A Multivariate Econometric Analysis for the Forecasting of the Interdependences Between the Housing Prices and the Socio-economic Factors in the City of Barcelona (Spain)

Pierluigi Morano[1], Francesco Tajani[2(✉)], Maria Rosaria Guarini[2], Felicia Di Liddo[2], and Debora Anelli[2]

[1] Department of Civil Engineering Sciences and Architecture, Polytechnic University of Bari, 70125 Bari, Italy
[2] Department of Architecture and Design, "Sapienza" University of Rome, 00196 Rome, Italy
francesco.tajani@uniromal.it

Abstract. The extreme volatility that has characterized the real estate market in recent years constitutes an important issue for researchers and experts, stimulating the development and the test of evaluation models able to predict future trends and to monitor the consequences of scenario evolutions that are different from those initially expected. With reference to the metropolitan area of Barcelona (Spain), the methodology implemented in this paper has allowed to make explicit the functional relationships between the residential properties prices and the socio-economic variables selected by the model (number of loans, unemployment level, market rent). The analysis carried out is "dynamic", i.e. it refers to a quarterly time series database covering a period of sixty-seven periods (1st quarter 2001-3rd quarter of 2017). The results obtained have shown the forecasted potentialities of the tool used, as support *(i)* for the investment decisions of private operators, *(ii)* for the fiscal decisions of central governments, *(iii)* for the selection of the most convenient urban transformation initiatives from local administrations.

Keywords: Econometric analysis · Spanish residential market · Socio-economic factors · Property values · Market rent

1 Introduction

In the last years, characterized by a situation of latent economic crisis, the non-performing bank loans have strongly increased. In the same category of the non-performing loans, the EU has reported the "unlikely to pay" (UTP) loans and the "bad debts" to be accounted in the banks' balance sheets. Therefore, there is a widespread need to promote transparency and reliability on the estimated values and on the times for the credit recovery, not only for the repayment of the debts, but also in order to ensure more effective mortgage loans [7].

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11622, pp. 13–22, 2019.
https://doi.org/10.1007/978-3-030-24305-0_2

In this sense, the Italian government has provided laws in terms of consumer default management, in relation to the transfer of the property, and in terms of bankruptcy regulations, in order to speed up the judicial procedures, that are particularly long - 1100 days in Italy - compared to other European countries - 410 in Germany and 550 in Spain.

With reference to the market value assessment of properties as guarantee of credit exposures and bad debts, the operators share the need, on the one hand, to obtain prudential valuations, by determining values that are sustainable in the long-term, and, on the other hand, to formulate value judgments that preclude the risk of a sell-off of the properties, preserving the guarantee value and effectively interpreting the conditions of the reference market [1, 4, 12].

The topic of the present research concerns the framework outlined. The work has analyzed the functional relationships between the housing selling prices and the main socio-economic variables relating to the metropolitan area of Barcelona (Spain), for which it has been possible to collect the corresponding time series, sufficiently representative of the evolutions of the characteristics identified. The output obtained is a model of simple interpretation, characterized by a high statistical performance and by the consistency with the expected empirical phenomena, and it represents a useful support for the forecasting of future market price trends, due to the different evolutions of the socio-economic factors in the analyzed context [2, 3, 6, 9, 10].

The research has been structured as follows. In the Sect. 2 the case study has been presented: the time series of the variables considered have been described and the main descriptive statistics have been illustrated. In the Sect. 3 the methodology implemented has been explained, highlighting the advantages of the algorithm used in a dynamic analysis compared to a classic Vector Autoregression (VAR). In Sect. 4 the application of the proposed method has been illustrated: the equations obtained have been specified, the best performing model from a statistical, practical and empirical point of view has been identified. In Sect. 5 the functional relationships of the chosen model have been interpreted and the predictive potentialities of the methodology used have been tested. Finally, the conclusions of the work have been discussed.

2 Case Study

The study sample considered for the present research consists of the time series of the main socio-economic factors collected for the metropolitan area of the city of Barcelona (Spain).

Taking into account the Spanish sources consulted on line, the time series of the variables refer to a period of approximately seventeen years, on a quarterly basis, from the first quarter of 2001 to the third quarter of the year 2017. In particular, the variables considered in the model have been as follows:

- unit selling prices of residential properties, expressed in €/m^2 [*Y*], which represents the dependent variable of the model. The average value of this variable for the sample measured in the Barcelona metropolitan area is equal to 2,167.38 €/m^2, with the minimum recorded value equal to 1,235.76 €/m^2 (first quarter of the year 2001) and the maximum value 3,280.06 €/m^2 (second quarter of the year 2007);

- number of total transactions of residential properties [*T*] observed in the relevant territorial context. In the study sample, an average value of 13,320 has been recorded, in a range between 4,948 (first quarter of the year 2013) and 28,812 transactions (second quarter of the year 2005);
- total number of residences [*N*] in the metropolitan area of Barcelona, grew by about 15% between the first quarter of the year 2001 (in which there were 2,280,334 residences) and the third quarter of the year 2017 (with a number of residences equal to 2,627,671);
- total population [*P*] in the metropolitan area of Barcelona. For this variable an increase in the analysis period of around 16% has been recorded (=4,804,606 in the first quarter of the year 2001 and 5,576,037 in the third quarter of the year 2017);
- total number of firms [*I*] in the territorial context of the metropolitan area of Barcelona. For this variable, the average number of firms for the selected study sample is equal to 438,518, oscillating from a minimum value of 374,027 (first quarter of the year 2001) to the maximum recorded value of 477,942 (first quarter of the year 2008);
- total number of "new" mortgages for housing purchases [*M*]. On average, over the analyzed period, a number of mortgages in each quarter has been detected equal to 25,733, with the minimum number recorded equal to 5,709 (fourth quarter of the year 2013) and the maximum number equal to 58,879 (first quarter of the year 2006);
- number of unemployed people [*D*] in the metropolitan area of Barcelona. The average number of unemployed people has been equal to 361,306, in a range between the minimum value of 193,467 (second quarter of the year 2007) and the maximum value of 492,288 (first quarter of the year 2013);
- number of employment contracts [*C*]. On average, the number of employment contracts signed during the period considered has been 156,713. The minimum value observed has been equal to 116,913 (first quarter of the year 2009) and the maximum value has been equal to 206,569 (second quarter of the year 2017);
- housing market rent [*L*], expressed in €/m^2 * month. The average value is equal to 11 €/m^2 * month, whereas the values recorded in the chosen period range from a minimum value of 8.85 €/m^2 * month (first quarter of the year 2014) to a maximum value of 13.24 €/m^2 * month (second quarter of the year 2017);
- average selling prices of the buildable soils located in the area of Barcelona metropolitan area [*S*]. The average value observed for the sample has been equal to 314.95 €/m^2, with a minimum value equal to 193.65 €/m^2 (third quarter of the year 2013) and a maximum value equal to 455.53 €/m^2 (fourth quarter of the year 2006).

In Table 1 the main descriptive statistics of the variables considered for the case study have been summarized.

Table 1. Main descriptive statistics of the variables considered in the study sample

Variable	Mean	Standard deviation	Max value	Min value
Unit selling price [€/m^2]	2,167.38	590.63	3,280.06	1,235.76
Number of total transactions of residential properties [n]	13,320.73	6,783.89	28,812	4,948
Total number of residences [n.]	2,513,672	116,166	2,627,671	2,280,334
Total population [n.]	5,346,839	239,760	5,576,037	4,804,606
Total number of firms [n.]	438,518	28,165	477,942	374,027
Total number of "new" mortgages for housing purchases [n.]	25,733	16,954	58,879	5,709
Number of unemployed people [n.]	361,306	101,190	492,288	193,467
Number of employment contracts [n.]	156,713	24,775	206,569	116,913
Housing market rent [€/m^2 * month]	11	1.37	13.24	8.85
Average selling prices of the buildable soils [€/m^2]	314.95	77.83	455.53	193.65

3 The Methodology

The methodology implemented for the dynamic analysis is *Evolutionary Polynomial Regression* (EPR). This is a data-driven hybrid method that combines polynomial structures together with evolutionary search in order to obtain symbolical expressions with exponents of polynomial form by means a simple genetic algorithm engine [5].

In particular, EPR searches for symbolic structures in the first stage by the genetic algorithm and estimates constant values in the second stage by solving a Least Squares linear problem, thus assuming a biunique relationship between a structure and its parameters. The general symbolic expression returned by EPR is reported in Eq. (1):

$$y = \sum_{j=1}^{m} F(X, f(X), a_j) + a_0 \tag{1}$$

where y is the estimated output of the process, a_j is an adjustable parameter determined by the method, F is a function constructed by the process, $\mathbf{X}$ is the matrix of the input variables, f is a function defined by the user, m is the length (number of terms) of the expression (bias excluded).

In the case of dynamic analysis, with reference to the study database in the present work, the dependent variable is the unit selling price at time t (Y_t), whereas the matrix $\mathbf{X}$ of the input variables includes the unit selling prices with the relative "lags" - $Y_{t-1},\ldots,$ Y_{t-p}, where p is the maximum eligible lag considered by the user - and the selected input variables, also with the related "lags" - $X_{i,t},\ldots, X_{i,t-p}$, where $i = 1,\ldots, k$ represents a specific variable among the k factors identified in the preliminary phase, and p is the maximum eligible lag considered appropriate for the analysis by the user.

Moreover, in order to obtain the maximum explanatory performance of the final function, EPR provides that each variable of the matrix **X** can be raised to a specific and appropriate exponent, within a range pre-established by the user and iteratively modified until the achievement of the statistically more reliable solution.

The choice of the best models is performed through the calculation of the Coefficient of Determination (COD) for each equation, defined in Eq. (2):

$$COD = 1 - \frac{N-1}{N} \cdot \frac{\sum_{N} \left(y_{estimated} - y_{detected}\right)^2}{\sum_{N} \left(y_{detected} - mean(y_{detected})\right)^2} \tag{2}$$

where $y_{estimated}$ are the values of the dependent variable estimated by the methodology, $y_{detected}$ are the collected values of the dependent variable, N is the sample size in analysis. The fitting of a model is higher when the COD is close to the unit value.

The implementation of a multi-objective genetic algorithm represents the fundamental advantage of the proposed methodology, as optimization strategy based on the Pareto dominance criterion. These objectives are conflictual, and aim at *(i)* the maximization of model accuracy, through the satisfaction of appropriate statistical criteria of verification of each equation; *(ii)* the maximization of model's parsimony, through the minimization of the number of terms (a_j) of the equation; *(iii)* the reduction of the complexity of the model, through the minimization of the number of the input variables (**X**) of the final equation. Therefore, the method allows to obtain, at the end of the modeling phase, a set of model solutions (i.e. the Pareto front of optimal models) for the three objectives considered.

In the case of dynamic analysis on time series, a further advantage of the implementation of the multi-objective genetic algorithm is represented by the identification of the best "lag" of each considered input variable, and consequently the best combination of the possible eligible lags of the variables in the model, in order to optimize the three objectives described above.

This ability of the methodology undoubtedly constitutes an important innovative feature, taking into account that in a classic multivariate econometric technique - for example, a VAR - it is the user that iteratively combines the variables with different lags, up to the model characterized by the best statistical reliability due to several tests (verification of the *t*-value for each parameter of the coefficient matrix, Portmanteau test to verify the reliability of the linear model, Jarque-Bera test to verify the residual normality, analysis of the autocorrelations among the residuals, etc.) [8].

Finally, the compositional logic with which the genetic algorithm operates (combination of several variables within the same additive term and admissibility of the exponents for each term) allows to overcome the constraint of the preliminary univariate analysis for the equations of a VAR (graphic analysis, unit root (ADF) test, etc.). The econometric technique allows, in this way, to gather a consistent range of solutions for the user, different from each other for COD and for complexity of the algebraic form, among which it is possible to select the most appropriate solution according to the specific needs, the knowledge of the phenomenon in analysis and the type of experimental data used [11].

4 Application of the Methodology

4.1 The Model

The methodology is implemented considering the additive function F and no function f selected among those allowed by the application software (logarithmic, semi-logarithmic, exponential, etc.). Each additive monomial term is assumed as a combination of the inputs ($Y_{t\text{-}1}$, ..., $Y_{t\text{-}p}$, $X_{i,t}$, ..., $X_{i,t\text{-}p}$) raised to the proper exponents. In particular, due to the limited number of the time series sequence, the analysis is carried out considering the autoregressive variable of input, exclusively the one associated with a single delay ($Y_{t\text{-}1}$), whereas for the other input variables an eligible delay of 12 (p = 12) has been assumed.

Candidate exponents belong to the set (−3; −2; −1; −0.5; 0; 0.5; 1; 2; 3). The genetic algorithm therefore, considers all the possible combinations according to the delay of the input parameters (variable from t to $t\text{-}12$ for each characteristic considered), of the selected exponent, of the variables number within the same additive term: the result is a wide range of solutions that define the Pareto frontier for the conflicting objectives imposed by the methodology, among which the best compromise between performance statistics, simplicity of the model algebraic form and the empirical reliability of the relationship between candidate inputs and the dependent variable should be identified.

The maximum number m of additive terms in final expressions set by the user in the preliminary phase, is equal to four. Finally, in order to eliminate the distortive effects that could be generated by the different numerical entities of the time series, a normalization of each historical series, compared to the highest numerical value detected for each variable, has been developed.

The implementation of the EPR methodology has generated several models, the main of which in Table 2 are set out. The equations show different direct and inverse functional relationships between the output variable of the model (selling prices at time t) and the selected inputs variables with their "delays".

All the models generated show a high level of statistical accuracy, with a COD that exceeds 85% in all cases. Among the main models included in Table 2, it should be observed that Eq. (3) presents a functional form characterized by additive terms of simple empirical interpretation than to the subsequent models, for which the functional links statistical improvement becomes not very significant.

Table 2. Equations obtained by the implementation of EPR

Eq. (n)	Model
(3)	$Y_t = +0.078128 \cdot L_{t-4}^{0.5} - 0.030354 \cdot L_{t-2}^{0.5} +$ $- 0.17738 \cdot M_{t-8}^{0.5} \cdot D_{t-8}^{0.5} + 1.0015 \cdot Y_{t-1} + 0.030369$
(4)	$Y_t = +0.017995 \cdot L_{t-11}^{0.5} - 0.024501 \cdot C_{t-6}^{0.5} -$ $+0.15475 \cdot M_{t-8} \cdot D_t \cdot D_{t-8} + 0.995544 \cdot Y_{t-1} + 0.034703$

(*continued*)

Table 2. (*continued*)

Eq. (n)	Model
(5)	$Y_t = -0.027627 \cdot C_{t-6} - 0.15812 \cdot M_{t-8} \cdot D_t \cdot D_{t-8} + + 0.15475 \cdot T_t \cdot T_{t-7} L_{t-11} + 0.98777 \cdot Y_{t-1} + 0.040871$
(6)	$Y_t = -0.030139 \cdot C_{t-6}^{0.5} \cdot S_t^{0.5} - 0.16632 \cdot M_{t-8} \cdot D_t \cdot D_{t-8} + + 0.16971 \cdot T_t \cdot T_{t-7} \cdot L_{t-11}^2 + 0.99263 \cdot Y_{t-1} + 0.037143$
(7)	$Y_t = -0.14719 \cdot I_{t-2}^2 \cdot M_{t-8}^{0.5} \cdot D_{t-8}^{0.5} + -0.030532 \cdot T_{t-6}^{0.5} \cdot L_{t-2} + + 0.15319 \cdot T_t^{0.5} \cdot T_{t-3}^{0.5} \cdot L_{t-4} \cdot S_{t-11}^{0.5} + 0.99079 \cdot Y_{t-1} + 0.038832$
(8)	$Y_t = -0.15532 \cdot I_{t-2}^2 \cdot M_{t-8}^{0.5} \cdot D_{t-8}^{0.5} + -0.032974 \cdot T_{t-6}^{0.5} \cdot C_{t-4} \cdot L_{t-2}^{0.5} + + 0.16108 \cdot T_t^{0.5} \cdot T_{t-3}^{0.5} \cdot L_{t-4} \cdot S_{t-11}^{0.5} + 0.98989 \cdot Y_{t-1} + 0.039316$

Therefore, Eq. (3) has been considered as the best model in terms of statistical performance and interpretive simplicity of functional relationships. Figure 1 shows the time series trends of the Y_t detected and of the Y_t estimated by Eq. (3). The graphical representation points out the statistical goodness of the results: an almost exact match of the values is widely verified between the two curves in the considered time period.

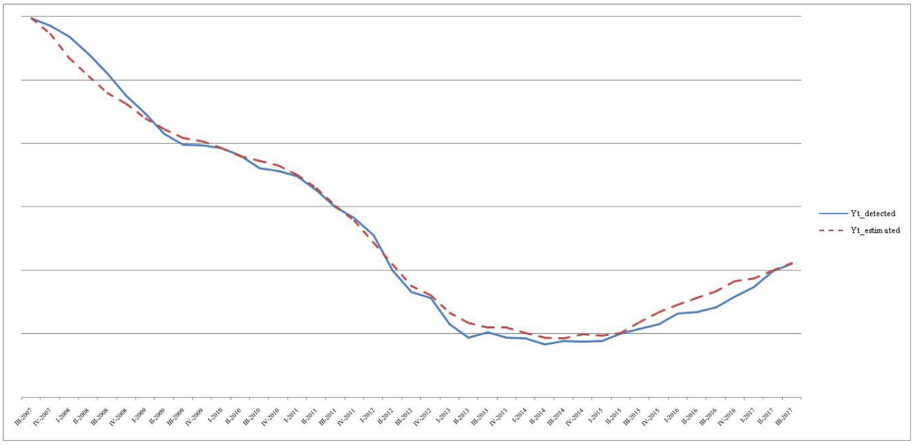


Fig. 1. Y_t detected (continuous line) and Y_t estimated (broken line) for the Barcelona case study

4.2 Interpretation of the Selected Model

The model of Eq. (3) selects only five variables as the most representative in the formation of selling prices at time *t*, and the related mathematical expression allows the immediate verification of the functional links between the selling prices at the time *t* (Y_t) and the following input parameters: selling prices at time *t-1* (Y_{t-1}) through a direct functional correlation; the housing market rent with a delay of four quarters (L_{t-4}), through a direct functional relationship; the housing market rent relate to the time t-2

($L_{t\text{-}2}$), with an inverse functional link; the total number of "new" mortgages for housing purchases with regard to time t-8 ($M_{t\text{-}8}$) and the number of unemployed people at time t −8 ($D_{t\text{-}8}$), combined into a single term, functionally connected to the variable Y_t through an inverse correlation.

The graph in Fig. 2 allows the qualitative representation of the functional correlations between the output variable Y_t and the input variables $L_{t\text{-}4}$, $L_{t\text{-}2}$, $M_{t\text{-}8}$ and $D_{t\text{-}8}$: the trends represented show that at negative phases for the total number of "new" mortgages for housing purchases and the level of unemployment followed increasing phases for selling prices in the metropolitan area of Barcelona after eight quarters (two years), whereas the positive trend for the housing market rents brings forward by four quarters (one year) the growing phase of residential selling prices.

The inverse relationship between the variable $L_{t\text{-}2}$ and the output variable, explained in Eq. (3), does not appear to match the two variables graphic trends, which instead seem to be characterized by fairly consistent trends.

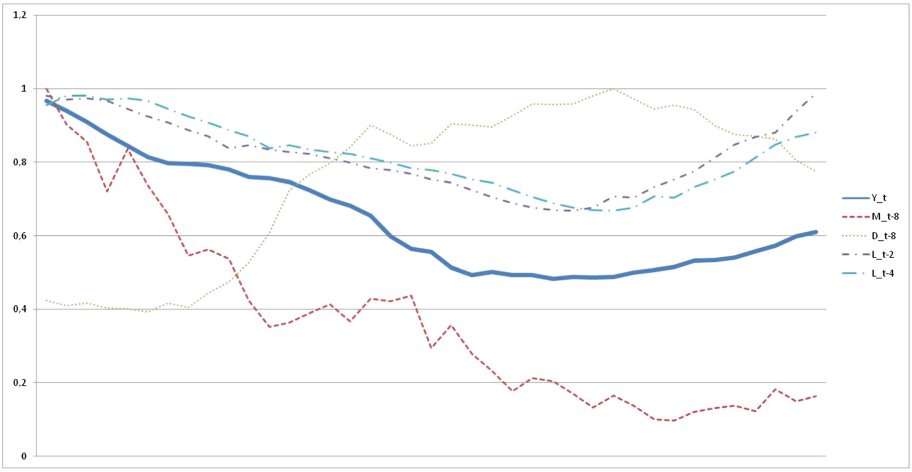


Fig. 2. Variable trends referred to the same time for the Barcelona case study

The forecasting capacity of selling prices represents an interesting peculiarity of the model obtained. In fact, since the selling prices at time t are linked to the number of mortgages and the level of unemployment at the time $t\text{-}8$, it is possible to exclusively assume different developments over time of the housing market rents and, consequently, to forecast future housing price quotations in the metropolitan area of Barcelona.

Figure 3 describes a sensitivity analysis of the selling prices, after the third quarter of 2017, and considering constant changes in the future market rents in a range [−9%; +9%]. First of all, the results obtained confirm the empirical evidence of a positive correlation between the selling prices and the housing market rent over time: this relationship is evident starting from the convergence point that occurs in the first quarter of 2019, after which positive variations in the housing market rents determine

an increase in selling prices, whereas negative changes in housing market rents adversely affect the trend in residential prices. Finally, in the two years considered, the maximum selling price gap between a "very" positive trend in leases (+9%) and a "very" negative trend (−9%) is still not very significant, being equal to about 4%.

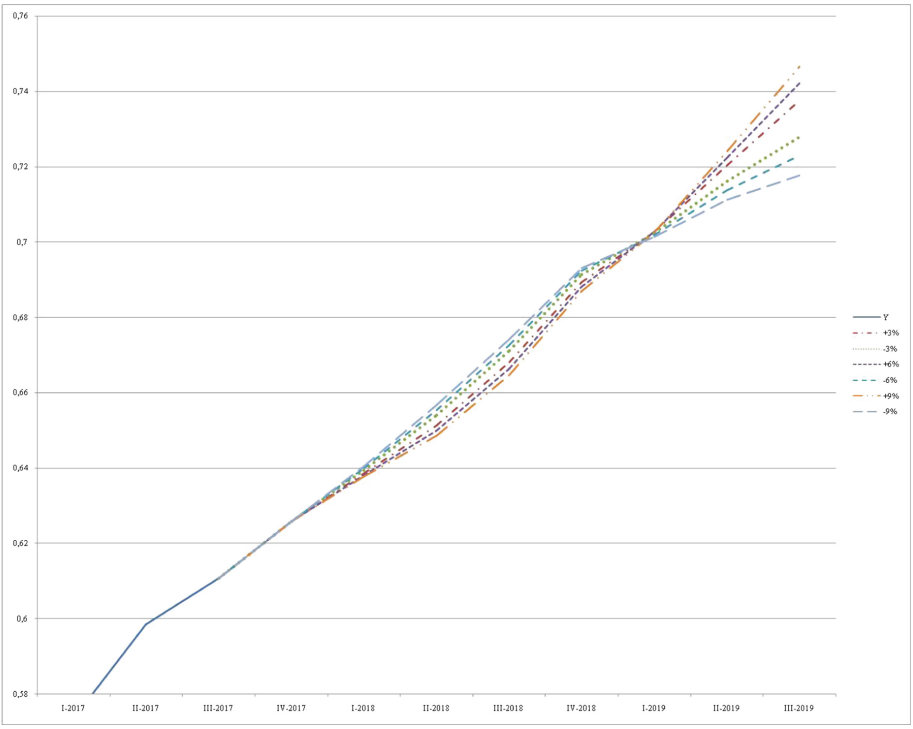


Fig. 3. Possible evolutions in the selling prices due to different variations in the housing market rent in Barcelona

5 Conclusions

In the present research the functional relationships between the residential properties selling prices and the main socio-economic variables related to the metropolitan area of Barcelona (Spain) have been analyzed. In particular, the dynamic analysis has referred to a quarterly time series database covering a period of sixty-seven periods (1st quarter 2001-3rd quarter of 2017). The methodology implemented, named EPR, has allowed to select the most significant correlations, and to generate a model characterized by a high statistical reliability and, at the same time, by a simplicity in the empirical interpretation of the functional relationships.

An important peculiarity of the model is the possibility to be used for forecasting purposes: in particular, as the historical data for the number of mortgages and the level of unemployment are sufficiently available, it is possible to assess the evolution of the

selling prices over a period of two years, for different possible variations of the housing market rents. In this sense, the methodology used, applied to the case of the city of Barcelona but implementable in any territorial context, represents a fundamental support (*i)* for the investment decisions of private operators, (*ii)* for the fiscal decisions of central governments, (*iii)* for the selection of the most convenient urban transformation initiatives from local administrations.

Further insights may address the implementation of the VAR method, and therefore the construction of a vector system of equations able to explain all the functional relationships between the variables considered. In this way, the variations of the factors over time can be endogenously obtained, by using the equations that define the model, thus avoiding the exogenous assumptions on the order of magnitude and of the variations signs.

References

1. Bourassa, S.C., Hendershott, P.H., Murphy, J.: Further evidence on the existence of housing market bubbles. J. Property Res. **18**(1), 1–19 (2001)
2. Mallamace, S., Calabrò, F., Meduri, T., Tramontana, C.: Unused real estate and enhancement of historic centers: legislative instruments and procedural ideas. In: Calabrò, F., Della Spina, L., Bevilacqua, C. (eds.) ISHT 2018. SIST, vol. 101, pp. 464–474. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-92102-0_49
3. Cosentino, C., Amato, F., Murgante, B.: Population-based simulation of urban growth: the Italian case study. Sustainability **10**(12), 4838 (2018)
4. Del Giudice, V., De Paola, P.: Real estate economics, management and investments: new perspectives and frontiers. Buildings **8**(3), 40 (2018)
5. Giustolisi, O., Savic, D.A.: A symbolic data-driven technique based on evolutionary polynomial regression. J. Hydroinform. **8**(3), 207–222 (2006)
6. Guarini, M.R., D'Addabbo, N., Morano, P., Tajani, F.: Multi-criteria analysis in compound decision processes: the AHP and the architectural competition for the Chamber of Deputies in Rome (Italy). Buildings **7**(2), 38 (2017)
7. Levy, D.S., Lee, C.K.C.: The influence of family members on housing purchase decisions. J. Property Investment Finan. **22**(4), 320–338 (2004)
8. Manganelli, B., Morano, P., Tajani, F.: House prices and rents: the Italian experience. WSEAS Trans. Bus. Econ. **11**, 219–226 (2014)
9. Rosato, P., Breil, M., Giupponi, C., Berto, R.: Assessing the impact of urban improvement on housing values: a hedonic pricing and multi-attribute analysis model for the historic Centre of Venice. Buildings **7**(4), 112 (2017)
10. Scorza, F., Pilogallo, A., Las Casas, G.: Investigating tourism attractiveness in inland areas: ecosystem services, open data and smart specializations. In: Calabrò, F., Della Spina, L., Bevilacqua, C. (eds.) ISHT 2018. SIST, vol. 100, pp. 30–38. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-92099-3_4
11. Tajani, F., Morano, P., Torre, C.M., Di Liddo, F.: An analysis of the influence of property tax on housing prices in the Apulia region (Italy). Buildings **7**(3), 67 (2017)
12. Vargas-Silva, C.: Monetary policy and the US housing market: a VAR analysis imposing sign restrictions. J. Macroecon. **30**(3), 977–990 (2008)