrast, lower transparency in the less impactful variables would produce a minor error in the estimate.

At the end of this research, the Authors identify the need to improve the sharing of data and information about real estate properties in Italy. Historical transaction prices should be made available, but also the descriptions of the assets should be much more precise and complete. Moreover, selling ads could require a minimum level of data before being considered ready for online selling. Besides, the information provided in the ads must be accurate and correct, especially regarding the property's location. Finally, selling and ads should be more transparent; even different websites could respect a minimum-shared layout to ensure completeness and clarity.

The purpose is twofold: first, to reduce the information asymmetry between the seller and the buyer so that the demand could be able to move more consciously in the real estate market. Secondly, to increase transparency in the market, since those data are used by companies that analyse the market, produce reports, and publish prices, and by valuers, who sometimes have to use asking prices because transaction prices cannot be found. Moreover, since the real estate market has become complex and has substantial implications for the rest of the economy, all operators must be assured of the quality of information collected.

The debate analysed here revolves around the concept of "**quality of information**". Of course, the market investigation cannot disregard the in-depth analysis of every comparable, whether buying, selling, or bidding. However, it also seems constructive to focus on the procedures rather than the type of data: appropriate methodologies allow even spurious information to be approached professionally and constructively, drawing meaningful insights.

In further developments of this research, the Authors intend to periodically apply the methodology adopted in other Italian real estate markets to map the different levels of opacity and to understand whether there is an evolution over time regarding access to information. In particular, it will be interesting to see whether the dynamics that have occurred in recent years (Covid-19 pandemic, the war in Ukraine, energy crisis, inflation) may, in some way, impact, not only the dynamics and prices of real estate, but also the level of transparency or opacity.

* **Laura Gabrielli,** *Department of Architecture, IUAV University of Venice*
e-mail: *laura.gabrielli@iuav.it*

** **Aurora Greta Ruggeri,** *Department of Architecture, IUAV University of Venice*
e-mail: *aurora.ruggeri@iuav.it*

*** **Massimiliano Scarpa,** *Department of Architecture, IUAV University of Venice*
e-mail: *massimiliano.scarpa@iuav.it*

Nomenclature

$[err]_{crawler}$ - error signal $ANN_{crawler}$

$[err]_{hand}$ - error signal from DB_{hand}

$[Xr]_{crawler}$ - column vector of the input neurons in $ANN_{crawler}$

[Xr]hand - column vector of the input neurons from DB_{hand}

$[Y\_expected]_{crawler}$ - target value $ANN_{crawler}$

$[Y\_expected]_{hand}$ - target value from DB_{hand}

$[Y\_forecast]_{crawler}$ - one-column and one-row vector of the output neuron $ANN_{crawler}$

$[Y\_forecast]_{hand}$ - output neuron from DB_{hand}

$ANN(B)_{crawler}$, - ANN developed on $DB_{crawler}$ for Bologna

$ANN(P)_{crawler}$, - ANN developed on $DB_{crawler}$ for Padova

$ANN(T)_{crawler}$ - ANN developed on $DB_{crawler}$ for Treviso

ANN/ANNs - Artificial Neural Network/Artificial Neural Networks

$ANN_{crawler}$ - Artificial Neural Network developed on the basis of the $DB_{crawler}$

bz – bias function in a neuron

$DB(B)_{crawler}$ - database collected via the web crawler for Bologna

$DB(B)_{hand}$ - database collected by-hand for Bologna

$DB(P)_{crawler}$ - database collected via the web crawler for Padova

$DB(P)_{hand}$ - database collected by-hand for Padova

$DB(T)_{crawler}$ - database collected via the web crawler for Treviso

$DB(T)_{hand}$ - database collected by-hand for Treviso

$DB_{crawler}$- database collected via the web crawler

DB_{hand} - database collected by-hand

RF - Random Forest

U - number of artificial synapsis entering in a neuron

wz,u - weight function in a neuron

xz,u - numerical inputs in a neuron

Yz - numerical output in a neuron

φz - activation function in a neuron

Bibliography

Akerloff G.A., *The Market for "Lemons": Quality Uncertainty and the Market Mechanism,* The Quarterly Journal of Economics, 84(3), 1970, pp. 488–500.

Anglin P.M., Rutherford R. and Springer T.M., *The trade-off between the selling price of residential properties and time-on-the-market: the impact of price setting*, The Journal of Real Estate Finance and Economics, Vol. 26, No. 1, 2003, pp. 95–111.

Arnott R., *Economic Theory and Housing. In Handbook of Regional and Urban Economics*, edited by E. Mills, London: Elsevier, 1987, pp. 959–988.

Beracha E. and Seiler M.J., *The effect of listing price strategy on transaction selling prices*, The Journal of Real Estate Finance and Economics, Vol. 49, No. 2, 2014, pp. 237–255.

Ćetković J., Lakić S., Lazarevska M., Žarković M., Vujoević S., Cvijović J. and Gogić M., *Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application*, Complexity, Vol. 2018, Article ID 1472957, 2018, p. 10.

Chiroma H., Herawan T., Fister I., Fister I., Abdulkareem S., Shuib L., Hamza M.F. et al., *Bio-inspired computation: Recent development on the modifications of the cuckoo search algorithm*, Applied Soft Computing, Vol. 61, 2017, pp. 149–173.

Curto R., Fregonara E. and Semeraro P., *Listing behaviour in the Italian real estate market,* International Journal of Housing Markets and Analysis, Vol. 8, No. 1, 2015, pp. 97–117.

Farzanegan M.R. and Fereidouni H.G., *Does real estate transparency matter for foreign real estate investments?*, International Journal of Strategic Property Management, Vol. 18, No. 4, 2014, pp. 317–331.

Farzanegan M.R., Gholipour H.F., *Does real estate transparency matter for foreign real estate investments?* Int. J. Strateg. Prop. Manag. 18 (4), 2014, pp. 317–331.

Feng J. and Zhu J., *Nonlinear regression model and option analysis of real estate price*, Dalian Ligong Daxue Xuebao/Journal of Dalian University of Technology, Vol. 57, No. 5, 2017, pp. 545–550.

Forte C. and De Rossi B., *Principi Di Economia Ed Estimo*, Etas., Milan, 1974.

French N., *Pricing to Market. An Investigation into the use of Comparable Evidence in Property Valuation*, TEGoVA The European Gropu of Valuers' Association, June, 2020.

French N., Crosby N. and Thorne C., *Pricing to market: market value - the enigma of misunderstanding*, Journal of Property Investment and Finance, Vol. 39, No. 5, 2021, pp. 492–499.

French N., Gabrielli L., *Pricing to market: Property valuation revisited: the hierarchy of valuation approaches, methods and models*, Journal of Property Investment & Finance, Vol. 36, No. 4, 2018, pp. 391–396.

Gholipour F.H., Tajaddini R., Pham T.N.T, *Real estate market transparency and default on mortgages* Research in International Business and Finance 53 10120, 2020.

Gholipour F.H., Masron A.T., *Real estate market factors and foreign real estate investment*. J. Econ. Stud. 40 (4), 2013, pp. 448–468.

Ghosh M., Guha R., Sarkar R. and Abraham A., *A wrapper-filter feature selection technique based on ant colony optimization*, Neural Computing and Applications, Vol. 32, No. 12, 2020, pp. 7839–7857.

Gordon B.L. and Winkler D.T., *The effect of listing price changes on the selling price of single family residential homes*, The Journal of Real Estate Finance and Economics, 2016, pp. 1–31.

Guerrieri G., *L'informazione per l'efficienza e la trasparenza del mercato immobiliare: l'esperienza italiana*, Territorio Italia, n. 1, 2011, pp. 88–102.

Hayunga D.K. and Pace R.K., *List prices in the US housing market*, The Journal of Real Estate Finance and Economics, 2016, pp. 1–30.

International Valuation Stardard Council IVSC, International Valuation Standars, London, 2020.

Ionaşcu E., Anghel I., *Improvement of the real estate transparency through digitalisation*, Proceedings of the International Conference on Business Excellence Vol. 14(1), July, 2020, pp. 371–384.

Ionaşcu E., Taltavull de La Paz P. and Mironiuc M., *The Relationship between Housing Prices and Market Transparency. Evidence from the Metropolitan European Markets*, Housing, Theory and Society, Vol. 38, No. 1, 2021, pp. 42–71.

John Lang Lasalle, *Global Real Estate Transparency Index,*

2022 - Transparency in an age of uncertainty, Real Estate Transparency Report, available at: www.joneslanglasalle.com (accessed July, 2022).

Lindqvist S., *The concept of transparency in the European Union's residential housing market: A theoretical framework*, International Journal of Law in the Built Environment, Vol. 4, 2012, pp. 99–115.

Mareli M. and Twala B., *An adaptive Cuckoo search algorithm for optimisation*, Applied Computing and Informatics, Vol. 14, No. 2, 2018, pp. 107–115.

Newell G., *The changing real estate market transparency in the European real estate markets*. J. Prop. Invest. Financ. 34 (4), 2016, pp. 407–420.

Pittarello M., Scarpa M., Ruggeri A.G., Gabrielli L. and Schibuola L., *Artificial Neural Networks to Optimize Zero Energy Building (ZEB) Projects from the Early Design Stages*, Appllied Sciences, 11, 2021, p. 5377.

Pozo A.G., *A nested housing market structure: additional evidence*, Housing Studies, Vol. 24, No. 3, 2009, pp. 373–395.

Razali M.N. and Adnan Y.M., *Transparency in Malaysian Property Companies*, Property Management, 30(5), 2012, pp. 398-415.

Sadayuki T., Harano K. and Yamazaki F., *Market transparency and international real estate investment*, Journal of Property Investment and Finance, Vol. 37, No. 5, 2019, pp. 503–518.

Schulte K.-W., Rottke N. and Pitchke C., *Transparency in the German real estate market*, Journal of Property Investment and Finance, Vol. 23, No. 1, 2005, pp. 90–108.

Siham A., Sara S. and Abdellah A., *Feature selection based on machine learning for credit scoring : An evaluation of filter and embedded methods*, International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2021, pp. 1–6.

Simonotti M., *Metodi Di Stima Immobiliare*, Flaccovio, Palermo, 2006.

Suresh S.M.S. and Narayanan A., *Improving Classification Accuracy Using Combined Filter+Wrapper Feature Selection Technique*, IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1–6.

Tatwani S. and Kumar E., *Parametric comparison of various feature selection techniques*, Journal of Advanced Research in Dynamical and Control Systems, Vol. 11, No. 10 Special Issue, 2019, pp. 1180–1190.

Ugolini M., *Metodologie di apprendimento automatico applicate alla generazione di dati 3d*, 2014, aviable at https://amslaurea.unibo.it/10415/.

UNI 11612:2015, *Determination of the market value of properties*, 2015.

UNI/PdR 53:2019, *Real estate market analysis - Guidelines for identifying the market segment and collecting real estate data*, 2019.

Wang A. and Xu Y., *Multiple linear regression analysis of real estate price*, in IEEE (Ed.), International Conference on Robots and Intelligent System, ICRIS 2018, Changsha (China), 2018, pp. 564–568.

Yassi M. and Moattar M.H., *Robust and stable feature selection by integrating ranking methods and wrapper technique in genetic data classification*, Biochemical and Biophysical Research Communications, Vol. 446, No. 4, 2014, pp. 850–856.

Yun L. and Chau K.W., *The impact of real estate market transparency on the linkages between indirect and direct real estate*, paper presented at ERES Conference, Vienna, July, 2013, pp. 3–6.

www.build.it

dei
TIPOGRAFIA DEL GENIO CIVILE

Identificare la trasparenza informativa nel mercato immobiliare italiano: un approccio *machine learning*

Laura Gabrielli, Aurora Greta Ruggeri**, Massimiliano Scarpa****

Parole chiave: Analisi del mercato immobiliare; Valore di mercato; *Asking price*; Trasparenza; *Machine learning; Artificial Neural Networks*

Abstract

Questa ricerca si pone come obiettivo il comprendere come la trasparenza del mercato e correttezza delle informazioni influenzino i procedimenti di stima e i processi decisionali nel mercato immobiliare italiano. Attraverso l'analisi di tre differenti mercati immobiliari e la verifica delle informazioni relative ai prezzi di offerta, il presente contributo vuole capire se, e in quale dimensione, l'impiego dei prezzi di offerta in luogo dei reali prezzi di compravendita possano portare a commettere errori valutativi, ad aumentare l'incertezza nelle stime e a pregiudicare il processo decisionale negli investimenti.

I risultati della ricerca evidenziano quali sono le fonti primarie di opacità informativa nel mercato immobiliare italiano, classificandole in base al loro impatto sulla stima del valore immobiliare.

La novità di questa ricerca risiede nell'uso integrato di tecniche di machine learning, programmazione informatica e procedure di stima multi-parametrica al fine di comprendere e gestire l'opacità informativa nel mercato immobiliare italiano, in particolare riguardo la stima del più probabile valore di mercato degli immobili appartenenti al segmento residenziale.

1. INTRODUZIONE

La misura della **trasparenza di un mercato** è legata alla disponibilità, completezza, qualità e affidabilità dei dati e delle informazioni. La trasparenza è uno dei presupposti del mercato perfettamente concorrenziale e condizione per il suo equilibrio e funzionamento. Quando il mercato è trasparente, gli agenti del mercato (in particolare, i consumatori) hanno una perfetta conoscenza dei prezzi e delle caratteristiche di beni e servizi.

In un mercato **trasparente**, le informazioni sono chiare, veritiere, affidabili e complete, ed esse sono accessibili sia da parte dalla domanda e dell'offerta. In tal modo, i soggetti sono in grado di prendere decisioni di consumo o di produzione sulla base dei dati a loro disposizione, senza necessitare di un elevato dispendio di risorse per la loro raccolta e analisi. Tale presupposto è cruciale per il corretto funzionamento del mercato e per il raggiungimento della sua efficienza.

Al contrario, se tali dati non sono disponibili, sono spuri o quantitativamente e qualitativamente scarsi, allora il mercato viene definito **opaco.**

Secondo Schulte et al. (2005), i mercati immobiliari sono trasparenti quando è chiaro il loro funzionamento e le variabili che influenzano tale meccanismo, ovvero quando, in ogni momento, è disponibile un numero molto ampio di informazioni per gli attori del mercato.

Per John Lang Lasalle (2022) il mercato immobiliare può essere considerato trasparente quando è un mercato libero, efficientemente organizzato, che opera all'interno di un quadro giuridico e normativo chiaro e coerente.

Il *Global Real Estate Transparency Index* (GRETI) 2022 di John Lang Lasalle (d'ora in avanti JLL) ha classificato 94 Paesi e 156 città in base ad un indice di trasparenza del mercato. Pubblicato per la prima volta nel 1999, questo indicatore prende in considerazione diversi parametri, tra cui la misurazione della *performance* degli investimenti, i fondamentali del mercato, la *governance* dei veicoli quotati, il contesto normativo e legale, il processo di compravendita e la trasparenza circa la sostenibilità degli immobili. Tale indice misura, inoltre, la qualità nel mondo delle valutazioni immobiliari. I mercati che nel 2022 si sono affermati per la loro trasparenza hanno incrementano la loro posizione in classifica grazie allo sviluppo tecnologico, alle iniziative per il cambiamento climatico, alla diversificazione dei mercati dei capitali e ai cambiamenti normativi. Inoltre, l'indice segna una crescente divergenza tra i mercati più trasparenti e quelli opachi, con molti di questi ultimi che si fermano o, addirittura, arretrano nello sviluppo della trasparenza dei loro mercati. La principale spinta della crescita della trasparenza nell'indice GRETI del 2022 è data dal sub-indicatore della sostenibilità, poiché molti Paesi hanno adottato standard obbligatori di misurazione dell'efficienza energetica e delle emissioni degli edifici. Dal punto di vista della digitalizzazione, la disponibilità di database sempre più affidabili e completi permette la crescita della trasparenza nel settore immobiliare, portando ad una maggiore comprensione delle dinamiche di funzionamento dei mercati e del comportamento degli edifici.

Da una approfondita analisi della bibliografia esistente, è possibile individuare un filone di ricerca che negli ultimi anni ha esaminato il rapporto tra trasparenza e mercato immobiliare, le valutazioni e gli investimenti, e che è stato poi oggetto di diverse pubblicazioni (Farzanegan e Gholipour, 2014; Gholipour e Masron, 2013; Newell, 2016). Gli studiosi concordano sul fatto che il problema della trasparenza nelle diverse nazioni può rendere complesso il processo di valutazione immobiliare, riducendo la capacità di scegliere gli investimenti più appetibili e compromettendo il dialogo tra i diversi attori del mercato. Solo un'elevata trasparenza informativa consente di gestire l'incertezza e il rischio, e, in generale, tutte le componenti stocastiche che sono presenti in ogni mercato.

Il presente lavoro si propone di discutere come **l'opacità informativa** rappresenti ancora un problema cruciale nel mercato immobiliare italiano, focalizzando l'attenzione sulle informazioni impiegate durante il processo valutativo della stima del valore di mercato degli immobili.

Per comprendere il livello di trasparenza del mercato immobiliare italiano e l'errore che è possibile commettere nelle stime immobiliari a causa della mancanza di essa, si propone di seguito un approccio innovativo per la verifica dell'opacità di mercato.

A tal fine, la trasparenza viene indagata con l'ausilio di tecniche di *Machine learning*, e si sviluppa un set di Reti Neurali Artificiali (*Artificial Neural Network*, le ANN) con l'obiettivo di individuare le principali fonti di opacità del mercato. In particolare, le ANN vengono impiegate come uno strumento di previsione multi-parametrica in grado di stimare il valore di mercato di un immobile in funzione delle diverse caratteristiche descrittive di ciascun edificio. Il *training* delle ANN si svolge su database ottenuti da specifici siti commerciali di offerte immobiliari disponibili online, grazie all'impiego di un processo di *downloading* automatico sviluppato ad-hoc con il linguaggio informatico Python®. Questi modelli di previsione riflettono l'«opacità del mercato» perché sono addestrati su dati relativi ai prezzi di offerta, come si illustrerà più avanti, che per loro natura contengono informazioni incomplete, fuorvianti o addirittura erronee. In un secondo momento, i database sono stati verificati puntualmente, al fine di controllare la correttezza di tutte le informazioni riportate negli annunci di vendita, producendo così database più trasparenti. Le ANN sono poi implementate sui database corretti, consentendo di conoscere quanto le informazioni poco attendibili influenzino la valutazione del valore di mercato. Inoltre, vengono isolate e analizzate le principali cause di opacità dei database.

La ricerca si sviluppa come segue. Il **Paragrafo 2** introduce e discute il concetto di trasparenza e come esso sia trattato nella letteratura. Il **Paragrafo 3** illustra l'approccio adottato per individuare l'opacità informativa nel mercato immobiliare italiano, mentre il **Paragrafo 4** presenta i tre casi-studio su cui verrà applicato il modello. Il **Paragrafo 5** descrive le procedure di training e testing delle Reti Neurali, mentre il **Paragrafo 6** chiarisce l'impatto delle variabili del modello sulla stima del valore di mercato. Infine, nel **Paragrafo 7** verranno presentate le conclusioni di questo lavoro.

2. BACKGROUND

2.1 Il concetto di trasparenza nei mercati immobiliari

Secondo i modelli teorici, un mercato è definito trasparente in senso stretto quando tutti gli attori del mercato (sia sul versante della domanda che dell'offerta) possiedono lo stesso grado di informazioni e la stessa possibilità di accesso alle fonti. Di converso, la mancanza di trasparenza causa asimmetria informativa nel mercato (Yun e Chau, 2013). L'asimmetria informativa si verifica quando alcuni attori sono più informati di altri durante tutte le fasi della compravendita immobiliare (Akerloff, 1970), e ciò determina distorsioni informative, maggiori costi di transazione e un incremento della rischiosità degli investimenti.

Sul significato di "trasparenza" non vi è convergenza, soprattutto se legato al particolare contesto immobiliare. Schulte et al. (2005) associano la trasparenza alla possibilità di raccogliere informazioni sul mercato immobiliare e

sui sotto-mercati in cui esso può essere segmentato per tipologia, funzioni, utenza, ecc. Per Linquivitz (2012), la trasparenza concerne principalmente le transazioni immobiliari e gli aspetti legali, la tassazione, i costi e le informazioni ad esse legati. Altri autori collegano direttamente il livello di trasparenza allo sviluppo tecnologico che pervade il mercato immobiliare: la trasparenza richiede un libero accesso a informazioni puntuali con ampia copertura geografica, ma allo stesso tempo implica completezza e accuratezza delle fonti informative (Ionaşcu e Anghel, 2020).

Nel settore immobiliare, si ritiene che un mercato più trasparente attragga più investimenti e investitori (Razali e Adnan, 2012). In particolare, maggiore è il livello di trasparenza (*Real Estate Transparency* - RET), maggiore è la partecipazione di investitori stranieri attratti dal settore immobiliare (*Foreign Investors in Real Estate* FREI). La RET è stata studiata in relazione anche all'insolvenza dei mutui (Gholipour et al., 2020).

A causa della globalizzazione delle transazioni immobiliari e della presenza di investitori stranieri, la richiesta di (migliori) informazioni sui dati immobiliari è aumentata in modo significativo (Farzanegan e Fereidouni, 2014), ma ciò non può essere applicato indistintamente in tutti i mercati, come sottolineato dall'ultimo rapporto di JLL del 2022.

Infatti, ogni mercato immobiliare presenta un livello di trasparenza specifico e i professionisti che operano nel campo della valutazione immobiliare devono confrontarsi con la diversa qualità e disponibilità delle fonti di dati che possono utilizzare. Se un professionista opera in un mercato opaco, dovrà necessariamente basare i propri giudizi di stima su dati di scarsa qualità, pregiudicando l'affidabilità della stima. Se lo stesso professionista operasse in un mercato trasparente, con la possibilità di disporre di numerose informazioni puntuali, la stima del più probabile valore di mercato potrebbe risultare più robusta ed attendibile. Quindi il concetto di trasparenza nel settore immobiliare deve essere analizzato in relazione alle peculiarità del segmento specifico, che determinano il diverso funzionamento del mercato immobiliare rispetto a qualsiasi altro mercato (Arnott, 1987).

Inoltre, nel campo degli investimenti immobiliari, l'opacità del mercato causata dalla mancanza o scarsità di informazioni dettagliate sulle caratteristiche e sui prezzi degli *asset* causa una inefficiente allocazione delle risorse ed un aumento della rischiosità dell'investimento. Ciò è legato al problema dell'asimmetria informativa, che coinvolge sia i consumatori, che non possono conoscere tutti gli aspetti legali, tecnici ed economici del bene che intendono acquistare, sia gli investitori, che affrontano maggiori rischi durante l'investimento. Inoltre, l'opacità del mercato rende più incerte le valutazioni degli immobili ostacolando così la verifica del rapporto tra qualità e prezzo degli immobili da parte degli acquirenti (Guerrieri, 2011).

Un problema rilevante relativo all'opacità del mercato, all'asimmetria del mercato e alla disponibilità di informazioni riguarda la valutazione immobiliare in Italia.

L'estimo in Italia si è allineato con il contesto internazionale fin dall'introduzione dei primi fondi di investimento immobiliare e di società quotate multinazionali verso la fine degli anni '90. L'arrivo e l'interesse degli investitori internazionali ha imposto un confronto tra le diverse impostazioni teoriche della dottrina estimativa e delle metodologie valutative. Gli Standard Internazionali tradotti e diffusi in Italia a partire dagli anni 2000 hanno introdotto standard etici e professionali, buone pratiche, principi, regole e procedure per garantire stime oggettive, ripercorribili e trasparenti. Se la disciplina estimativa italiana ha una fisionomia ben delimitata, il cui impianto teorico poggia su una serie di postulati aventi validità generale, gli Standard Internazionali hanno invece cercato di individuare linguaggi e procedure condivise, regole di natura metodologica e applicative, *best practice* per la rilevazione delle informazioni del mercato, dei contenuti e i passaggi di una valutazione immobiliare.

Tuttavia, l'adozione di Standard Internazionali è molto complessa, se non impossibile, senza prima stabilire delle regole per la raccolta e l'analisi delle informazioni di mercato. Ciò ha richiesto uno sforzo maggiore in relazione all'accesso, all'utilizzo e alla diffusione dei dati e delle informazioni immobiliari.

Indubbiamente, negli anni 2000, in concomitanza con lo sviluppo dei fondi immobiliari italiani e la rapida crescita del mercato, si è assistito a un netto miglioramento della qualità e della quantità delle informazioni economiche. Grazie alla creazione di serie storiche dei prezzi, alla divulgazione di studi e di ricerche, alla crescita della copertura territoriale degli osservatori del mercato in grado di fornire quotazioni per diverse categorie, tipologie e microzone si sono diffuse una serie di buone pratiche per quanto concerne l'analisi di mercato.

In particolare, nell'ultimo decennio, la Normativa UNI (UNI/PdR 53:2019, 2019), insieme ad altre pubblicazioni e regolamenti, ha focalizzato l'attenzione sulla raccolta dei dati immobiliari e la loro analisi. La Norma Uni, in particolare, ha come obiettivo l'individuazione di linee guida sulla metodologia di rilevazione dei dati economici e immobiliari, per conseguire obiettivi di accuratezza, completezza dei dati rilevati, e conseguire trasparenza e oggettività nella stima immobiliare.

Come conseguenza di questo nuovo contesto di mercato, è possibile notare che, secondo il *Global Real Estate Transparency Index* di JLL, il livello di trasparenza del mercato in Italia è cresciuto negli anni. L'indice di JLL viene espresso attraverso un numero indice: più basso è tale valore, meno opaco è il mercato. L'indice attribuito al mercato italiano nel 2001 era pari a 2,82 (collocato al 23° posto su 41 Paesi, e considerato un mercato semi - trasparente). Nel 2022 tale indice è sceso a 2,12 (mercato trasparente), portando l'Italia al 19° posto su 94 Nazioni ora incluse nel paniere monitorato dal JLL. L'Italia ha perso due posti rispetto al 2020 (quando si collocava al 17° posto), ed è ora

al 12° posto tra i 29 Paesi Europei analizzati.

I punteggi parziali (Fig. 1) per le singole categorie di cui si compone l'Indice di JLL riguardano la *Performance* dell'investimento, i Fondamentali del mercato, i Veicoli di investimento quotati, le Leggi e Regolamenti, il Processo di transazione immobiliare e la Sostenibilità. I punteggi migliori raggiunti dall'Italia riguardano il processo di compravendita (1,33), in particolar modo le informazioni prevendita, la commercializzazione del bene nel mercato, gli standard professionali degli agenti immobiliari e la normativa antiriciclaggio. La performance peggiore si registra nell'indice di sostenibilità (2,80), che riguarda le certificazioni di bioedilizia, la rendicontazione energetica, gli standard di efficienza, il calcolo delle emissioni gassose, i contratti di locazione verdi, ecc. Nel grafico, gli indici con il valore più basso sono quelli che indicano una maggiore trasparenza nel mercato italiano.

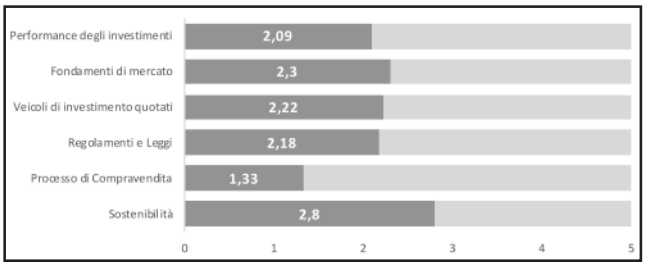

Figura 1 - *Andamento dei sottoindici del mercato italiano* - Fonte: *John Lang Lasalle*.

Tuttavia, anche se in Italia sono stati fatti grandi passi in avanti a partire dagli anni '90 ad oggi, la difficoltà di reperimento e di diffusione delle informazioni e dei prezzi di mercato impattano sulla qualità delle valutazioni immobiliari, sulle analisi di mercato e sulle valutazioni delle opportunità di investimento nel nostro Paese. Senza dubbio, il mercato immobiliare deve continuare il processo di crescita del livello di trasparenza cominciato oramai da qualche decennio.

2.2 Prezzo di compravendita e valore di mercato

Ai fini della presente ricerca, che ha come obiettivo l'analisi e la comprensione della trasparenza dei dati e dell'uso delle informazioni nel mercato immobiliare italiano, appare fondamentale distinguere tra due concetti chiave nell'ambito delle valutazioni immobiliari: prezzo e valore.

La differenza tra prezzo e valore è puntualmente illustrata da (Forte e De Rossi, 1974; French et al., 2021; International Valuation Standard Council, 2020): il **prezzo** rappresenta la quantità di denaro realmente pagata da un acquirente per un immobile che viene scambiato nel mercato, un dato storico che può essere osservato solo a transazione avvenuta. Il **valore** di mercato di un bene è, invece, un giudizio di stima, una previsione, l'importo al quale l'immobile potrebbe essere venduto alle normali condizioni mercato alla data della stima. Il prezzo, pertanto, è un dato storico definito dal mercato, una quantità di denaro effettivamente pagata per l'acquisto di un immobile, mentre il valore è un giudizio di stima che esprime, in termini ipotetici, il futuro prezzo del bene. Se ne deduce che prezzo e valore sono concetti profondamenti differenti, che si verificano in tempi differenti della negoziazione e assumono ruoli diversi nel mercato.

La differenza monetaria tra prezzi e valori è misurata dall'**accuratezza** della stima. Come descritto in (French et al., 2021), si definisce accuratezza della valutazione la misura in cui la stima del valore di mercato differisce dal prezzo di compravendita raggiunto dalla contrattazione, misurando, in qualche modo, l'affidabilità della valutazione immobiliare. Di converso, la differenza tra due (o più) stime immobiliari del medesimo bene, redatte da professionisti diversi nello stesso momento e nelle stesse condizioni di mercato, è chiamata **variabilità.** Si parla, quindi, di accuratezza della stima quando i valori di mercato individuati differiscono dai prezzi di compravendita effettivi, mentre si parla di variabilità delle stime quando si confrontano valutazioni dello stesso bene effettuate da soggetti diversi.

L'accuratezza e la variabilità della valutazione, in una certa misura, indicano entrambe l'«errore» commesso nel processo di stima della grandezza economica rispetto al prezzo di vendita effettivo.

La stima del più probabile valore di mercato è una indicazione di un possibile prezzo futuro basata sulle informazioni disponibili al perito al momento della stima. Ciò, tuttavia, non si riduce ad una mera elaborazione matematica dei dati compiuta attraverso un più o meno sofisticato algoritmo, che porta alla formulazione esatta dell'entità economica (prezzo, o costo). Pur potendosi avvalere oggi di complessi modelli econometrici capaci di riprodurre con sufficiente attendibilità i meccanismi del mercato, nessuno di essi potrà pervenire alla indicazione di un prezzo "certo", in quanto, per sua natura, la stima ha carattere probabilistico, non deterministico, e le stime sono sempre connotate da un certo livello di aleatorietà, che abilità, intuizione, competenza e esperienza del valutatore non possono eliminare completamente.

Le ragioni degli errori possono essere diverse, alcune controllabili dal valutatore, altre no. Infatti, il valore del bene può essere sovrastimato o sottostimato a causa di una scarsa analisi di mercato, di convinzioni personali fuorvianti o errate, di ipotesi sbagliate, o anche di una interpretazione erronea dei dati reperiti nel mercato. Tuttavia, al di là di alcune cause che possono essere diretta responsabilità del valutatore, la scarsa correttezza delle stime potrebbe essere legata a fattori non controllabili, quali il rischio, l'incertezza e l'azione delle componenti stocastiche presenti nel mercato.

Come sarà illustrato nel presente contributo, la principale causa di una bassa correttezza della valutazione si ha quando un professionista deve operare in un mercato con un livello elevato di **opacità informativa**. Poiché la trasparenza/opacità informativa è legata alla disponibilità e alla correttezza dei dati ottenibili in uno specifico mercato, l'accuratezza e la variabilità della valutazione dipendono fortemente dall'affidabilità delle informazioni fornite.

2.3 La comparazione e i dati di mercato

La trasparenza del mercato è direttamente correlata alla disponibilità, alla correttezza e all'affidabilità dei campioni significativi dei prezzi di vendita e delle caratteristiche dei beni analoghi all'oggetto di stima che il professionista può impiegare durante la perizia. Nella disciplina estimativa italiana, uno dei principi dell'estimo afferma che il metodo di stima è unico ed è basato sulla **comparazione** (Forte e De Rossi, 1974).

Anche nel contesto internazionale e all'interno della scuola di matrice anglosassone, la comparazione costituisce il fondamento di tutti i procedimenti di stima riconosciuti dagli Standard Internazionali, ovvero il *Market Comparison Approach, l'Income Approach, il Cost Approach* (Simonotti, 2006).

La comparazione si basa sul confronto dell'immobile da stimare con **beni simili**, per i quali risultano noti il prezzo, il reddito o i costi, ricadenti nello stesso segmento di mercato. Oltre alle grandezze economiche, debbono essere note anche le caratteristiche estrinseche ed intrinseche degli immobili che compongono il campione significativo di comparazione, quali, per citarne alcune, l'ubicazione, il livello di manutenzione, la superficie, l'affaccio, il livello di piano o la tipologia edilizia.

Per tale ragione, i dati immobiliari degli elementi impiegati nella comparazione costituiscono un elemento cruciale per la stima, siano essi fatti economici (prezzi, costi) o dati tecnici (caratteristiche quantitative e qualitative). In base alla qualità e alla quantità dei dati disponibili, il valutatore sceglierà il procedimento (e, successivamente, il modello matematico) più idoneo per giungere alla formulazione del giudizio di stima (French e Gabrielli, 2018).

Diverse tipologie di informazioni (prezzi effettivi, prezzi richiesti, prezzi di locazione, costi) possono rappresentare dati comparabili che i professionisti del settore immobiliare utilizzano per formulare un giudizio di stima. In relazione allo scopo della stima e al procedimento adottato, i dati relativi ad un campione significativo possono essere un immobile la cui transazione di vendita/affitto è avvenuta di recente (prezzo storico), un immobile attualmente in vendita (prezzo richiesto), così come qualsiasi informazione contenuta nei dati catastali, nelle quotazioni di mercato, nelle indagini di mercato o nelle perizie effettuate da altri professionisti.

È chiaro che i diversi dati immobiliari disponibili, ovvero i prezzi o i redditi di mercato, possono dare un diverso livello di qualità informativa, da cui dipendono (e derivano) giudizi di stima più o meno robusti.

A questo proposito, un recente rapporto di TEGOVA (French, 2020) sottolinea come i comparabili giochino un ruolo cruciale nella qualità e nell'affidabilità delle valutazioni immobiliari. Lo scritto, inoltre, illustra come diversi Paesi europei stiano adottando legislazioni, strumenti e approcci molto diversi relativamente alle valutazioni immobiliari. Il rapporto TEGOVA distingue due livelli di informazioni da cui i professionisti possono desumere i dati per la stima, ovvero informazioni «hard» e «soft». Il termine «***hard information***» si riferisce ai dati storici relativi alla vendita e locazione di un immobile quando sono noti tutti gli aspetti economici (prezzo) e tecnici (dimensioni, stato di manutenzione, numero di stanze, piano, ecc.). Il reperimento di queste informazioni è legato a un documento di transazione immobiliare (atto di compravendita o locazione), solitamente depositato presso gli studi notarili, archivi notarili, la Conservatoria o l'Agenzia delle Entrate.

Al contrario, le «***soft information***» indicano i rapporti di società di ricerca, i dati delle agenzie, le banche dati pubbliche o le riviste specializzate. In questo caso, il valutatore non può conoscere come i dati siano stati raccolti e elaborati, ma piuttosto sono disponibili poche indicazioni molto generali circa la loro determinazione. Di solito, le informazioni *soft* prendono forma di dati di mercato aggregati, pur a differente livello (locale, regionale, nazionale).

Nella tradizione estimativa italiana, questa doppia classificazione potrebbe essere identificata attraverso l'articolazione in «**fonti dirette**» e «**fonti indirette**», anche se non c'è una totale convergenza tra le due classificazioni. Le fonti dirette comprendono anche i prezzi di immobili offerti nel mercato contenuti negli annunci di vendita e nelle inserzioni online. Al contrario, nel rapporto TEGOVA, gli *asking price* sono classificati come soft information, insieme alle informazioni provenienti da rapporti e ricerche di società che realizzano analisi di mercato.

Questa differenza tra la classificazione delle fonti informative in ciascun Paese è fortemente condizionata dalla disponibilità di dati puntuali, dalla legislazione locale e dal grado di trasparenza del mercato immobiliare.

Tuttavia, la qualità dei dati immobiliari non è analoga in ogni Paese (come ampiamente dimostrato dall'Indice di JLL), ma non lo è neppure all'interno di ogni Nazione, come nel caso delle diverse **informazioni disponibili** nelle grandi città rispetto a quelle di piccole dimensioni (Ionaşcu et al., 2021). Ciò implica che gli Standard di Valutazione Internazionali non dovrebbero essere eccessivamente prescrittivi nel codificare l'uso appropriato dei dati comparabili, poiché ogni fonte informativa può svolgere ruoli diversi a seconda del livello di trasparenza del mercato locale (Sadayuki et al., 2019).

2.4 La stima attraverso i prezzi richiesti *(asking price)*

I **prezzi richiesti** o di **prezzi offerta**[1] (in inglese, *asking prices*) sono i prezzi indicati dai venditori nella speranza di

[1] In italiano i termini "prezzi richiesti" o "di offerta" vengono impiegati in modo interscambiabile. In inglese gli "*asking prices*" (o anche, più raramente, *listing prices*) sono i prezzi richiesti dal venditore, mentre gli "*offer prices*" sono i prezzi proposti dall'acquirente al momento della trattativa e che riflettono la sua disponibilità a pagare per un dato bene.

attirare potenziali acquirenti disposti a pagare tale importo. Il prezzo viene individuato attraverso stime o analisi di mercato o, ancora, dagli stessi agenti immobiliari. Il prezzo di offerta è collocato all'inizio della vendita o della trattativa di vendita e non sempre è un prezzo definitivo, essendo soggetto, nella maggior parte dei casi, ad un **margine di trattativa**. Si tratta, quindi di un prezzo temporaneo e unilaterale. Solo il prezzo di compravendita, invece, si colloca alla fine della trattativa tra le due parti, acquirente e venditore, ed è un dato storico e immodificabile. I due prezzi sono, pertanto, diacronici.

I prezzi di offerta sono impiegati in diverse ricerche per analizzare i trend e le dinamiche del mercato immobiliare, e per la costruzione di indici (Pozo, 2009; Hayunga e Pace, 2016; Gordon e Winkler, 2016). Inoltre, alcuni studi hanno messo in relazione e confronto i prezzi di richiesta con i prezzi di vendita nei mercati immobiliari (Anglin et al., 2003; Beracha e Seiler, 2014; Knight, 2002). L'articolo di Curto et al. (2015) illustra come l'uso dei prezzi di offerta per la valutazione degli immobili sia molto diffuso in Italia, e tale diffusione sia provocata principalmente dalla scarsa accessibilità ai dati relativi ai prezzi di compravendita. Lo stesso studio rileva che l'utilizzo di tali valori può rendere più difficoltosa la valutazione degli immobili.

Nell'estate del 2017 si è svolto un animato dibattito tra studiosi e professionisti del mercato immobiliare sulla rivista online italiana «Monitor Immobiliare». Oggetto della discussione sono stati proprio gli *asking price* e il loro possibile uso nelle valutazioni. Le opinioni non sono convergenti: alcuni professionisti affermano che l'uso si rende necessario quando i prezzi richiesti non sono disponibili. Altri rifiutano l'uso del prezzo di offerta perché non può in alcun modo riflettere il valore di mercato dell'immobile, mentre altri ancora propongono una sorta di compromesso tra le due posizioni opposte. In generale, nella discussione viene sollevato il problema che tra ciò che è scritto nei manuali, nelle norme e nei regolamenti, e il mercato immobiliare vi sia molta distanza, per cui gli operatori devono trovare forme di adattamento necessarie alla loro "sopravvivenza" professionale. Infatti, nonostante la maggiore trasparenza del mercato e la costruzione di banche dati storiche, in Italia l'accesso ai dati sulle compravendite è tuttora limitato o estremamente costoso. Non mancano i difensori dell'impiego degli *asking price* nelle stime, adducendo come motivo il fatto che i prezzi di compravendita, essendo storici, riflettono solo il passato e precludano lo sguardo verso il futuro. Tuttavia, un acquirente, un investitore, è disposto a pagare una somma di denaro al momento della compravendita in grado di riflettere le attese future di valore e rendimento. Quindi i prezzi di compravendita sono in grado di riflettere le attuali aspettative rispetto all'evoluzione del mercato nel futuro. La valutazione è quindi la migliore stima del (futuro) prezzo di compravendita alla data della stima, e tale valore deve tenere conto del particolare ciclo immobiliare, dell'andamento della domanda e della sostenibilità di tale valore nel medio periodo. E per fare questo è necessario un adeguato patrimonio informativo.

È pratica diffusa dai valutatori e **professionisti italiani** l'uso dei **prezzi di offerta** individuati nei siti immobiliari specializzati. L'ampia reperibilità, il facile accesso e la gratuità sono tra le cause del largo impiego dei prezzi di immobili offerti in vendita. Gli Standard Internazionali, insieme a diversi regolamenti e norme, hanno indicato quali dati potrebbero essere impiegati nelle stime. In Italia, la Norma UNI del 2015 (UNI 11612:2015, 2015) ha ammesso, pur in alcuni casi residuali, l'impiego degli *asking price*. In particolare, nel caso di valutazioni di immobili in zone in cui non siano avvenute sufficienti transazioni, e non sia possibile individuare prezzi di compravendita recenti di un campione significativo analogo al bene da stimare, o che tali prezzi di compravendita individuati risultino inaffidabili, potranno essere impiegate le richieste di prezzi di immobili simili offerti in vendita. L'impiego e la rilevanza di tali informazioni deve essere chiaramente definita, analizzata criticamente e giustificata nel rapporto di valutazione fornito al Committente. Tale pratica, però, si estende oltre l'eccezionalità indicata dalla Norma.

Tuttavia, è importante sottolineare che anche i siti di vendita immobiliare disponibili online riflettono l'opacità connaturata al mercato italiano. A titolo di esempio, è possibile confrontare i dati disponibili online in Italia con quelli presenti nei siti, per esempio, di vendita immobiliare statunitensi. Da questi ultimi è possibile reperire informazioni di quali e quante volte un immobile è stato compravenduto, i diversi prezzi di compravendita, l'ammontare di tasse di proprietà o una stima dei costi di gestione. La disponibilità di dati è conseguenza di un sistema trasparente che permette alle informazioni contenute nel catasto, che, in questo caso, sono totalmente disponibili online a chiunque.

Di contro, gli atti di compravendita in Italia non sono fruibili telematicamente su larga scala dagli operatori del settore immobiliare e spesso le informazioni contenute in essi sono **scarse**. Anche interrogando i dati dell'Agenzia delle Entrate, nella sezione "Consultazione valori immobiliari dichiarati", è possibile disporre delle informazioni economiche (prezzo), ma i dati circa l'unità immobiliare compravenduta sono scarse (si hanno indicazioni limitatamente a dimensione, categoria catastale, e data dell'atto di compravendita).

Inoltre, anche se è possibile raccogliere una casistica di dati storici relativi a beni analoghi a quello di stima, non è possibile garantire la **veridicità** delle informazioni dichiarate negli atti di compravendita. La ricerca di campioni significativi di immobili compravenduti presso la Conservatoria può portare a **informazioni incomplete** o, peggio, **errate** relative alla compravendita, al prezzo o alle caratteristiche dell'immobile. Le caratteristiche quantitative e qualitative quindi, considerate come variabili esplicative del prezzo, sono un problema informativo nella raccolta dei prezzi reali di compravendita in Italia.

Gli *asking price* devono quindi essere considerati dati non adeguati perle stime immobiliari? Quello che è

certo è che un valutatore deve essere consapevole del livello di opacità che essi comportano.

Obiettivo di questa ricerca è l'analisi e la verifica della correttezza degli *asking price* qualora fossero impiegati come unica fonte nella stima del più probabile valore di mercato. In particolare, gli immobili comparabili e utilizzati per il *training* delle Reti Neurali sono rappresentati dagli annunci online raccolti su specifici siti web specializzati nella vendita di immobili in Italia.

Comunque, prima di esaminare se il prezzo richiesto possa essere un valore congruo a rappresentare il (futuro) prezzo di compravendita, è necessario capire quanto siano veritiere le caratteristiche descritte negli annunci pubblicitari, e come queste, se erroneamente descritte per diverse ragioni, possano influenzare il procedimento di stima del più probabile valore di mercato.

In che modo gli immobili descritti sommariamente in un annuncio immobiliare possono essere considerati dati attendibili per le stime?

Le consistenze, lo stato di manutenzione e le caratteristiche dell'immobile possono essere dedotte con un adeguato grado di affidabilità dagli annunci immobiliari?

Quali errori commettono i valutatori se utilizzano i prezzi richiesti dai venditori in luogo dei dati desunti dalle transazioni storiche?

Quali sono le principali **fonti di opacità** del mercato?

3. UN APPROCCIO METODOLOGICO

La metodologia di analisi adottata in questa ricerca integra procedure di programmazione informatica, tecniche di *machine learning* e valutazione immobiliare multi-parametrica, intendendo comprendere come e quanto l'opacità nel mercato immobiliare italiano influenzi le previsioni dei valori di mercato. Per un determinato mercato (caso studio), vengono definiti i seguenti passaggi.

- Nella prima fase viene sviluppato un **software di crawling (scansione) automatizzato** in linguaggio informatico Python®. Un *web crawler* è un software in grado di aprire e analizzare i contenuti delle pagine web in maniera strutturata e automatizzata. Si crea un *web crawler* nell'ambito della presente ricerca al fine di analizzare specifici siti web di vendita immobiliare in Italia e scaricare automaticamente gli *asking price* delle proprietà attualmente in vendita. Contestualmente agli *asking price* anche alcune caratteristiche descrittive degli immobili vengono raccolte tramite il medesimo *web crawler*. Grazie a questa procedura è possibile ottenere agilmente migliaia di dati. Ogni dato (osservazione) rappresenta un immobile in vendita di cui si conoscono il prezzo di offerta e alcune caratteristiche descrittive. Il database, raccolto tramite il *web crawler* per un dato mercato, è generalmente denominato $DB_{crawler}$.
- Nella seconda fase della ricerca, viene sviluppata una **rete neurale artificiale** (o *Artificial Neural Network*) a partire dal $DB_{crawler}$, e conseguentemente denominata $ANN_{crawler}$, con lo scopo di prevedere il valore di un immobile in funzione delle sue caratteristiche all'interno del dato mercato.

Nel campo dell'apprendimento automatico, le ANNs sono sistemi computazionali in grado di apprendere *pattern* e procedure (Ćetković et al., 2018). Le reti neurali sono costituite da neuroni artificiali e sinapsi artificiali: i neuroni sono le unità computazionali, mentre la sinapsi collegano i neuroni tra di loro. I neuroni artificiali sono organizzati in più *layer* separati in modo che i neuroni di input siano visualizzati nel *layer* di input e i neuroni di output siano contenuti nel *layer* di output. Tra di essi sono presenti vari *layer* di neuroni chiamati *layer* nascosti. Essendo una procedura di apprendimento automatico, una rete neurale può «**imparare**» come specifiche informazioni di input fluiscono dai neuroni di input per generare una risposta di output. Una rete può essere addestrata su qualsiasi database di informazioni strutturato nella forma input-output, producendo di conseguenza un modello predittivo (Pittarello et al., 2021). In questo articolo, i neuroni di input dell'$ANN_{crawler}$ sono le caratteristiche descrittive degli immobili, mentre il neurone di output rappresenta il suo corrispondente valore di mercato.

In particolare, le ANN qui assumono il ruolo di una **tecnica di valutazione multi-parametrica del valore di mercato**. Le ANN vengono infatti utilizzate per analizzare diverse caratteristiche dell'edificio e prevederne il valore di mercato. Nella funzione di previsione, ogni variabile indipendente (le caratteristiche dell'edificio) contribuisce in modo diverso alla stima del prezzo (la variabile dipendente) (Simonotti, 2006). In questo modo, sarà possibile isolare il contributo che ogni variabile apporta al prezzo e capire che tipo di errore nella stima potrebbe produrre un diverso livello di opacità delle informazioni per ogni caratteristica dell'immobile.

Il vettore colonna dei neuroni di input è chiamato $[X_r]_{crawler}$, con $1 \leq r \leq R$, dove R è il numero totale di osservazioni contenute nel database. Il neurone di output è un vettore a una colonna e una riga indicato come $[Y\_forecast]_{crawler}$. $[Y\_forecast]_{crawler}$ può anche essere rappresentato come $[Y\_forecast]_{crawler} = f([X_r]_{crawler})$, ovvero una funzione del vettore $[X_r]_{crawler}$ definita attraverso i layer nascosti.

Quando una rete viene addestrata (*training*) su un database che mette in relazione degli input a degli output, la rete «impara» come i dati di input sono correlati all'output corrispondente. Di conseguenza la rete ne costruisce un modello previsionale. Durante il *training* su un database, infatti, la rete modifica iterativamente i suoi parametri liberi, vale a dire i pesi (w) e i *biases* (b) della rete, fino a identificare il miglior modello previsionale.

Alla scala del singolo neurone, durante in training, ogni z^{esimo} neurone ottiene uno o più input numerici, qui chiamati $x_{z,u}$, con $1 \leq u \leq U$, dove U è il numero di sinapsi artificiali che entrano nel neurone. Quindi le in-

formazioni ricevute vengono tutte combinate all'interno del z^{esimo} neurone attraverso una funzione definita da un peso ($w_{z,u}$) e da un *bias* (b_z), in modo che venga prodotto un output numerico (Y_z). A questo punto, una funzione di attivazione (φ_z) trasforma il valore del neurone in un valore di risposta sotto forma di somma ponderata, così come definito attraverso l'**Equazione 1**:

$$\forall \text{ neurone } z^{esimo}, \quad Y_z = \varphi z[(\sum_{u=1}^{U}[w_{z,u} * x_{z,u}) + b_z]) \quad (1)$$

Durante l'addestramento della rete i pesi e i *biases* vengono quindi modificati iterativamente fino al raggiungimento del miglior modello di previsione o, in altre parole, fino a quando $[Y\_forecast]_{crawler}$ diventa il più simile possibile a $[Y\_expected]_{crawler}$. In questa ricerca $[Y\_expected]_{crawler}$ è il valore target, ovvero il valore di mercato contenuto nel database di *training* ($DB_{crawler}$), e $[Y\_forecast]_{crawler}$, come definito in precedenza, è il valore prodotto da $ANN_{crawler}$. La differenza tra il valore target meno la previsione produce un segnale di errore denominato $[err]_{crawler}$ come dall'**Equazione 2**:

$$[err]_{crawler} = [Y\_expected]_{crawler} - [Y\_forecast]_{crawler}. \quad (2)$$

Durante l'addestramento della $ANN_{crawler}$ attraverso il $DB_{crawler}$, l'obiettivo è quindi ridurre al minimo la somma dei segnali di errore prodotti.

- Nella terza fase, gli Autori intendono verificare il livello di trasparenza delle informazioni raccolte automaticamente tramite il *web crawler*. Infatti, oltre all'incertezza propria dell'uso degli *asking price* nelle valutazioni immobiliari, è possibile verificare che gli annunci di vendita contengono anche false dichiarazioni, informazioni errate o dati incompleti.
  Poiché il presente articolo mira a valutare come e quanto l'opacità nelle informazioni influenzi la corretta stima del valore di mercato dell'immobile, nella terza fase di sviluppo della ricerca, un **set di controllo** dei dati viene nuovamente raccolto, ma questa volta **a mano**. Per quanto sia estremamente dispendiosa in termini di tempo, questa procedura di raccolta manuale rappresenta l'unico modo per verificare la correttezza di tutte le informazioni contenute negli annunci di vendita, in modo tale da escludere gli annunci incompleti, i cui dati non possono essere verificati, completare le informazioni mancanti ma verificabili e correggere i dati errati. Questo set di controllo raccolto manualmente viene denominato DB_{hand}.
- Nella quarta fase, l'$ANN_{crawler}$ precedentemente sviluppata viene testata sul DB_{hand}. Questa volta, i neuroni di input sono le caratteristiche degli immobili raccolte e corrette a mano, quindi denominate $[X_r]_{hand}$, mentre il neurone di output è la conseguente previsione di un valore di mercato $[Y\_forecast]_{hand}$. Il valore di mercato atteso, o valore target, è il prezzo indicato negli annunci di vendita, denominato $[Y\_expected]_{hand}$. Chiaramente, $[Y\_expected]_{hand} = [Y\_expected]_{crawler}$. La differenza tra il valore di mercato atteso e la sua previsione rappresenta una stima dell'errore prodotto a causa dell'opacità informativa contenuta negli annunci online (opacità del mercato), così come illustrato nell' **Equazione 3:**

$$[err]_{hand} = [Y\_expected]_{hand/crawler} - [Y\_forecast]_{hand} \quad (3)$$

4. TRE CASI STUDIO RAPPRESENTATIVI NEL NORD ITALIA

4.1 Download dei database: *il web crawler*

Per questa ricerca è stato scelto di analizzare l'opacità informativa in tre diversi mercati italiani, ovvero i mercati immobiliari di:

- Bologna;
- Padova;
- Treviso.

Secondo l'opinione degli Autori, queste tre città potrebbero essere casi di studio rappresentativi perché incarnano tre differenti dimensioni di mercato, dove per dimensione di mercato qui ci si riferisce sia alle dimensioni stesse della città in analisi, sia al numero di elementi (osservazioni) che costituiscono la domanda e l'offerta, sia al volume totale delle transazioni. Inoltre, sebbene sia chiaro che ogni mercato immobiliare rappresenti un caso specifico e unico, nessuna di queste tre città descriva un caso fortemente a sé stante come, ad esempio, potrebbero essere i mercati di Venezia, Roma o Milano. I risultati possono quindi essere significativi anche per altri mercati immobiliari simili in Italia.

Il primo problema è stato capire come realisticamente raccogliere i dati e le informazioni necessarie per sviluppare le ANNs. Come già introdotto nel **Capitolo 3** è stato sviluppato un **software di *crawling* automatizzato** tramite il linguaggio di programmazione Python al fine di analizzare le proprietà immobiliari in vendita, a Bologna, Padova e Treviso nei siti web di vendita in modo da scaricarne automaticamente gli *asking price* e le caratteristiche.

La ricerca online del *web crawler* deve essere targettizzata attraverso la definizione di tre diversi **domini di ricerca**, dove ogni dominio è specifico per una città. I tre domini comprendono tutti gli immobili residenziali in vendita a Bologna, Padova e Treviso. Si sceglie di escludere da questo studio le proprietà non residenziali, come le proprietà commerciali e direzionali, così come anche i locali posti in affitto. Per quanto riguarda la ti-

pologia edilizia, sono incluse nel dominio di ricerca sia le nuove costruzioni che gli edifici esistenti, quali appartamenti, soffitte, villette a schiera e ville pluri/bi/unifamiliari.

Per il primo dominio di ricerca, tutte le 10 aree di Bologna sono incluse nella procedura di *download*, mentre il secondo dominio contiene tutte le 14 aree in cui è suddiviso il Comune di Padova. Infine, le 7 aree di Treviso sono altresì tutte comprese nel terzo dominio di ricerca.

Per consentire al *web crawler* di estrarre le informazioni da ogni annuncio di vendita, è necessario inserire all'interno del codice Python *l'Uniform Resource Locator* (URL) di ogni annuncio sotto forma di indirizzo web, come «https://...».

Dal momento che, chiaramente, aprire tutti gli annunci in vendita online per copiare e incollare i loro URL all'interno del codice Python sarebbe un'operazione manualmente impraticabile, l'obiettivo è capire come produrre gli URL in modo automatico.

È stato notato che ogni annuncio di vendita mostra un URL che è formato dalla combinazione dell'URL della home page più un numero di serie, dove la home page può essere definita come l'elenco di tutti gli annunci di vendita nel sito risultante dalla ricerca online all'interno dei dati domini.

Pertanto, il *web crawler* è stato programmato per estrarre prima l'URL della *home page*, poi identificare i numeri di serie degli annunci in essa contenuti, e infine creare automaticamente gli URL corrispondenti.

Successivamente, è stata implementata la libreria Python «Beautiful Soup» per sfogliare e analizzare tutte le pagine HTML degli annunci in vendita. «***Beautiful Soup***» è un pacchetto sviluppato da Leonard Richardson proprio al fine di analizzare i documenti in formato HTML. Con l'aiuto di questa libreria, è stato possibile estrarre dati e informazioni dai testi HTML, creando così un albero di analisi per tutte le pagine web identificate. È stata poi costruita in Python una **classe** di oggetti e funzioni per definire l'insieme di informazioni da estrarre da ogni annuncio. La classe utilizzata in ciascun dominio di ricerca è illustrata in **Tabella 1**.

La **classe** degli oggetti e delle funzioni sopra definiti è stata determinata sia in base ad una specifica analisi delle informazioni disponibili contenute nei siti di vendita immobiliare, sia in base alle caratteristiche degli edifici più comunemente utilizzate nelle procedure di stima **multi-parametrica** del valore di mercato (Feng e Zhu, 2017; Wang e Xu, 2018). Tra gli attributi di stima si considerano caratteristiche strutturali/fisiche dell'edificio, qualità del quartiere e localizzazione.

In seguito all'analisi delle pagine HTML degli annunci di vendita, è stata poi implementata in Python la libreria di analisi dei dati «***Pandas***». Sviluppata da Wes McKinney, la libreria «Pandas» viene utilizzata per estrarre un file in formato .xls a seguito di una procedura di scansione di siti web con lo scopo di visualizzare i dati e le informazioni estratte sotto forma di tabella. Ogni riga della tabella rappresenta una diversa osservazione, mentre le colonne indicano gli elementi della classe scaricati per ogni osservazione. Il database di Bologna, chiamato $DB(B)_{crawler}$, presenta 2.455 osservazioni, nel database scaricato per Padova, denominato $DB(P)_{crawler}$, ci sono 2.884 osservazioni, mentre il database di Treviso, definito come $DB(T)_{crawler}$, mostra 1.473 osservazioni.

Tabella 1 - *Classe di oggetti e funzioni*

Elemento o funzione della classe	u.d.m.	Elemento o funzione della classe	u.d.m.	Elemento o funzione della classe	u.d.m.
IDENTIFICAZIONE		PREZZO		Ascensore	binario
Web URL	testo	Prezzo	€	Giardino privato	binario
Numero annuncio	numero	Sup. commerciale	mq	Garage privato	binario
Codice identificativo	numero	Prezzo unitario	€/mq	Giardino comune	binario
LOCALIZZAZIONE		CARATTERISTICHE		Parcheggio	binario
Zona	testo	n. bagni	numero	Cantina	binario
Latitudine	coordinata	n. camere	numero	Terrazza	binario
Longitudine	coordinata	Piano min.	numero	Domotica	binario
TIPOLOGIA EDILIZIA		Piano max.	numero	Riscaldamento centralizzato	binario
Appartamento	binario	Ultimo piano	binary	Impianto fotovoltaico	binario
Soffitta	binario	CONDIZIONI		Ventilazione meccanica	binario
Casa a schiera	binario	Classe energetica	numero da(1 a 10)	Aria condizionata	binario
Villa pluri-familiare	binario	Manutenzione	numero (da 1 a 4)	Fibra ottica	binario
Villa bi-familiare	binario	Anno di costruzione	numero (anno)	Camino	binario
Villa singola	binario	ATTRIBUTI		Allarme	binario

4.2 Pulizia dei database: rimozione delle osservazioni incomplete e degli *outlier*

Dopo la procedura di *download*, è stato necessario pulire i tre *database* di *training* delle ANNs dalle osservazioni incomplete o evidentemente errate. Come prima cosa sono stati esclusi gli annunci incompleti. In un secondo momento sono poi state escluse le osservazioni contenenti errori evidenti o valori anomali (*outlier*), come, ad esempio, quando contenenti un prezzo di vendita nullo o una superficie commerciale nulla. In particolare, le percentuali di osservazioni incomplete/errate sono illustrate in **Tabella 2**, definite per ogni città ed elemento della classe. Pertanto, le percentuali espresse nella Tabella 2 rappresentano le osservazioni che sono state eliminate dal campionamento perché i dati non erano completi (o corretti) in tutte le classi di oggetti. Ad esempio, la «classe energetica» e l'«anno di costruzione» rappresentano la più alta percentuale di osservazioni perse a causa di informazioni mancanti.

Pertanto, i database di *training* scaricati via *web crawler* hanno dovuto essere ridotti, rispettivamente, a 1.665, 2.122 e 867 elementi.

Tabella 2 - *Percentuale di annunci persi per classe*

Classe	Percentuale di dati eliminati		
	Bologna	Padova	Treviso
Zona	1,37%	0,43%	5,06%
Latitudine	1,34%	2,77%	5,50%
Longitudine	1,55%	0,73%	5,57%
Appartamento	0,00%	0,00%	0,00%
Soffitta	0,00%	0,00%	0,00%
Villa pluri-familiare	0,00%	0,00%	0,00%
Villa singola	0,00%	0,00%	0,00%
Casa a schiera	0,00%	0,00%	0,00%
Villa bi-familiare	0,00%	0,00%	0,00%
Prezzo	2,44%	1,80%	2,65%
Sup. commerciale	1,96%	1,32%	0,41%
n. bagni	2,61%	3,16%	3,73%
n. camere	4,56%	3,47%	3,39%
Piano min.	2,32%	2,70%	5,84%
Piano max.	2,32%	2,70%	5,84%
Ultimo piano	1,18%	0,45%	0,20%
Classe energetica	24,52%	22,33%	28,45%
Manutenzione	4,56%	6,14%	7,06%
Anno di costruzione	19,85%	28,54%	21,90%
Ascensore	1,18%	0,45%	0,20%
Giardino privato	1,18%	0,45%	0,20%
Garage privato	1,18%	0,45%	0,20%
Giardino comune	1,18%	0,45%	0,20%
Parcheggio	1,18%	0,45%	0,20%
Cantina	1,18%	0,45%	0,20%
Terrazza	1,18%	0,45%	0,20%
Domotica	1,18%	0,45%	0,20%
Riscaldamento centralizzato	1,18%	0,45%	0,20%
Impianto fotovoltaico	1,18%	0,45%	0,20%
Ventilazione meccanica	1,18%	0,45%	0,20%
Aria condizionata	1,18%	0,45%	0,20%
Fibra ottica	1,18%	0,45%	0,20%
Camino	1,18%	0,45%	0,20%
Allarme	1,18%	0,45%	0,20%

5. LE TRE RETI NEURALI

5.1 Il *training* tramite l'algoritmo di ottimizzazione "Cuckoo"

Ripuliti i tre database ($DBs_{crawler}$) dalle osservazioni statisticamente non significative è stato possibile eseguire il *training* delle tre corrispondenti $ANNs_{crawler}$. I **set di *training*** impiegati per addestrare le reti neurali sono stati costituiti selezionando casualmente il 60% delle osservazioni dei $DB_{crawler}$ per ciascuna città. I **set di selezione** sono definiti prendendo in modo casuale un altro 20% delle istanze rimanenti, mentre i dati residui creano i rispettivi **set di test**. Sono state valutate tre diverse proporzioni di suddivisione dei database: 80%-10%-10%, 70%-15%-15% o 60%20%20%. La proporzione di suddivisione definitiva è stata poi scelta coerentemente al numero di osservazioni e input presenti nei dataset e nei modelli di rete neurale. In questo caso, i database sono abbastanza ridotti, tuttavia i numerosi neuroni di input rendono il modello molto complesso. Questo ha reso fondamentale mantenere un numero adeguatamente sufficiente di osservazioni nei set di selezione e test.

La procedura di *training* è stata sviluppata in codice Python e implementata separatamente per ogni città. I set di addestramento vengono inizialmente utilizzati per generare diversi modelli di ANN. In seguito questi modelli sono implementati sui set di selezione al fine di identificare quale ANN performi meglio anche su quest'ultimo set. Il set di test viene infine utilizzato per calcolare l'errore sulle previsioni.

Il processo di *training* viene eseguito all'interno di una procedura di ottimizzazione che consente di testare diversi modelli di rete neurale per ridurre al minimo l'errore sulle previsioni. L'algoritmo di ottimizzazione utilizzato durante il *training* delle reti è il **l'algoritmo di ottimizzazione "Cuckoo"** (Chiroma et al., 2017; Mareli e Twala, 2018). Esso è un algoritmo di ottimizzazione ispirato alla natura

che identifica e confronta tutte le possibili architetture delle ANNs che hanno un punto di ottimo relativo (cioè l'errore minimo) fino ad identificare quella che presenta l'ottimo globale. L'ottimizzazione tramite l'algoritmo "Cuckoo" è infatti adatta a risolvere i problemi di ottimizzazione globale poiché impiega il processo casuale dei "*random walks*".

Per semplicità, la rete neurale sviluppata per Bologna sarà indicata come $ANN(B)_{crawler}$, per Padova come $ANN(P)_{crawler}$, e per Treviso come $ANN(T)_{crawler}$. I risultati delle ANNs sviluppate sono presentati in **Tabella 3.** Gli errori medi prodotti sul set di test sono del 9,50% per Bologna, dell'11,75% per Padova e del 13,55% per Treviso.

Gli errori sono circa dello stesso ordine di grandezza. Tuttavia, possiamo notare una correlazione inversa rispetto alle dimensioni del mercato: più grande/attivo è il mercato minore è l'errore, e viceversa.

Tabella 3 - *Caratteristiche delle ANN*

NNs	N. di inputs	Output	Numero di layer nascosti	Numero di neuroni per layer nascosto	Funzione di attivazione	Scalarizzazione	Training
ANN(B)crawler	32	Valore di Mercato	7	184	ReLu	min-max	errore medio quadrato
ANN(P)crawler	32	Valore di Mercato	7	144	ReLu	min-max	errore medio quadrato
ANN(T)crawler	32	Valore di Mercato	7	144	ReLu	min-max	errore medio quadrato

5.2 Verifica della trasparenza informativa

Questa sezione è dedicata all'utilizzo delle $ANNs_{crawler}$ sviluppate in precedenza per discutere come la trasparenza del mercato (o meglio la sua opacità) influenzi l'affidabilità delle stime nel mercato immobiliare italiano.

A tale scopo, i tre database sono stati raccolti nuovamente, ma questa seconda volta a mano, per verificare la correttezza delle informazioni che in precedenza erano state scaricate automaticamente dal *web crawler* e utilizzate per costruire le reti. I database raccolti a mano sono chiamati $DB(B)_{hand}$ per Bologna, $DB(P)_{hand}$ per Padova, e $DB(T)_{hand}$ per Treviso.

Ovviamente, per definire i tre domini di ricerca web sono stati utilizzati gli stessi criteri di selezione già utilizzati per la precedente ricerca online. I domini si limitano così a cercare gli immobili residenziali in vendita (esclusi gli affitti e altri tipi di contratto), comprendendo sia edifici esistenti che nuove costruzioni.

Durante questa seconda raccolta dati viene verificata la correttezza di tutte le informazioni, e tutti gli immobili di cui non è possibile verificare le caratteristiche vengono direttamente esclusi dai DBs_{hand}.

Certamente questo metodo di raccolta dati si è rivelato un processo molto complesso e dispendioso in termini di tempo, ma, allo stesso tempo, l'unico possibile per testare e verificare la trasparenza del mercato, la veridicità dei dati e la disponibilità di informazioni.

I tre database DBs_{hand} sono stati implementati attraverso le $ANNs_{crawler}$ in modo che le **caratteristiche corrette** a mano degli immobili ne costituiscano i neuroni di input per poi produrre una nuova stima del loro valore di mercato. Come previsto dagli Autori, l'errore prodotto su queste ultime previsioni è superiore rispetto a prima. L'errore medio per Bologna è del 19,45%, per Padova è del 25,57% e per Treviso è del 26,97%. Questo aumento dell'errore sulle previsioni è dovuto all'opacità del mercato o, in altre parole, è dovuto alle informazioni errate riportate negli annunci. Più grande è il mercato immobiliare, minore è la percentuale di errore medio: i mercati più piccoli mostrano minore trasparenza nei dati descritti dagli annunci immobiliari.

La **localizzazione** si rivela essere la principale fonte di opacità. Solo alcuni tra gli annunci controllati mostrano la posizione esatta dell'edificio, mentre la maggior parte di essi sono posizionati in una strada o quartiere non corretti.

In generale questo problema si è rivelato più marcato a Padova e Treviso rispetto Bologna. Altri annunci, inoltre, sono stati proprio esclusi da questa analisi perché non fornivano informazioni sufficienti (immagini o descrizioni) per consentire l'esatta localizzazione della reale posizione dell'immobile. Certamente, le ragioni alla base di questa mancanza di trasparenza nella localizzazione dell'immobile possono risiedere sia nella privacy richiesta dai proprietari, sia nella strategia commerciale adottata dall'agenzia immobiliare. Infatti, diviene necessaria una intermediazione qualora un potenziale acquirente non sia in grado di identificare la posizione dell'immobile di interesse. Ciò nonostante, la mancanza di chiarezza riguardo la localizzazione rimane eccessiva e fuorviante, e diversi annunci specificano una posizione degli immobili che si è rivelata completamente sbagliata, confondendo persino aree centrali, semi-centrali e periferiche.

Un'altra importante fonte di opacità si rivela essere il **livello di manutenzione** poiché, spesso, non corrisponde alle altre informazioni riportate negli annunci stessi come, ad esempio, la classe energetica, l'anno di costruzione o gli impianti disponibili. In generale, le condizioni di manutenzione non sono coerenti con il vero stato in cui versa l'immobile. In alcuni casi le immagini mostrano livelli di manutenzione chiaramente molto peggiori di

quanto dichiarato negli annunci. Questa valutazione rimane comunque qualitativa e soggettiva, e gli Autori hanno ritenuto di modificare le condizioni di manutenzione dichiarate solo quando indubbiamente diverse dalle immagini fornite.

Altri errori spesso contenuti negli annunci riguardano l'identificazione e la definizione della **tipologia edilizia**, la disponibilità di una **cantina**, il **livello di piano**, la presenza di un **garage** e di un **giardino** privato/comune. In particolare, gli annunci o omettono queste informazioni oppure non forniscono dati coerenti. Gli annunci, infatti, che sono costituiti sia una descrizione verbale che da una sintesi in tabella, presentano alcune informazioni nella descrizione nettamente in contrasto rispetto quelle inserite in tabella.

Ancora, la definizione di **attico** rimane vaga e manca di trasparenza. La parola italiana utilizzata negli annunci per indicare un attico dovrebbe essere utilizzata per gli appartamenti di lusso e posti all'ultimo piano di un edificio. Tuttavia, in alcuni annunci, il termine «attico» è anche utilizzato per appartamenti e soffitte di categoria normale o economica. A causa dell'incertezza data da questo parametro, gli Autori hanno deciso di segnalare solo lo stato di «ultimo piano», senza alcun riferimento al livello qualitativo dell'immobile.

Altri errori riguardano gli **impianti**: l'aria condizionata o gli impianti di ventilazione meccanica sono spesso dichiarati presenti nella casa anche quando ne esiste solamente una predisposizione di tubi e condutture per una futura ipotetica installazione.

C'è poi anche un'altra area di opacità delle informazioni che non è direttamente correlata alle caratteristiche descrittive degli edifici. Gli annunci online subiscono un alto **tasso di modifica/decorrenza**. Di conseguenza, molti annunci sono stati rimossi, reintegrati o modificati in un arco temporale di poche settimane. Ciò ha reso necessari frequenti aggiornamenti e nuovi *download* dei database, e per questo motivo il *web crawler* automatizzato e sviluppato nell'ambito di questa ricerca si è rivelato estremamente utile. Tuttavia, anche i controlli manuali devono essere eseguiti più volte, aumentando significativamente il carico di lavoro totale.

Un'altra questione riguarda la **distribuzione asimmetrica** delle informazioni. Ad esempio, gli annunci per gli edifici di nuova costruzione sono generalmente più ricchi di dati rispetto al caso di edifici vecchi. Di conseguenza, è stato necessario scartare durante il controllo manuale un numero maggiore di annunci per edifici vetusti a causa della mancanza di informazioni cruciali. Per questo motivo, i tre database presentano una maggiore frequenza di annunci di offerte di edifici nuovi.

6. DEFINIRE L'IMPORTANZA DELLE VARIABILI

Nel paragrafo precedente, sono state discusse le fonti primarie di opacità. Tuttavia, è da considerare che non tutti gli errori contenuti negli annunci producono lo stesso impatto sulla stima del valore di mercato. L'opacità in alcune informazioni è infatti molto più significativa di altre. Per questo motivo, può essere utile determinare quali parametri di input mostrano il maggiore impatto sul valore di mercato attraverso una ***feature importance analysis*** (o analisi dell'influenza delle variabili).

Tra gli approcci che aiutano a calcolare l'impatto delle variabili sull'output ci sono i metodi *Filter-based*, i *Wrapper methods*, e gli *Embedded methods* (Tatwani e Kumar, 2019).

I metodi *Filter-based* si fondano sulla statistica univariata, come il coefficiente di correlazione di Pearson, il test di chi-quadro, il *Fisher's Score*, la *Variance Threshold*, la *Dispersion ratio* o la *Mean Absolute Difference*. I *Wrapper methods* impostano la selezione di un set di caratteristiche a guisa di un problema di ricerca (Ghosh et al., 2020; Suresh e Narayanan, 2019; Yassi e Moattar, 2014), quali la *Forward Feature Selectio*n, la *Backward Feature Elimination*, *l'Exhaustive Feature Selection* o la *Recursive Feature Elimination*. Infine, gli *Embedded methods* combinano le caratteristiche dei metodi *Filter* e *Wrapper*, quali ad esempio la LASSO *Regularization* e la *Random Forest*.

In questo articolo viene valutata l'influenza delle caratteristiche attraverso la metodologia della ***Random Forest*** (RF) o foresta casuale. È stato scelto questo approccio perché gli *Embedded methods* sono altamente precisi e mostrano eccellenti proprietà di generalizzazione (Siham et al., 2021).

Una foresta casuale è un particolare **classificatore** formato da un insieme di alberi decisionali (classificatori semplici) (Ugolini, 2014), dove un albero decisionale, nel campo dell'informatica, è una struttura dati costituita da nodi e archi. Un albero decisionale viene letto dall'alto verso il basso. I nodi dell'albero sono gli elementi che contengono le informazioni, mentre gli archi sono le connessioni tra i nodi. Il nodo iniziale è la radice e non ha archi di input, mentre i nodi terminali sono chiamati foglie e non mostrano archi in uscita.

Ogni albero decisionale, durante una procedura di foresta casuale, viene costruito (cioè addestrato) a partire da un sottoinsieme definito in maniera casuale del set di *training*. In questo caso, i tre set di *training* sono le banche dati di informazioni descrittive degli edifici raccolte rispettivamente per Bologna, Padova e Treviso. Ogni albero decisionale è costruito su un'estrazione casuale delle caratteristiche analizzate (cioè le caratteristiche degli immobili). Questa casualità nella selezione delle caratteristiche e delle osservazioni è una parte fondamentale durante la costruzione dei classificatori e ha lo scopo di aumentare l'eterogeneità e diminuire la correlazione tra variabili.

Per definire l'importanza di ogni caratteristica, è necessario misurare quanto diminuisce l'impurità di ogni caratteristica durante il *training*. Infatti, più una variabile diminuisce la propria impurità, più significativa risulta essere quella variabile. Nella classificazione (variabili discrete), l'impurità è data dall'impurità di Gini o dall'aumento/riduzione di entropia. Nelle regressioni (variabili continue),

invece, l'impurità è quantificata attraverso dalla varianza.

È stata definita una matrice colonna in cui le assi x rappresentano le caratteristiche degli immobili (le variabili) e l'asse y mostra l'obiettivo (il valore di mercato). È stata utilizzata in Python la libreria "**Numpy**" per eseguire il regressore-RF. Il regressore-RF è in grado di calcolare i coefficienti di importanza per ogni caratteristica. È stato utilizzato come set di addestramento il 70% delle osservazioni, mentre il restante 30% si impiega come set di test. Durante la procedura RF sono stati costruiti 2.000 alberi e il *threshold* impostato è pari allo 0,75 del valore medio dei coefficienti di importanza calcolati. In questo caso, la diminuzione dell'impurità di ogni caratteristica viene valutata come la media delle diminuzioni date da ciascun albero costituente la foresta. In questo modo viene stimata l'importanza finale di ciascuna variabile. I coefficienti di importanza calcolati dal regressore-RF sono illustrati in **Tabella 4**.

Supponiamo che un'informazione errata negli annunci riguardi i dati di maggiore impatto, come la latitudine e la longitudine (quindi la localizzazione), il livello di manutenzione o la superficie commerciale. In tal caso, verrà commesso un notevole errore nella previsione del valore di mercato. Al contrario, le variabili meno impattanti sono gli impianti e le tecnologie come la climatizzazione, la ventilazione meccanica, l'allarme, l'ascensore o la domotica. Tra le variabili meno impattanti ci sono anche il piano interrato, il giardino condominiale e il livello di piano.

Tabella 4 - *Coefficienti di importanza RF*

Elemento o funzione della classe	coefficienti RF		
	Bologna	Padova	Treviso
Variabile	%	%	%
Latitudine	22,43%	19,69%	15,38%
Longitudine	29,30%	19,52%	13,38%
Tipologia	0,75%	0,72%	1,05%
Sup. commerciale	8,17%	7,35%	7,71%
n. bagni	0,87%	0,93%	1,20%
n. camere	1,50%	1,12%	0,97%
Piano min.	2,03%	2,07%	4,02%
Piano max.	1,53%	1,63%	1,43%
Ultimo piano	0,61%	0,40%	0,37%
Classe energetica	3,19%	3,31%	6,99%
Manutenzione	14,76%	20,19%	25,80%
Anno di costruzione	7,55%	15,46%	10,03%
Ascensore	0,62%	1,00%	0,49%
Giardino privato	0,65%	0,35%	0,34%
Garage privato	0,50%	0,53%	0,99%
Giardino comune	0,53%	1,01%	4,11%
Parcheggio	0,61%	0,59%	0,51%
Cantina	0,81%	0,78%	0,52%
Terrazza	0,60%	0,53%	0,78%
Domotica	0,15%	0,09%	0,12%
Riscaldamento centralizzato	0,62%	0,40%	1,34%
Impianto fotovoltaico	0,08%	0,22%	0,29%
Ventilazione meccanica	0,00%	0,16%	0,23%
Aria condizionata	0,46%	0,29%	0,23%
Fibra ottica	0,59%	0,54%	0,50%
Camino	0,50%	0,45%	0,59%
Allarme	0,58%	0,68%	0,64%

7. DISCUSSIONE E CONCLUSIONI

Questo lavoro ha integrato l'analisi di mercato con tecniche multi-parametriche di stima del valore di mercato, programmazione informatica e procedure di apprendimento automatico per **comprendere l'opacità del mercato in Italia**.

In primo luogo, un *web crawler* automatizzato, sviluppato in linguaggio Python, ha aiutato a raccogliere rapidamente una notevole quantità di dati descrittivi riguardo gli immobili in vendita a Bologna, Padova e Treviso. Sulla base di questi tre database sono state addestrate in Python tre corrispondenti reti neurali artificiali al fine di prevedere il valore di mercato di una proprietà in funzione di 32 caratteristiche descrittive di input, tra cui, la localizzazione, il livello di manutenzione, gli impianti, la tipologia edilizia, la presenza di terrazza, garage o giardino. A questo punto i tre database sono stati raccolti una seconda volta manualmente. Questa procedura ha permesso di controllare la veridicità di ogni informazione indicata negli annunci. Le informazioni errate sono state corrette, mentre le osservazioni non verificabili sono state escluse. Questi tre database «corretti» sono stati implementati nelle reti neurali artificiali sviluppate in precedenza: l'errore prodotto sulle previsioni rappresenta l'errore nella stima del valore di mercato dovuto all'opacità del mercato stesso (o, in altre parole, l'errore dovuto alle informazioni non corrette contenute negli annunci). Infine, è stata eseguita un'analisi dell'influenza delle caratteristiche descrittive basata sulla metodologia *Random Forest*.

In conclusione, questa ricerca può aiutare a capire come e quanto l'opacità del mercato in Italia influisce sull'affidabilità delle stime immobiliari. Le reti neurali artificiali sono efficaci procedure statistiche di previsione e i modelli di rete neurale sono in grado di descrivere accuratamente qualsiasi relazione input-output. L'utilizzo di un modello di rete neurale per confrontare i risultati di un database opaco rispetto a un database «corretto» ha portato a determinante quanto un valutatore possa essere indotto a sbagliare la stima del più probabile valore di mercato di un immobile solo a causa della mancanza di trasparenza informativa.

Inoltre, questa analisi multi-parametrica ha anche permesso di identificare le variabili che più impattano sulla stima del valore di mercato. Questo è un aspetto cruciale in quanto i dati informativi dell'edificio devono essere controllati attentamente per valutare in modo corretto il valore di mercato dell'immobile. Chiaramente, una maggiore opacità nelle variabili più impattanti porterà a un errore più elevato nella previsione, mentre una minore trasparenza nelle variabili meno influenti produrrà un errore minore nella stima.

Al termine di questa ricerca gli Autori indicano la necessità di migliorare la condivisione dei dati e delle informazioni sulle proprietà immobiliari in Italia. Non solo i prezzi di compravendita dovrebbero essere resi disponibili, ma anche le descrizioni degli immobili dovrebbero essere molto più precise e complete. Inoltre, gli annunci di vendita potrebbero richiedere un livello minimo di dati da fornire prima di essere considerati pronti ad essere messi online. Ancora, le informazioni fornite negli annunci dovrebbero essere precise e corrette, soprattutto per quanto riguarda la localizzazione della proprietà. Infine, gli annunci dovrebbero essere più trasparenti: anche siti web diversi tra loro potrebbero comunque rispettare un *layout* condiviso per garantire completezza e chiarezza dell'informazione.

Lo scopo è duplice: in primo luogo, ridurre l'asimmetria informativa tra venditore e acquirente per far sì che la domanda possa essere in grado di muoversi in maniera più consapevole all'interno del mercato immobiliare. In secondo luogo, lo scopo è aumentare la trasparenza del mercato, poiché tali dati sono utilizzati sia da società che analizzano i mercati, producendo relazioni e pubblicando prezzi, sia dai valutatori che sono talvolta costretti a basarsi sugli *asking price* in quanto i prezzi di compravendita non sono disponibili. Inoltre, poiché il mercato immobiliare è complesso e ha implicazioni sostanziali per il resto dell'economia, tutti gli operatori dovrebbero essere certi della qualità delle informazioni che utilizzano.

La problematica affrontata in questa ricerca ruota attorno al concetto di «**qualità dell'informazione**». Naturalmente, un'indagine di mercato non può prescindere dall'analisi approfondita di ogni dato di riferimento, che si tratti di un acquisto, una vendita o un'offerta. Oltre a questo, sembra poi utile concentrarsi anche sulle procedure di stima oltre che sul tipo di dato: una metodologia appropriata consente di approcciare in modo professionale e costruttivo anche informazioni spurie, traendone intuizioni significative.

Per gli ulteriori sviluppi di questa ricerca, gli Autori intendono applicare periodicamente la metodologia adottata in altri mercati immobiliari italiani al fine di mappare i diversi livelli di opacità nei mercati e per capire se vi sia un'evoluzione nel tempo per quanto riguarda l'accesso alle informazioni. In particolare, potrebbe essere interessante comprendere se le dinamiche che si sono verificate negli ultimi anni (pandemia di Covid-19, guerra in Ucraina, crisi energetica) possono, in qualche modo, impattare non solo sulle dinamiche e sui prezzi degli immobili ma anche sul livello di trasparenza o opacità.

* **Laura Gabrielli,** *Dipartimento di Culture del Progetto, IUAV Università di Venezia*
e-mail: *laura.gabrielli@iuav.it*

** **Aurora Greta Ruggeri,** *Dipartimento di Culture del Progetto, IUAV Università di Venezia*
e-mail: *aurora.ruggeri@iuav.it*

*** **Massimiliano Scarpa,** *Dipartimento di Culture del Progetto, IUAV Università di Venezia*
e-mail: *massimiliano.scarpa@iuav.it*

Nomenclatura

$[err]_{crawler}$ - errore in $ANN_{crawler}$

$[err]_{hand}$ – errore dal DB_{hand}

$[X_r]_{crawler}$ – vettore colonna dei neuroni di input di $ANN_{crawler}$

$[X_r]_{hand}$ - vettore colonna dei neuroni di input da DB_{hand}

$[Y\_expected]_{crawler}$ - valore target in $ANN_{crawler}$

$[Y\_expected]_{hand}$ - valore target da DB_{hand}

$[Y\_forecast]_{crawler}$ vettore una a riga e una colonna del neurone di output $ANN_{crawler}$

$[Y\_forecast]_{hand}$ – neurone di output da DB_{hand}

$ANN(B)_{crawler}$, - ANN sviluppata da $DB_{crawler}$ per Bologna

$ANN(P)_{crawler}$, - ANN sviluppata da $DB_{crawler}$ per Padova

$ANN(T)_{crawler}$ - ANN sviluppata da $DB_{crawler}$ per Treviso

ANN/ANNs - Artificial Neural Network/Artificial Neural Networks

$ANN_{crawler}$ - Artificial Neural Network sviluppata da $DB_{crawler}$

b_z - bias

$DB(B)_{crawler}$ - database raccolto tramite web crawler per Bologna

$DB(B)_{hand}$ - database raccolto a mano per Bologna

$DB(P)_{crawler}$ - database raccolto tramite web crawler per Padova

$DB(P)_{hand}$ - database raccolto a mano per Padova

$DB(T)_{crawler}$ - database raccolto tramite web crawler per Treviso

$DB(T)_{hand}$ - database raccolto a mano per Treviso

$DB_{crawler}$ - database raccolto tramite web crawler

DB_{hand} - database raccolto a mano

RF - Random Forest

U - numero di sinapsi artificiali entranti nel neurone

$w_{z,u}$ - funzione dei pesi nel neurone

$x_{z,u}$ - input numerico nel neurone

Y_z - output numerico nel neurone

φ_z - funzione di attivazione nel neurone

Bibliografia

Akerloff G.A., *The Market for "Lemons": Quality Uncertainty and the Market Mechanism,* The Quarterly Journal of Economics, 84(3), 1970, pp. 488–500.

Anglin P.M., Rutherford R. and Springer T.M., *The trade-off between the selling price of residential properties and time-on-the-market: the impact of price setting*, The Journal of Real Estate Finance and Economics, Vol. 26, No. 1, 2003, pp. 95–111.

Arnott R., *Economic Theory and Housing. In Handbook of Regional and Urban Economics*, edited by E. Mills, London: Elsevier, 1987, pp. 959–988.

Beracha E. and Seiler M.J., *The effect of listing price strategy on transaction selling prices*, The Journal of Real Estate Finance and Economics, Vol. 49, No. 2, 2014, pp. 237–255.

Ćetković J., Lakić S., Lazarevska M., Žarković M., Vujoević S., Cvijović J. and Gogić M., *Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application*, Complexity, Vol. 2018, Article ID 1472957, 2018, p. 10.

Chiroma H., Herawan T., Fister I., Fister I., Abdulkareem S., Shuib L., Hamza M.F. et al., *Bio-inspired computation: Recent development on the modifications of the cuckoo search algorithm*, Applied Soft Computing, Vol. 61, 2017, pp. 149–173.

Curto R., Fregonara E. and Semeraro P., *Listing behaviour in the Italian real estate market*, International Journal of Housing Markets and Analysis, Vol. 8, No. 1, 2015, pp. 97–117.

Farzanegan M.R. and Fereidouni H.G., *Does real estate transparency matter for foreign real estate investments?*, International Journal of Strategic Property Management, Vol. 18, No. 4, 2014, pp. 317–331.

Farzanegan M.R., Gholipour H.F., *Does real estate transparency matter for foreign real estate investments?* Int. J. Strateg. Prop. Manag. 18 (4), 2014, pp. 317–331.

Feng J. and Zhu J., *Nonlinear regression model and option analysis of real estate price*, Dalian Ligong Daxue Xuebao/Journal of Dalian University of Technology, Vol. 57, No. 5, 2017, pp. 545–550.

Forte C. and De Rossi B., *Principi Di Economia Ed Estimo*, Etas., Milan, 1974.

French N., *Pricing to Market. An Investigation into the use of Comparable Evidence in Property Valuation*, TEGoVA The European Gropu of Valuers' Association, June, 2020.

French N., Crosby N. and Thorne C., *Pricing to market: market value - the enigma of misunderstanding*, Journal of Property Investment and Finance, Vol. 39, No. 5, 2021, pp. 492–499.

French N., Gabrielli L., *Pricing to market: Property valuation revisited: the hierarchy of valuation approaches, methods and models*, Journal of Property Investment & Finance, Vol. 36, No. 4, 2018, pp. 391–396.

Gholipour F.H., Tajaddini R., Pham T.N.T, *Real estate market transparency and default on mortgages* Research in International Business and Finance 53 10120, 2020.

Gholipour F.H., Masron A.T., *Real estate market factors and foreign real estate investment*. J. Econ. Stud. 40 (4), 2013, pp. 448–468.

Ghosh M., Guha R., Sarkar R. and Abraham A., *A wrapper-filter feature selection technique based on ant colony optimization*, Neural Computing and Applications, Vol. 32, No. 12, 2020, pp. 7839–7857.

Gordon B.L. and Winkler D.T., *The effect of listing price changes on the selling price of single family residential homes*, The Journal of Real Estate Finance and Economics, 2016, pp. 1–31.

Guerrieri G., *L'informazione per l'efficienza e la trasparenza del mercato immobiliare: l'esperienza italiana*, Territorio Italia, n. 1, 2011, pp. 88–102.

Hayunga D.K. and Pace R.K., *List prices in the US housing market*, The Journal of Real Estate Finance and Economics, 2016, pp. 1–30.

International Valuation Stardard Council IVSC, International Valuation Standars, London, 2020.

Ionaşcu E., Anghel I., *Improvement of the real estate transparency through digitalisation*, Proceedings of the International Conference on Business Excellence Vol. 14(1), July, 2020, pp. 371–384.

Ionaşcu E., Taltavull de La Paz P. and Mironiuc M., *The Relationship between Housing Prices and Market Transparency. Evidence from the Metropolitan European Markets*, Housing, Theory and Society, Vol. 38, No. 1, 2021, pp. 42–71.

John Lang Lasalle, *Global Real Estate Transparency Index, 2022 - Transparency in an age of uncertainty*, Real Estate Transparency Report, available at: www.joneslanglasalle.com (accessed July, 2022).

Lindqvist S., *The concept of transparency in the European Union's residential housing market: A theoretical framework*, International Journal of Law in the Built Environment, Vol. 4, 2012, pp. 99–115.

Mareli M. and Twala B., *An adaptive Cuckoo search algorithm for optimisation*, Applied Computing and Informatics, Vol. 14, No. 2, 2018, pp. 107–115.

Newell G., *The changing real estate market transparency in the European real estate markets*. J. Prop. Invest. Financ. 34 (4), 2016, pp. 407–420.

Pittarello M., Scarpa M., Ruggeri A.G., Gabrielli L. and

Schibuola L., *Artificial Neural Networks to Optimize Zero Energy Building (ZEB) Projects from the Early Design Stages*, Appllied Sciences, 11, 2021, p. 5377.

Pozo A.G., *A nested housing market structure: additional evidence*, Housing Studies, Vol. 24, No. 3, 2009, pp. 373–395.

Razali M.N. and Adnan Y.M., *Transparency in Malaysian Property Companies*, Property Management, 30(5), 2012, pp. 398-415.

Sadayuki T., Harano K. and Yamazaki F., *Market transparency and international real estate investment*, Journal of Property Investment and Finance, Vol. 37, No. 5, 2019, pp. 503–518.

Schulte K.-W., Rottke N. and Pitchke C., *Transparency in the German real estate market*, Journal of Property Investment and Finance, Vol. 23, No. 1, 2005, pp. 90–108.

Siham A., Sara S. and Abdellah A., *Feature selection based on machine learning for credit scoring : An evaluation of filter and embedded methods*, International Conference on INnovations in Intelligent SysTems and Applications (INISTA), 2021, pp. 1–6.

Simonotti M., *Metodi Di Stima Immobiliare*, Flaccovio, Palermo, 2006.

Suresh S.M.S. and Narayanan A., *Improving Classification Accuracy Using Combined Filter+Wrapper Feature Selection Technique*, IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1–6.

Tatwani S. and Kumar E., *Parametric comparison of various feature selection techniques*, Journal of Advanced Research in Dynamical and Control Systems, Vol. 11, No. 10 Special Issue, 2019, pp. 1180–1190.

Ugolini M., *Metodologie di apprendimento automatico applicate alla generazione di dati 3d*, 2014, aviable at https://amslaurea.unibo.it/10415/.

UNI 11612:2015, *Determination of the market value of properties*, 2015.

UNI/PdR 53:2019, *Real estate market analysis - Guidelines for identifying the market segment and collecting real estate data*, 2019.

Wang A. and Xu Y., *Multiple linear regression analysis of real estate price*, in IEEE (Ed.), International Conference on Robots and Intelligent System, ICRIS 2018, Changsha (China), 2018, pp. 564–568.

Yassi M. and Moattar M.H., *Robust and stable feature selection by integrating ranking methods and wrapper technique in genetic data classification*, Biochemical and Biophysical Research Communications, Vol. 446, No. 4, 2014, pp. 850–856.

Yun L. and Chau K.W., *The impact of real estate market transparency on the linkages between indirect and direct real estate*, paper presented at ERES Conference, Vienna, July, 2013, pp. 3–6.