Location-based data driven model for real estate market value analysis based on energy performance certification

Location-based data driven model for real estate market value analysis based on energy performance certification

Nicola Moretti[1], Lavinia Chiara Tagliabue[2], Mario Claudio Dejaco[1], Fulvio Re Cecconi[1]

[1]Department of Architecture, Built Environment and Construction Engineering, Politecnico di Milano, Via Ponzio 31, 20133 Milano, Italy

[2]Department of Civil, Environmental, Architectural Engineering and Mathematics, University of Brescia, Via Branze 43, 25123 Brescia, Italy

fulvio.rececconi@polimi.it

Abstract. The market value of physical assets varies over time as a function of different factors. These factors are related, for instance, to the correct execution of construction works and installation of systems, the procedures for the operation of systems and equipment, the continuous implementation of maintenance operations and to the location and the perceived value. However, there is no clear evidence, in Italy, of how buildings' market value is affected by the energy performance measured by the Energy Performance Certification (EPC), promoted thanks to the European Energy Performance of Buildings Directive in order to drive the assets' energy efficiency. The aim of the research is the definition of a methodology for the identification of the impact of the energy performance on the market value of the assets. An analysis of the market values of assets has been carried out based on a public database. On this basis, an investigation of homogeneous assets' price areas has been performed. The results have been geolocated and further processed, which allowed to associate differences in the market values to the energy performance retrieved from the *Certificazione Energetica degli EDifici* (CENED) database (the Italian regional EPC database). The buildings included in the analysis are residential buildings distributed in the regional territory (Lombardy Region). The methodology has been tested at national level but is potentially scalable in different contexts. It has been developed and applied in two case studies, cities with different size and characteristics located in the northern Italy.

1. Introduction

The real estate evaluations in Italy are often mainly based on the evaluation of the attractiveness of some specific areas, as a function of the proximity to the city center and to a set of services compared to peripheral areas. This results in a price difference of the assets (in this case the building), especially when the location of the assets is characterized by a low quality of life both perceived and actual (e.g. satellite neighborhoods, social housing, etc.). The market values of the properties are therefore referred to specific urban areas (historic center; semi-center; suburbs etc.). In addition, the prices related to the different homogeneous areas vary according to the condition of the building, generally defined in a qualitative way as new or well maintained, in average condition or old without maintenance. For each municipality in Italy there are databases in which position, conditions and average value of the buildings in specific areas are correlated. For the existing buildings, the availability of economic incentives for

systems and envelope upgrade leading to remarkable energy savings, led owners to carry out retrofit interventions in compliance with the recent European Energy Performance of Buildings Directive (EPBD) [1][2][3]. One of the consequences of these recent directives is the creation of a local/national geolocated database concerning the Energy Performance Certification (EPC) of buildings. The EU directives and budget limitations of small/medium investors seem to lead the market towards a stronger interest in the energy performance of the assets. However, this economic sector is not very dynamic, thus it is hard to promote new practices, even though they have been introduced a decade ago. Since the databases of real estate values [4] and energy certification of buildings are geolocated, there is an opportunity to verify whether the changes in market values are affected by the new the EPC [5]. The aim of this research is the development of a methodology for the evaluation of the impact of the EPC on the market value of the residential buildings in Italy. This methodology has been developed thanks to the availability of open data on EPC and market values, enabling the implementation of a Principal Component Analysis (PCA): a method able to determine the impact of a set of variables on the dataset's trend.

2. Methods

The development of the research methodology is enabled by the availability of public market data concerning the buildings' prices and EPC. Data availability promoted a cross evaluation of the assets, locating them in the territory to evaluate the weight of the location and energy quality, influencing the asset costs. The methodological approach includes two crucial steps: the first one is the data cleaning, carried out in order to improve the data reliability used to apply the PCA; the second is the geo-clusterization of the data enabling a geographical visualization of the data and unveiling the correlation between cost, location, energy quality. The location-based data has been employed for the organisation and visualisation of the assets' information in the analysed territories to propose a replicable and scalable methodology applicable to different locations and territory extensions.

2.1. Data collection and data cleaning

The first step concerns the acquisition of the market value data for residential buildings. Italian buildings' market values can be obtained and related to specific relevant areas (clusters). These data are consolidated and periodically updated by the local Chamber of Commerce. Moreover, the energy certifications data of new buildings or renovated buildings have been retrieved from the regional EPC database CErtificazione ENergetica degli EDifici (CENED) [6]. The EPC provides the EPh (primary energy for heating) of the individual building, the description of the spatial and dimensional features (i.e. number of rooms, uses, gross and net volume, surfaces, orientations, etc.) as well as the address. These data can be related to the clusters of market values. However, it is necessary to verify the data quality of CENED. For the real estate market (and for Italian laws) it is required to provide CENED data, for example in transactions of the leasing phase of the assets, however, there is no indication of the operator who processed/entered data, therefore misleading information is possibly recorded in the database. For this reason, checking EPh values in the database, it is possible identify clearly incorrect values (e.g. EPh values attributed to incongruent classes) that must be removed. This operation led to a database equal to about 75% of the initial total CENED, providing an acceptable quality level which allows the implementation of the following methodological steps.

2.2. Correlation between cost for dimensional unit and energy quality based on EPh value

It is clear that the most influential factor in the property price could be identified as its location, thus in order to limit the effect of the location on the Price/EPh ratio, the second phase of the research concerns the calculation of the significant correlation between cost per square meter (€/m^2) and EPh value (kWh/m^2 year) within the homogeneous geographic cluster defined using an unsupervised neural network. In the context of social sciences (i.e. where phenomena in which the human variable has a noteworthy influence are analysed), two variables are considered well correlated when their correlation

index (Pearson) is greater than the value 0.4 (or less than -0.4 if the correlation is inverse). This phase of investigation allowed to select the clusters adopted to develop the following phase of research.

2.3. Principal Component Analysis

The Principal Component Analysis (PCA) has been adopted to define the main influential components on the price of the buildings. The PCA calculation concerns the third phase of the proposed methodological approach. This method is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Depending on the application, PCA is named differently, for example in mechanical engineering it is called proper orthogonal decomposition (POD). PCA is used by almost all scientific disciplines [7] and for many purposes, from analyzing signals [8] to the identification of parameters in elastoplastic material models [9] to solving heat transfer problems [10]. Large or massive data sets are increasingly common and often include measurements on many variables. Despite the constantly increasing computational power of modern computers, it is frequently necessary to simplify the analysis of these data sets. PCA may allow to remarkably reduce the number of variables, while preserving much of the information in the original data set [11]. In order to achieve its goal, reducing the number of variables, PCA computes new variables called principal components which are obtained as linear combinations of the original variables. The first principal component is required to have the largest possible variance. The second component is computed under the constraint of being orthogonal to the first component and to have the largest possible inertia. The other components are computed likewise [7]. The PCA procedure is based on the following four steps:

1. The data is normalised by reducing each column to an average of 0 and standard deviation 1 (1):

$$x_{std,i} = \frac{(x_i - \bar{x})}{\sigma_x} \tag{1}$$

Where: $\bar{x}$ is the mean of X, x_i is the i value of X and σ_x is the standard deviation of X.

2. The covariance matrix (2) is calculated. Covariance provides a measure of the strength of the correlation between two or more sets of random variates. The covariance for two random variates X and Y, each with sample size N, is defined by the expectation value:

$$Covariance(X,Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x}) - (y_i - \bar{y})}{N} \tag{2}$$

Where: $\bar{x}$ is the mean of X, x_i is the i value of X, $\bar{y}$ is the mean of Y and y_i is the i value of Y. If the variables are correlated in some way, then their covariance will be nonzero, otherwise, if the variable are uncorrelated the covariance is zero. In fact, if Covariance(X,Y)>0, then Y tends to increase as X increases, and if Covariance(X,Y)<0, then Y tends to decrease as X increases. For a matrix m with p columns, Covariance[m] is a p×p matrix of the covariances between columns of m.

3. The eigenvectors and eigenvalues of the covariance matrix are calculated. Eigenvectors represent the main components and eigenvalues corresponding to the contribution (weight) of the eigenvectors in describing the variability of the data. The eigenvalues of a matrix m are the values λ_i for which one can find nonzero vectors v_i such that $m\ v_i = \lambda_I\ v_i$. The eigenvectors are the vectors v_i.
4. Only the main components are considered.

3. Case study and results

Two cities have been adopted as case studies. The two cities are located in northern Italy, in the Lombardy Region and they have similar climate conditions: the first is Varese (approx. 81,000 inhabitants), the second is Brescia (approx. 197,000 inhabitants).

3.1. Database analysis

The analysed data concerned 1.054 samples for the city of Varese and 869 for the city of Brescia, for which a breakdown by energy performance class was carried out. The analysis of these first data shows that about 74% of the samples of Varese have poor energy performance (13% in class F; 61% in class G, the lowest), highlighting that this geographic area could present a substantial share of historical heritage buildings or old construction. Concerning Brescia, about 55% of the samples have poor energy performance (15% in class F; 40% in class G), but it is worthy to note that about 11% of the buildings are in class B: these are most likely recent buildings, in compliance with the new requirements in terms of energy consumption. The samples seem to have a limited reliability although they come from the open and public regional database. Some values, obviously wrong, have been eliminated (about 15%) but there is still the possibility of incorrect data.

3.2. Correlation between price per square meter and EPh

In order to limit the effect of the location in the calculation procedure of the Price/EPh ratio, the correlation between the price per square meter and the EPh of the assets has been calculated within the same geographical cluster. The results of this correlation are represented in Table 1 where the green colour highlights the optimal correlations.

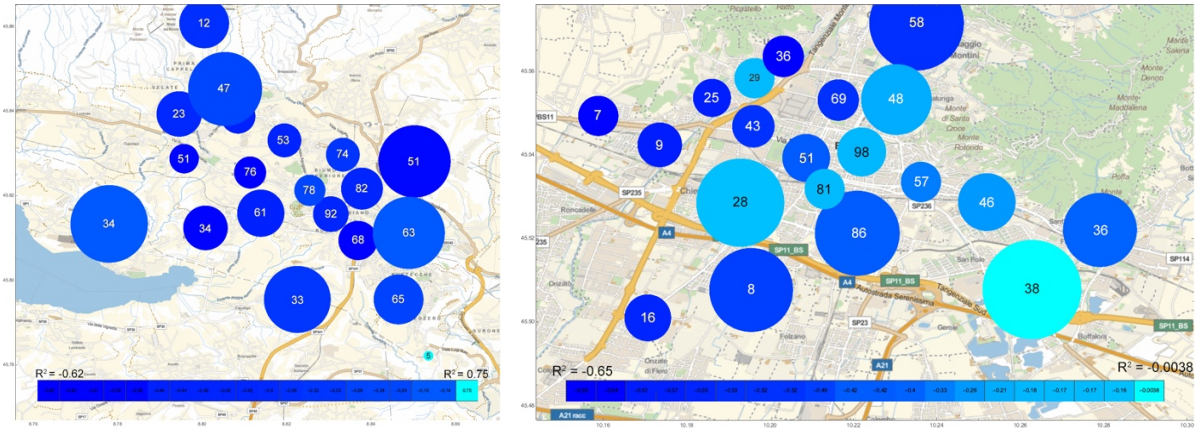
Table 1. Pearson correlation index between cost [€/m^2] and EPh [kWh/m^{2}year] for the clusters in the two cities.

N° samples	76	68	34	51	51	61	52	53	85	92	23	12	74	36	65	78	52	63	34	5
Varese	-0,61898	-0,6154	-0,59959	-0,58917	-0,58133	-0,43924	-0,43739	-0,42435	-0,41961	-0,41539	-0,40329	-0,36486	-0,33329	-0,33036	-0,2563	-0,24324	-0,24322	-0,14555	-0,14262	-0,754978
N° samples	36	7	9	25	16	58	69	43	8	51	86	36	57	46	48	98	81	29	28	38
Brescia	-0,6458	-0,6381	-0,57331	-0,56849	-0,55288	-0,52639	-0,52167	-0,52102	-0,48621	-0,41641	-0,41628	-0,39827	-0,33019	-0,26065	-0,21079	-0,18325	-0,1737	-0,17347	-0,16078	-0,0038

The data of Table 1 can be associated with the geographical location as shown in Figure 1. Figure 1 shows the clusters on a map, which represents the correlation between market values and EPh of the assets, indicating the number of the samples in the cluster. The color of the circles refers to the correlation which is higher for darker colors. Figure 2 shows the correlation asset cost/energy quality developed for the single cluster where it is possible to validate the correlation index: for higher price/cost the primary energy is lower in both case studies.

3.3. PCA application

In the context of this research, 7 variables have been considered as input data. These variables are: WGS X; WGS Y; Surface; N° of rooms; N° of bathrooms; EPh [kWh/m^{2}y]; Price (€). Through the application of the PCA method as presented in Paragraph 2.3, the variables are reduced to 3 components representing more than the 81% of the variability of the initial data for the city of Varese (Component 1: 46.5%; Component 2: 20.1%; Component 3: 14.8%) and 77% for the city of Brescia (Component 1: 44.3%; Component 2: 17.9%; Component 3: 14.6%. This means that the seven input variables can be traced back to the three main components which, in the present research, can also be physically interpreted according to the contribution of the seven initial variables: Component 1: consists mainly of Surface, N° of rooms and bathrooms, Price. We can say that it is representative of the "consistency of the property"; Component 2: consists mainly of Cartesian coordinates. We can say that it is representative of the "location of the building"; Component 3: it is almost entirely made up of the EPh. We can say that it is representative of the "energy quality of the building".

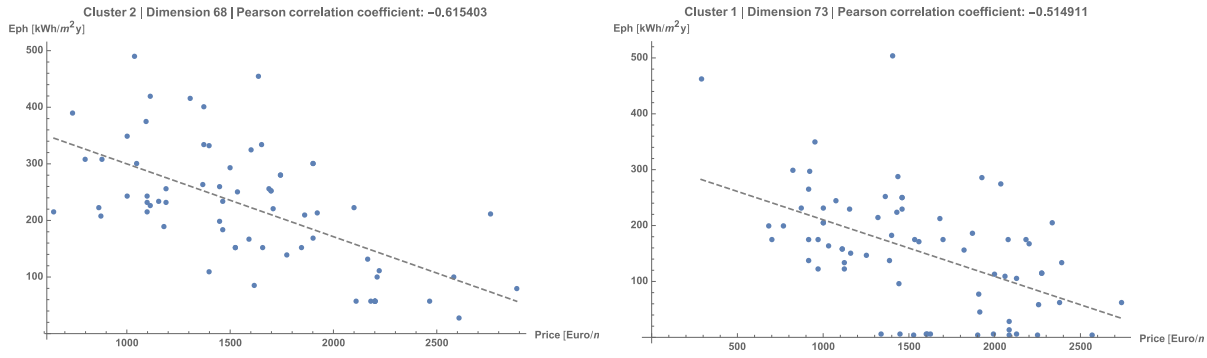


Varese Brescia

Figure 1. Correlation between price and EPh for Brescia and Varese. Each cluster is represented by a circle. The color represents the correlation (the darker it is the higher is the Price/EPh correlation) and radius is proportional to average price. The inner number refers to the number of samples.

Therefore, analyzing table 2 and figures 1 and 2, it can be stated that:

- Thanks to the presented methodology it is possible to identify homogeneous clusters of buildings which present a remarkable correlation price/eph
- for the city of Varese, the correlation price/EPh is mostly valid and strongly evident;
- for Brescia it is possible to identify how the city center and the social housing settlements might have a lower correlation due to low performance or low cost or, on the contrary, high costs and high performance due to refurbishments or recent developments;
- thanks to the PCA it is possible to better investigate the assets dataset identifying three vectors grouping different variables responsible for the performance of the sample of assets.



Varese Brescia

Figure 2. Correlation between cost/energy quality for two clusters.

4. Discussion and conclusions

The methodology deployed in the present research has been developed to investigate how EU policies implemented in the last 10 years have impacted on the Italian market of residential buildings. The results demonstrate that there is a significant impact of the energy performance on the price of the buildings. The methodology is original, and the use of location of buildings allows to understand the specificity of the territory and mainly to locate the strategies and interpret the results. Further parameters representing the energy performance of the buildings could be encompassed in the assessment, in order to achieve a more precise energy profile of the buildings included in the analysis. The energy certifications are not always precise and a data cleaning process has to be performed before the analysis. The first findings

that can be highlighted are related to the effectiveness of the EU energy policies which introduced the energy certification to drive the asset market.

Table 2. First three eigenvectors of PCA or the two case studies.

	Varese			Brescia		
Variables	Component 1	Component 2	Component 3	Component 1	Component 2	Component 3
WGS X	-0.13927	0.683471	-0.00768	0.095493	0.663397	-0.23586
WGS Y	0.009364	-0.69467	-0.27468	0.061741	-0.69998	0.131854
Surface [m^2]	0.507426	0.068117	-0.07327	0.504897	-0.06195	-0.08098
N° of rooms	0.499274	0.106205	-0.0822	0.490722	-0.00485	-0.12398
N° of bathrooms	0.499407	0.032958	0.04004	0.492114	0.016914	0.103829
EPh [kWh/m^2y]	0.004703	0.18233	-0.94467	-0.0047	-0.25635	-0.93784
Price [€]	0.473636	-0.00678	0.135483	0.49914	0.008	0.121413

In the research a first policy survey has been carried out according to the proposed methodology, providing insights on a large scale. For a more in-depth evaluation of the actual correlation between market values and certifications, it seems appropriate to carry out a re-aggregation of territorial areas not necessarily by province but according to criteria of attractiveness, recalibrating within them the range of assets' values to be related to certification as well as with the age of construction and level of renovation of the buildings. Further analyses will be performed to include the age of the buildings as well as the importance of the urban aggregation considered. In other words, the application of the method to larger cities or "rural" areas should certainly include the different population densities with a redefinition of the size of the clusters. A further development could be also achieved by a better knowledge of the technological and typological features of the analyzed buildings.

References

[1] Direttiva 2010/31/UE del Parlamento Europeo e del Consiglio del 19 maggio 2010 sulla prestazione energetica nell'edilizia (rifusione), Gazzetta ufficiale dell'Unione europea 18.6.2010, L 153/13., 2010
[2] Direttiva 2012/27/UE del Parlamento Europeo e del Consiglio del 25 ottobre 2012 sull'efficienza energetica, che modifica le direttive 2009/125/CE e 2010/30/UE e abroga le direttive 2004/8/CE e 2006/32/CE gazzetta ufficiale dell'unione europea 14.11.2012.
[3] Direttiva (UE) 2018/844 del Parlamento Europeo e del Consiglio del 30 maggio 2018 che modifica la direttiva 2010/31/UE sulla prestazione energetica nell'edilizia e la direttiva 2012/27/UE sull'efficienza energetica (Testo rilevante ai fini del SEE) Gazz.
[4] Real Estate market values database: https://www.borsinoimmobiliare.it
[5] CRESME Ricerche S.p.A., XXVI Rapporto Congiunturale e previsionale CRESME, Il Mercato delle Costruzioni 2019, Lo scenario di medio periodo 2018-2023.
[6] CENED open data website: http://www.cened.it/opendata_cenedplus2
[7] Abdi H., Williams L. J., (2010) Principal component analysis, https://doi.org/10.1002/wics.101
[8] Moore B. (1981) Principal component analysis in linear systems: Controllability, observability, and model reduction, https://doi.org/10.1109/TAC.1981.1102568
[9] Buljak V., Cocchetti G., Cornaggia A., Maier G. (2018) Parameter identification in elastoplastic material models by Small Punch Tests and inverse analysis with model reduction, https://doi.org/10.1007/s11012-018-0914-3
[10] Ostrowski Z, Białecki RA, Kassab AJ (2008) Solving inverse heat conduction problems using trained POD-RBF network inverse method. Inverse Probl Sci Eng 16(1):35–54. https://doi.org/10.1080/17415970701198290
[11] Jolliffe I. (2011) Principal Component Analysis. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg