Retrospective Analysis of the Dwelling Price by Means of STAR Models with Neighborhood Effects

Authors A. Beamonte, P. Gargallo, and M.J. Salvador

Abstract

The need of retrospective predictions arises in situations such as litigations, inheritances or fiscal purposes to provide the fair market value of a dwelling of given characteristics and specific location. In this paper, a Bayesian method to carry out this retrospective analysis is proposed by using a Spatio-Temporal Autoregressive (STAR) model with neighborhood effects and heteroscedastic errors. Moreover, the developed methodology allows the construction of a retrospective price index that shows the evolution of the dwelling price level in a specific geographical area. The methodology is illustrated with an application to the real estate market in the Spanish city of Zaragoza.

The evolution of the housing market is a subject of extraordinary interest from both a theoretical and practical point of view. In addition to reasons such as the number of jobs, the financial assets generated, the assessment of new mortgages, and the formulation of housing policies highlight the importance of the building sector in national economies.

The dwelling price has been the variable that traditionally has explained the behavior of the real-estate market, and the hedonic pricing model (Court, 1939; Rosen, 1974) the instrument to analyze and forecast its value. The hedonic model assumes that the dwelling price is a function of its structural characteristics and location. The accuracy of the price valuations is affected by the selection of the characteristics, the functional form of the hedonic function, and the assumptions made on the random error terms. Although considerable research efforts have been made in modeling and forecasting housing prices, there are still problems with the models. Different models, procedures, and techniques have coped with these issues such as the regression spatial models (Dubin, 1988, 1992, 1998; Basu and Thibodeau, 1998; Osland, 2010), kriging and cokriging procedures (Chica-Olmo, 2007; Montero and Larraz, 2011), grid techniques (Pace and Gilley, 1998), replication methods (Lai, Vandell, Wang, and Welke, 2008), artificial intelligence tools (Nguyen and Cripps, 2001; Zurada, Levitan, and Guan, 2011) or adaptive neuro-fuzzy modeling (Guan, Zurada, and Levitan, 2008), among others.
The proliferation of large housing price data sets, spatially and temporally indexed, has raised the need to analyze the spatial nature of the price property valuation, as well as the price changes over time. From a methodological point of view, the introduction of time into spatial modeling brings a substantial increase in the scope of the work because separate decisions regarding spatial and temporal correlations and how space and time interact in the data must be made. Such modeling will also carry an obvious associated increase in notational and computational complexity (Banerjee, Carlin, and Gelfand, 2004). In this context, new families of statistical models have emerged such as spatio-temporal autoregressive models (Pace, Barry, Clapp, and Rodriguez, 1998; Pace, Barry, Gilley, and Sirmans, 2000), local regression models (Clapp, 2004) or Bayesian hierarchical spatial models (Gelfand, Ghosh, Knight, and Sirmans, 1998; Gelfand, Ecker, Knight, and Sirmans, 2004), which allow the description of the spatio-temporal dependences present in data in a flexible and realistic way (Case, Clapp, Dubin, and Rodriguez, 2004).

Our starting point in this work is the Spatio-Temporal Autoregressive (STAR) model with neighborhood effects first introduced in Pace, Barry, Clapp, and Rodriguez (1998) and Pace, Barry, Gilley, and Sirmans (2000). This model permits the capture of a significant number of local spatio-temporal effects that are not explicitly included in the hedonic part of the model by means of the use of nearest neighbor methods. These authors estimated the model parameters with low-cost OLS. Later, Sun, Tu, and Yu (2005), Beamonte, Gargallo, and Beamonte (2008), Alberto, Beamonte, Gargallo, Mateo, and Salvador (2010), and Beamonte, Gargallo, and Salvador (2010 a, b), proposed the use of Bayesian techniques based on MCMC methods that provide more robust estimations of the parameters and, particularly, correct the heteroscedasticity problems in order to weaken the influence of atypical observations. Moreover, these estimation procedures do not depend on asymptotic results and allow the incorporation of restrictions on the model parameters as well as to build predictive intervals taking into account the uncertainty associated with the model selection and the estimation process. Furthermore, these authors elaborated procedures to build prospective predictions (i.e., forward in time) of the price levels, which are used to carry out the outsampling comparison (Alberto, Beamonte, Gargallo, Mateo, and Salvador, 2010) and validation of the proposed models (Beamonte, Gargallo, and Salvador, 2010 a, b).

However, some situations (litigations, inheritances, fiscal purposes) demand retrospective forecasts to provide the fair market value of a dwelling of given characteristics and specific location. The need of retrospective predictions also arises in other contexts, such as the construction of price indexes where some smoothing of the elaborated series might help to identify trends. An example is the Sale Price Appraisal Ratio (SPAR) method (Bourassa, Hoesli, and Sun, 2006; De Hann et al., 2009) where retrospective appraisals of dwelling prices are used to build constant quality measures of housing prices, avoiding the inclusion of information about some characteristics of the houses in the sample.
Taking into consideration this background, the objective of the paper is the elaboration of a price dwelling forecast for a transaction that took place in a specific location at a fixed date in the past, not observed but included in the period of the considered data. To that aim, a Bayesian method based on a STAR model with neighborhood effects and heteroscedastic errors is proposed. The retrospective probability distribution of the dwelling prices is calculated, and point price forecasts and Bayesian credibility intervals that incorporate the uncertainty associated with the estimation of the model parameters are obtained. The proposed approach let us elaborate a price index to show the retrospective joint evolution of prices of a representative dwellings basket. The methodology is applied to a dataset containing the prices of all the dwellings transactions reported in a specific area of the city of Zaragoza (Spain) during a two-year period.

The paper is organized as follows: Section 2 introduces the STAR model with neighborhood effects. Section 3 describes the Bayesian statistical methodology to retrospectively estimate the price of a dwelling and the elaboration of a price index. Section 4 shows the application to the real estate market of Zaragoza. Finally, Section 5 presents the main conclusions and future lines of research. The Appendix contains the mathematical details of the paper.

### The STAR Model with Neighborhood Effects

#### The Model

The data correspond to a set of \( n \) dwelling transactions in a fixed period of time within a determined geographical area, ordered by date, from the oldest to the most recent. Let \( Y = (y_i; \ i = 1,...,n) \) be the \((n \times 1)\) log-prices vector with \( y_i = \log p_i \), where \( p_i \) is the price of the \( i \)th transaction.

The STAR model with neighborhood effects used in the paper is given by:

\[
Y = Z\alpha + X\beta + \phi_T(T(Y - X\beta)) + \phi_S(S(Y - X\beta))
+ \phi_{ST}(ST(Y - X\beta)) + \phi_{TS}(TS(Y - X\beta)) + \varepsilon, \tag{1}
\]

where:

- \( Z = (z_1,...,z_n)' \) is the \((n \times q)\) matrix of observations corresponding to \( Z_1,...,Z_q \) where \( z_i = (z_{i1},...,z_{iq})' \) are the characteristics of the \( i \)th transaction of a dwelling for \( i = 1,...,n \). These variables do not present spatio-temporal lags.
- \( \alpha = (\alpha_1,...,\alpha_q)' \) is the coefficient vector of the hedonic regression which determines the type and degree of influence exerted by the independent variables \( Z_1,...,Z_q \) on the dwelling price.
\( \mathbf{X} = (x_1, \ldots, x_n)' \) is the \((n \times u)\) matrix of observations corresponding to \(X_1, \ldots, X_u\) where \(x_i = (x_{i1}, \ldots, x_{iu})'\). These variables present spatio-temporal lags.

\( \mathbf{B} = (\beta_1, \ldots, \beta_u)' \) is the coefficient vector of the hedonic regression which determines the type and degree of influence exerted by the independent variables \(X_1, \ldots, X_u\) on the dwelling price.

\( \phi_j, j \in \{T, S, ST, TS\} \) are the autoregressive coefficients of the model that determine the spatio-temporal neighborhood effects transmitted by the previous transactions.

\( \epsilon \sim N_u(0, \sigma^2 \mathbf{V}) \) is the error vector of the model where \( \mathbf{V} = \text{diag}(v_1, \ldots, v_n) \) \((v_i > 0 \ i = 1, \ldots, n \ \sigma^2 > 0)\) and \(N_p(\mathbf{\mu}, \Sigma)\) denotes the \(p\)-dimensional multivariate normal distribution whose mean vector and variance and covariance matrix are \( \mathbf{\mu} \) and \( \Sigma \), respectively.

The roles of the terms \( \mathbf{Z} \alpha \) and \( \mathbf{X} \beta \) correspond to the hedonic part of the model; and \( \phi_T \mathbf{Y} \mathbf{X} \beta, \phi_S \mathbf{Y} \mathbf{X} \beta, \phi_{ST} \mathbf{Y} \mathbf{X} \beta, \phi_{TS} \mathbf{Y} \mathbf{X} \beta \) capture the spatial, temporal and compound spatio-temporal influence trends, measured through errors and not collected by the hedonic part.

Matrix \( \mathbf{S} = \Sigma_{t=1}^{m_S} \lambda^t \mathbf{S}_t / \sum_{t=1}^{m_S} \lambda^t \) contains the spatial weights with \( 0 < \lambda \leq 1 \) and \( m_S \in \mathbb{N} \) where \( \mathbf{S}_t = (s_{ij}^t) \) with \( s_{ij}^t = 1 \) if the \( j \)th transaction is the \( t \)th closest in space to the \( i \)th transaction in the past time, and 0 otherwise.

Matrix \( \mathbf{T} = \Sigma_{t=1}^{m_T} \gamma^t \mathbf{T}_t / \sum_{t=1}^{m_T} \gamma^t \) encloses the temporal weights with \( 0 < \gamma \leq 1 \) and \( m_T \in \mathbb{N} \) where \( \mathbf{T}_t = (t_{ij}^t) \) with \( t_{ij}^t = 1 \) if the \( j \)th transaction is the \( t \)th closest in time to the \( i \)th transaction in the past time, and 0 otherwise. For further details on these matrices see Beamonte et al. (2008).

**Estimation of the Model**

We use the Bayesian methodology described in Beamonte, Gargallo, and Salvador (2010b), which uses a prior distribution over the model parameters given by:

\[
\begin{align*}
\alpha & \sim N_d(0, \Sigma_\alpha); \ \beta \sim N_k(0, \Sigma_\beta); \ \phi_j \sim U(-1, 1), \\
\ j \in \{T, S, ST, TS\}; \\
\tau_i & = \frac{1}{v_i} \sim \chi^2 \left( \frac{r}{\tau}, i = 1, \ldots, n; \ r \sim \text{gamma}(m_1, m_2) \right); \\
\tau & = \frac{1}{\sigma^2} \sim \text{gamma} \left( \frac{d_0}{2}, \frac{d_0 s_0}{2} \right); \\
m_T & \sim \mathcal{U}_D(\mathbf{m}_{\text{max}}); \ \gamma \sim \mathcal{U}(\gamma_{\text{min}}, \gamma_{\text{max}}); \ m_S \sim \mathcal{U}_D(\mathbf{m}_{\text{net}}) \text{ and} \\
\lambda & \sim \mathcal{U}(\lambda_{\text{min}}, \lambda_{\text{max}}).
\end{align*}
\]
These distributions are independent; \( \Sigma_\alpha = s_\alpha^2 I_q \) and \( \Sigma_\beta = s_\beta^2 I_q \); \( U(a, b) \) denotes the uniform distribution in \((a, b)\) with \( a < b \in \mathbb{R} \); \( \chi^2_r \) denotes the chi-square distribution with \( r \) degrees of freedom; \( \text{gamma}(a, b) \) is the gamma distribution with mean \( a/b \) and variance \( a/b^2 \); \( U_p(A) \) is the discrete uniform distribution in the set \( A = \{a_1, ..., a_g\} \), \( m_1, m_2, d_0, s_0 > 0, 0 \leq \gamma_{\min} \leq \gamma_{\max} \leq 1 \), \( 0 \leq \lambda_{\min} \leq \lambda_{\max} \leq 1 \); and \( m_{T_{\text{net}}} = (m_{T_{\text{net},1}}, ..., m_{T_{\text{net},t}}) \) with \( 0 \leq m_{T_{\text{net},1}} < ... < m_{T_{\text{net},t}} < \infty \) and \( m_{T_{\text{net},i}} \in \mathbb{N} \cup \{0\} \), \( m_{S_{\text{net}}} = (m_{S_{\text{net},1}}, ..., m_{S_{\text{net},s}}) \) with \( 0 \leq m_{S_{\text{net},1}} < ... < m_{S_{\text{net},s}} < \infty \) and \( m_{S_{\text{net},i}} \in \mathbb{N} \cup \{0\} \) are discrete nets of the most feasible values of the parameters \( m_T \) and \( m_S \).

Distributions for \( \alpha, \beta, \) and \( \phi \) are standard in Bayesian literature and are diffuse if \( s_\alpha^2 \rightarrow \infty \) and \( s_\beta^2 \rightarrow \infty \). Distributions for \( \tau, r, \) and \( \tau \) follow the treatment given by Geweke (1993), LeSage (1999), and Sun, Tu, and Yu (2005) to the problem of heteroscedasticity and are diffuse if \( m_1, m_2, \) and \( d_0 \rightarrow 0 \).

The scale factors \( \{\tau_i, i = 1,...,n\} \) model the conditional heteroscedasticity of the error terms. They have a mean equal to 1; therefore, the prior distribution supposes the model to be homoscedastic. Nevertheless, the variance of these factors is \( 2/r \), in such a way that the lower the value, the lower the initial confidence in the homoscedasticity of the model. Besides, \( e_i | r \sim t_r \) where \( t_r \) denotes the Student \( t \) distribution with \( r \) degrees of freedom and so the lower the \( r \), the higher the leptokurtosis of the error distribution and the heavier are its tails. Small values of \( r \) are associated with the prior belief that observations with abnormally large errors, in absolute value, exist. This treatment of the error distribution will weaken their influence in the estimation process.

Let \( \theta = (\alpha', \beta', \phi', \tau, r, m_T, m_S, \gamma, \lambda)' \) the vector of model parameters where \( \phi = (\phi_T, \phi_S, \phi_{ST}, \phi_{S\phi})' \) and \( \tau = (\tau_1, ..., \tau_n)' \) with \( \tau_i = 1/v_i \); \( i = 1,...,n \). The estimation of the components of \( \theta \) is carried out from the posterior distribution \( \theta | Y, Z, X \) calculated via the Bayes Theorem. Given that this distribution is not analytically tractable, it is obtained applying Monte Carlo Markov Chain (MCMC) methods [see, for example, Robert and Casella (2004) for a review]. The algorithm used is described in Beamonte, Gargallo, and Salvador (2010b) and as a result we get a sample of the posterior distribution:

\[
\{\theta^{(l)}; \ l = 1,...,n_{\text{sample}}\} \\
= \{(\alpha^{(l)}, \beta^{(l)}, \phi^{(l)}, \tau^{(l)}, r^{(l)}, \tau_{\phi}^{(l)}, m_T^{(l)}, \gamma^{(l)}, m_S^{(l)}, \lambda^{(l)})\} \\
\ l = 1,...,n_{\text{sample}}\] (3)

from which inferences about the model parameters can be made.
Retrospective Analysis

In this section, we describe a procedure to predict retrospectively the price of a set of dwellings of given characteristics with the information provided by the sample (3). First, we show how to forecast the price of a dwelling. We then extend this result to a set of dwellings and build a retrospective price index. From here forward, \([Y]\) will denote the marginal distribution of random variable \(Y\) and \([Y|X]\) the density of the distribution of \(Y\) conditioning by \(X\).

Retrospective Predictions of a Price Dwelling

Our objective is the elaboration of a price dwelling forecast for a transaction that took place at a certain point in the past included in the period of the observed data. Let \(z = (z_1, \ldots, z_q)'\) and \(x = (x_1, \ldots, x_u)'\) be the hedonic characteristics of the dwelling whose price \(p\) we want to forecast. This retrospective prediction is carried out from the posterior predictive distribution given by:

\[
y|Y, Z, z, X, x \sim \int [y|Y, Z, z, X, x, \theta][\theta|Y, Z, X]d\theta,
\]

where \(y = \log(p)\). From this distribution, it is possible, in particular, to obtain retrospective point predictions of the prices by means of their posterior median and to calculate Bayesian credibility intervals from the appropriate quantiles. However, (4) is analytically intractable. For this reason, we appeal, again, to the Monte Carlo method and we obtain a sample of (4) by using the composition sampling from (3). The procedure needs the following intermediate result to be accomplished.

**Proposition 1.** Let \(M\) be the model (1) for a given value of \(\theta\) and let \(i\) be the position according to the date assigned to the transaction whose price we want to predict\(^2\). So, we have that: \(y|Y, Z, z, X, x, \theta \sim N(m(\theta), s^2(\theta))\):

\[
\begin{align*}
\text{with } m(\theta) &= \frac{\tau_i \mu_i + \sum_{j \neq i} \tau_j f_{ji} (y_j - a_{ji})}{\tau_i + \sum_{j \neq i} \tau_j f_{ji}^2}; \\
&\quad s^2(\theta) = \frac{\sigma^2}{\tau_i + \sum_{j \neq i} \tau_j f_{ji}^2}.
\end{align*}
\]
where:

\[
a_{ji} = \mu_j - f_{ji}y
\]

with \(\mu_j = \begin{cases} E[y_j|Y_j, Z_j, X_j, \theta] & \text{if } j < i \\ E[y_j|Y_j, Z_j, z, X_j, x, \theta] & \text{if } j = i \\ E[y_j|Y_j, y, Z_j, z, X_j, x, \theta] & \text{if } j > i \end{cases} \quad (6)
\]

\[
f_{ji} = \phi_Tt_{ji} + \phi_Ss_{ji} + \phi_{ST}(st)_{ji} + \phi_{TS}(ts)_{ji}, \quad (7)
\]

with \(Y_j = (y_l : l < j), Z_j = (z_l : l < j), X_j = (x_l : l < j), T = (t_{ij}), S = (s_{ij}), ST = ((st)_{ij})\), and \(TS = ((ts)_{ij})\).

From the previous proposition, and using sample (3), we obtain an estimation of the density \([y|Y, Z, z, X, x]\) applying the following expression:

\[
\frac{1}{n_{\text{sample}}} \sum_{l=1}^{n_{\text{sample}}} [y|Y, Z, z, X, x, \theta^{(l)}], \quad (8)
\]

where \(y|Y, Z, z, X, x, \theta^{(l)} \sim N(m(\theta^{(l)}), s^2(\theta^{(l})) \quad (9)
\]

with \(m(\theta^{(l)})\) and \(s^2(\theta^{(l)})\) given by the expression (8) calculated in \(\theta = \theta^{(l)}; \ell = 1, \ldots, n_{\text{sample}}\).

Using (8) it is possible to obtain a point forecast of \(p\) by means of the posterior median, \(p(0.5)\), of (8), and a Bayesian credibility interval of a given confidence level \(1 - \alpha (0 < \alpha < 0.5)\), which limits are given by \(p(\alpha/2)\) and \(p(1 - \alpha/2)\), posterior \(\alpha/2\) and \(1 - \alpha/2\) quantiles of (8), respectively. The calculation of these quantities will be made from a sample of (8), which is drawn using the composition sampling. Algorithm 2 describes the details.

**Algorithm 2**

**Step 0 (Initialization)**

Put \(\ell = 1\);

**Step 1 (Sampling of \(p\))**

Draw \(y^{(l)}\) from \(N(m(\theta^{(l)}), s^2(\theta^{(l)}))\), where \(m(\theta^{(l)})\) and \(s^2(\theta^{(l)})\) given by the expressions (5) calculated in \(\theta = \theta^{(l)}\). Calculate \(p^{(l)} = e^{y^{(l)}}\).

**Step 2 (Stopping Rule)**

Set \(\ell = \ell + 1\). If \(\ell > n_{\text{sample}}\), go to Step 3. Otherwise go to Step 1.

**Step 3 (Calculation of point forecast and Bayesian credibility interval)**

Calculate \(\hat{p}(0.5), \hat{p}(\alpha/2)\) and \(\hat{p}(1 - \alpha/2)\) as the median, \(\alpha/2\) and \(1 - \alpha/2\) quantiles of \(\{p^{(l)}; \ell = 1, \ldots, n_{\text{sample}}\}\), respectively.
Elaboration of a Retrospective Price Index

The evaluation of some fiscal or economic policies related to the evolution of the real estate market makes a retrospective estimation of the dwelling price level sometimes necessary in a specific area of interest at a fixed date in the past. This objective can be accomplished by means of index numbers that are able to show the joint evolution of prices of a representative dwellings basket (David, Dubujet, Gourieroux, and Laferrière, 2002; De Hann, Van der Wal, and De Vries, 2009).

Let \( \{1 < t_1 < \ldots < t_K < T\} \) be the periods in which the values of the price index are going to be calculated and \( 1 \leq t_0 < t_1 \) be the reference period. In order to constitute the “basket” to build the price index, we take a dwelling sample with characteristics \( Z_{k}^{\text{basket}}, R \) and \( X_{k}^{\text{basket}}, k \). Therefore, \( (z_{k}^{\text{basket}}, x_{k}^{\text{basket}}) \)' is the vector of hedonic covariates for the \( r \)-th basket dwelling in period \( t_k; k = 1, \ldots, K \).

Let \( p_{k,r}^{\text{basket}} \) be the price of the \( r \)-th dwelling of the basket in period \( t_k; k = 0, \ldots, K; r = 1, \ldots, R \) and let \( p_{k}^{\text{basket}} = (p_{k,1}^{\text{basket}}, \ldots, p_{k,R}^{\text{basket}})' \) and \( y_{k}^{\text{basket}} = \log(p_{k}^{\text{basket}}); k = 0, \ldots, K \).

The value of the price index is calculated for the periods \( \{t_k; k = 1, \ldots, K\} \) by means of the expressions:

\[
\{IP_{t_0}^k = 100 \frac{V_k}{V_0}; k = 1, \ldots, K\},
\]

where \( V_k = \sum_{r=1}^{R} p_{k,r}^{\text{basket}} \) is the value of the dwellings basket in the period \( t_k; k = 0, \ldots, K \).

The value of \( V_0 \) is known at the reference period. However, \( p_{k,r}^{\text{basket}}; k = 1, \ldots, K; r = 1, \ldots, R \) and, hence, \( V_k k = 1, \ldots, K \) is unknown and must be estimated. To that aim, we use the retrospective distributions \( y_{k}^{\text{basket}}|Y_{k-1}, Z_k, X_k, \theta \)

\[
\sim N_\theta(m_{k}(Y_{k-1}, Z_k, X_k, \theta), \sigma^2 S_{k}(Y_{k-1}, Z_k, X_k, \theta)); k = 1, \ldots, K,
\]

Proposition 3. Let \( i_k \) be the position in the vector \( Y_k \) assigned to the first dwelling of the basket; \( k = 1, \ldots, K \). Under the conditions of Proposition 3, we have that:
where:

\[ m_k(Y_{k-1}, Z_k, X_k, \theta) = \left( D_k + \sum_{u \geq i_k + R} \tau_u f_u f_u^\prime \right)^{-1} \left( D_k \mu_k + \sum_{u \geq i_k + R} \tau_u f_u (y_u - a_u) \right). \]  

\[ S_k(Y_{k-1}, Z_k, X_k, \theta) = \left( D_k + \sum_{u \geq i_k + R} \tau_u f_u f_u^\prime \right)^{-1}. \]  

with \( D_k = \text{diag}(\tau_{ik}, \tau_{ik+1}, ..., \tau_{ik+R-1}) \), \( \mu_k = (\mu_{u,i_k}, \mu_{u,i_k+1}, ..., \mu_{u,i_k+R-1})' \), \( f_u = (f_{u,i_k}, f_{u,i_k+1}, ..., f_{u,i_k+R-1})' \) with \( \mu_j \) and \( f_{ji} \) given by (6) and (7) and \( a_u = \mu_u - f_u y_{basket} \).

With this result, we can draw a sample from \((p_{k,r}^{basket}; k = 1, ..., K; r = 1, ..., R)\) \(|Y, Z, X\) and, hence, from \((V_1, V_{k})\) \(|Y, Z, X\) using Algorithm 4.

**Algorithm 4**

**Step 0 (Initialization)**

Put \( p_{0,r}^{basket} = p_{0,r}^{basket}; r = 1, ..., R; \ell = 1, ..., n_{sample}. \)

Calculate \( y_{0,r}^{basket} = \log(p_{0,r}^{basket}); r = 1, ..., R; \ell = 1, ..., n_{sample}. \)

**Step 1 (Sampling of \( p_{k,r}^{basket}; k = 1, ..., K \))**

For \( k = 1, ..., K \) and using (11) and (12), draw \( y_{k,r}^{basket}(\ell) \) from:

\[ N_R \left( m_k(Y_{k-1}^{(\ell)}, Z_k, X_k, \theta^{(\ell)}), \frac{1}{p^{(\ell)}} S_k(Y_{k-1}^{(\ell)}, Z_k, X_k, \theta^{(\ell)}) \right) \]

\[ r = 1, ..., R; \ell = 1, ..., n_{sample} \]

where: \{ \( \theta^{(\ell)}; \ell = 1, ..., n_{sample} \) \} is the sample (3).

Calculate \( p_{k,r}^{basket}(\ell) = \exp(y_{k,r}^{basket}(\ell)); r = 1, ..., R; \ell = 1, ..., n_{sample} \)

**Step 2 (Sampling of \( V_k; k = 1, ..., K \))**

For \( k = 1, ..., K \), calculate \( V_k^{(\ell)} = \Sigma_{r=1}^R p_{k,r}^{basket}(\ell), \ell = 1, ..., n_{sample}. \)

Using these samples we calculate:

\[ \left\{ IP_{t_0}^{(\ell)} = 100 \frac{V_k^{(\ell)}}{V_0}; \ell = 1, ..., n_{sample} \right\} \text{ for } k = 1, ..., K. \]  

\[ \text{for } k = 1, ..., K. \]
From (13), we estimate the value of the price index (10) using the median values and calculate approximated Bayesian credibility intervals from the appropriate quantiles.

Application to the Real Estate Market of Zaragoza

We apply the described methodology to the analysis of the housing prices evolution in Zaragoza, Spain. The data correspond to a sample of \( n = 788 \) transactions carried out between November 2002 and December 2004 in an area of Zaragoza (Exhibit 1) and are given with the specific date on which these transactions were registered and are described in Beamonte, Gargallo, and Salvador (2008).

We use the STAR model (1) that contains \( q = 4 \) explanatory variables included in \( Z \): constant, time, the two principal components of the UTM coordinates, and \( u = 14 \) hedonic variables \( X \), namely: living area (logarithm), age (logarithm), subsidized character, form of acquisition, lift, parking, porter, type of heating (two variables), air conditioner, and number of rooms (three variables).

A time indicator is included to collect the evolution of prices given by external factors not gathered by the hedonic variables. Furthermore, in order to avoid multicollinearity problems and after a principal component analysis of these variables, we incorporate two components, PC1 and PC2, instead of the UTM coordinates (Exhibit 1). Finally, and given that a 28% of transactions are affected by incomplete information on some of the independent variables, we include a missing indicator to analyze the existence of this possible bias (Beamonte, Gargallo, and Salvador, 2010b).

Exhibit 2 displays the point estimations of the parameters obtained from the posterior median, as well as the limits of the 95% Bayesian credibility intervals (Beamonte, Gargallo, and Salvador, 2010b). The hedonic variables with significant influence on the model are living area, age of the building, form of acquisition, lifts, and porter services. The signs of the regression coefficients are as expected. Note the systematic upward quarterly trend in the price level of around 2.85%. Global spatial trends, not explained by the hedonic characteristics, are appreciated through the significant effect of PC1 (positive), in the way that the nearer the center of the city, the higher the price of the dwelling. Finally, the missing indicator coefficient is not significant, reflecting a lack of systematic bias due to missing data exhibit that present great uncertainty; however, the parameter \( m_s \) is clearly different from 0, revealing significant spatial neighborhood effects. The estimation of \( \lambda \), the weight of the spatial neighbors, presents high values, around 0.88, roughly assigning the same importance to the information provided by every influential neighbor over the price of a dwelling. The value of \( \gamma \), the weight of temporal neighbors, is around 0.72. The estimation of the autoregressive parameters, \( \phi \), shows that only \( \phi_s \), which collect spatial effects, are significantly positive, which matches the statistical significance of \( m_s \) and \( \lambda \) previously
Exhibit 1 | Zaragoza, Spain

Left: Map of the city of Zaragoza. The area under study is within the black rectangle. Right: The area under study with the locations of the specific transactions (scale 1:25,000), along with the principal components. The right panel highlights the existence of a strong negative correlation between the UTM X and Y coordinates of the dwellings in the dataset. For that reason, a principal component analysis was carried out in order to avoid multicollinearity problems between these two independent variables in the model (1).
### Exhibit 2 | Estimation of the Parameters of the STAR model (1)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Q 2.5</th>
<th>Median</th>
<th>Q 97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>Constant in the model</td>
<td>-0.6476</td>
<td>5.1031**</td>
<td>106.2440</td>
</tr>
<tr>
<td>Time linear trend</td>
<td>Temporal linear trend (expressed in quarters)</td>
<td>0.0100*</td>
<td>0.0285*</td>
<td>0.0499*</td>
</tr>
<tr>
<td>PC1</td>
<td>Components obtained on a Principal Component Analysis from the UTM coordinates</td>
<td>0.0001*</td>
<td>0.0002*</td>
<td>0.0002*</td>
</tr>
<tr>
<td>PC2</td>
<td></td>
<td>-0.0003</td>
<td>-0.0001</td>
<td>0.0001</td>
</tr>
<tr>
<td>Log (area)</td>
<td>Living area of the building (m²)</td>
<td>0.6626*</td>
<td>0.7513*</td>
<td>0.8367*</td>
</tr>
<tr>
<td>Log (age)</td>
<td>Age of the building (years)</td>
<td>-0.2226*</td>
<td>-0.1699*</td>
<td>-0.1254*</td>
</tr>
<tr>
<td>Subsidized</td>
<td>State subsidized dwelling</td>
<td>-0.0246</td>
<td>0.0279</td>
<td>0.0780</td>
</tr>
<tr>
<td>Form of acquisition</td>
<td>Form of acquisition of the dwelling (purchase or other)</td>
<td>0.3999*</td>
<td>0.4628*</td>
<td>0.5266*</td>
</tr>
<tr>
<td>Air conditioner</td>
<td>Air conditioning in the dwelling</td>
<td>-0.0263</td>
<td>0.0532</td>
<td>0.1262</td>
</tr>
<tr>
<td>Collective heating</td>
<td>Type of heating in the dwelling</td>
<td>-0.0906</td>
<td>-0.0096</td>
<td>0.0700</td>
</tr>
<tr>
<td>No heating</td>
<td></td>
<td>-0.1236</td>
<td>-0.0498</td>
<td>0.0213</td>
</tr>
<tr>
<td># rooms &lt; 4</td>
<td>Number of rooms in the dwelling</td>
<td>-0.0941</td>
<td>0.0086</td>
<td>0.1201</td>
</tr>
<tr>
<td># rooms = 4</td>
<td></td>
<td>-0.1033</td>
<td>-0.0381</td>
<td>0.0283</td>
</tr>
<tr>
<td># rooms &gt; 5</td>
<td></td>
<td>-0.1077</td>
<td>-0.0054</td>
<td>0.1022</td>
</tr>
<tr>
<td>Porter</td>
<td>Porter in the building (1 = Yes, 0 = No)</td>
<td>-0.0030</td>
<td>0.0696**</td>
<td>0.1359</td>
</tr>
<tr>
<td>Garage</td>
<td>Garage in the building (1 = Yes, 0 = No)</td>
<td>-0.1180</td>
<td>-0.0396</td>
<td>0.0375</td>
</tr>
<tr>
<td>Lift</td>
<td>Lift in the building (1 = Yes, 0 = No)</td>
<td>0.1159*</td>
<td>0.1729*</td>
<td>0.2302*</td>
</tr>
<tr>
<td>Ind. missing</td>
<td>Missing indicator (1 = Yes, 0 = No)</td>
<td>-0.0919</td>
<td>-0.0125</td>
<td>0.0658</td>
</tr>
<tr>
<td>$\phi_7$</td>
<td>Spatial, temporal and spatio-temporal autoregressive coefficients</td>
<td>-0.1181</td>
<td>0.0722</td>
<td>0.2199</td>
</tr>
<tr>
<td>$\phi_5$</td>
<td></td>
<td>0.1219*</td>
<td>0.3319*</td>
<td>0.6124*</td>
</tr>
</tbody>
</table>
**Exhibit 2 | (continued)**

Estimation of the Parameters of the STAR model (1)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Q 2.5</th>
<th>Median</th>
<th>Q 97.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_{ST}$</td>
<td></td>
<td>$-0.4603$</td>
<td>$-0.0210$</td>
<td>$0.4229$</td>
</tr>
<tr>
<td>$\phi_{TS}$</td>
<td></td>
<td>$-0.5465$</td>
<td>$0.0309$</td>
<td>$0.5562$</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Parameters of the error distribution</td>
<td>$122.541$</td>
<td>$150.717$</td>
<td>$186.544$</td>
</tr>
<tr>
<td>$r$</td>
<td></td>
<td>$20.977$</td>
<td>$26.59$</td>
<td>$35.138$</td>
</tr>
<tr>
<td>$m_T$</td>
<td>Number of spatial ($m_s$) and temporal ($m_t$) nearest neighbors and the weights assigned to them ($\gamma, \lambda$)</td>
<td>$2$</td>
<td>$29$</td>
<td>$49$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td></td>
<td>$0.1079$</td>
<td>$0.7238$</td>
<td>$0.9863$</td>
</tr>
<tr>
<td>$m_s$</td>
<td></td>
<td>$6$</td>
<td>$31$</td>
<td>$50$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td></td>
<td>$0.6750$</td>
<td>$0.8768$</td>
<td>$0.9769$</td>
</tr>
</tbody>
</table>

Notes: The table shows the point estimation and the lower and upper limits of the 95% Bayesian credibility interval for the parameters of the model (1). The point estimations correspond to the posterior median calculated from the sample (3) and the lower and upper limits of the interval correspond to the 2.5 (Q2.5) and 97.5 (Q97.5) posterior quantiles. When both limits are positive (negative), the value of the parameter is 95% significantly positive (negative). Different signs indicate the non-significance of the parameter. The hedonic variables with significant influence on the model are: living area, age of the building, form of acquisition, lift and porter services. Furthermore, the autoregressive parameter $\phi_{ST}$, which collects spatial effects is significant with positive sign.

* Significant estimation at the 95% credibility level.
** Significant estimations at the 90% credibility level.
Exhibit 3 | Retrospective Comparison of Models (in bold the best behavior)

<table>
<thead>
<tr>
<th>Compared Models</th>
<th>Criteria</th>
<th>Hedonic Homoscedastic</th>
<th>Hedonic Heteroscedastic</th>
<th>STAR Homoscedastic</th>
<th>STAR Heteroscedastic</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.4431</td>
<td>0.4422</td>
<td>0.4451</td>
<td>0.3945</td>
<td></td>
</tr>
<tr>
<td>MAD</td>
<td>0.2947</td>
<td>0.2951</td>
<td>0.3116</td>
<td>0.2790</td>
<td></td>
</tr>
<tr>
<td>LPRED</td>
<td>−625.43</td>
<td>−170.66</td>
<td>−448.92</td>
<td>−167.05</td>
<td></td>
</tr>
<tr>
<td>%LPRED(M, M₀)</td>
<td>49.45%</td>
<td>49.32%</td>
<td>19.77%</td>
<td>—</td>
<td></td>
</tr>
<tr>
<td>COV₀.₉₅</td>
<td>85.21%*</td>
<td>96.03%</td>
<td>95.07%</td>
<td>93.42%</td>
<td></td>
</tr>
<tr>
<td>COV₀.₉₉</td>
<td>89.86%**</td>
<td>99.04%</td>
<td>97.40%**</td>
<td>97.26%**</td>
<td></td>
</tr>
<tr>
<td>WIDTH₀.₉₅</td>
<td>1.1224</td>
<td>1.1974</td>
<td>1.7513</td>
<td>1.1664</td>
<td></td>
</tr>
<tr>
<td>WIDTH₀.₉₉</td>
<td>1.4545</td>
<td>1.5395</td>
<td>2.2620</td>
<td>1.4972</td>
<td></td>
</tr>
</tbody>
</table>

Notes: We have taken M₀ = STAR heteroscedastic as the reference model. The best point forecast behavior (RMSE and MAD) corresponds to the STAR heteroscedastic model (1) with the lowest values. From a global viewpoint (LPRED and %LPRED) the best behavior corresponds, again, to the STAR heteroscedastic model (1) with the largest value on the LPRED criterion and, when taken as the reference model, the %LPRED for the rest of the models are lower than 50%. Heteroscedastic models have an adequate behavior with respect to the COV₀.₉₅ because their values do not differ significantly from 0.95. In addition, and despite the fact that only the hedonic heteroscedastic model has an appropriate behavior with respect to the COV₀.₉₉, the widths of the predictive intervals of the STAR heteroscedastic are lower. Therefore, we conclude that the STAR heteroscedastic model presents the most balanced behavior with respect to all the criteria.

*The best behavior.

The low value of the degrees of freedom parameter r (around 2.66), due to the presence of some observations with large negative error terms, eᵢ (Beamonte, Gargallo, and Salvador, 2010b) that correspond to dwellings acquired at an extremely low price.

Retrospective Analysis

Exhibits 3–5 show the results of a leaving-one-out cross-validation procedure of comparison of models described in the Appendix. We compare four models: the STAR model (1), the hedonic version of this model (with no spatio-temporal
Notes: The continuous line is the observed log-prices; the dashed dotted lines are the limits of 99% Bayesian credibility intervals. The figures show, for each compared model, the log-price forecasts together with the lower and upper limits of the 99% Bayesian credibility intervals calculated as described in the Appendix. Concretely, the log-price forecasts correspond to the posterior means of the leave-one-out cross-validation distributions and have been calculated from expression (A.2). The limits of 99% Bayesian credibility intervals are the 0.5% and 99.5% quantiles of the weighted sample (A.1). The heteroscedastic models provide more adjusted intervals with wider fluctuations closely adapted to the observed transactions prices.

The best predictive behavior corresponds to the heteroscedastic models, both point (RMSE and MAD) and globally (LPRED and %LPRED). In particular, the best
Exhibit 5 | Boxplots of Widths of the Retrospective Cross-validation leaving-one-out 99% Bayesian Credibility Intervals Forecasts for the Models

Note: The figure displays, for every compared model, the boxplot of the 99% Bayesian credibility intervals widths used in the calculation of the WIDTH0.99 criterion described in (A.3) of the Appendix. The distributions corresponding to the heteroscedastic models present more variability with a heavier right tail, better collecting the existence of dwellings with extremely high prices.

Score in RMSE (0.3945), MAD (0.2790), and LPRED (−167.05) belong to the heteroscedastic STAR model which, in addition, dominates the rest of the models according to the %LPRED criterion.

The heavy tails of the predictive distributions of the heteroscedastic models allow the capture of transactions with extremely high or low prices (Exhibits 4 and 5), which provides predictive intervals more adjusted to the evolution of transaction prices, with a good behavior regarding the empirical coverage. Between the heteroscedastic models, the STAR presents the smallest interval widths (Exhibit 3). All these facts show that the spatio-temporal dependences of local character collected by the heteroscedastic STAR model enhance its predictive ability.
A hedonic model that thoroughly uses spatial information (e.g., neighborhood dummies, distances to a variety of locations) and temporal variation in the mean function would have, probably, performed better than the STAR model, making the computation and interpretation of the results easier [see Case, Clapp, Dubin, and Rodriguez (2004) for some relevant results]. However, in some occasions, those types of information are, for different reasons, not available. STAR models can be, in such a situation, an appropriate alternative because the non-parametric
terms are flexible enough to capture the omitted spatio-temporal effects not included in the hedonic part of the model.

Finally, Exhibit 6 displays the retrospective estimated quarterly evolution and the 90% limits of the Bayesian credibility intervals of the dwelling price for 2003 and 2004. The estimation is carried out by means of an index number calculated by using Algorithm 4, with base 100 for the first quarter of 2003. The transactions (\(R = 57\)) that took place in the last quarter of 2002 were used as a representative sample of the sold dwelling set. From this sample, the price is retrospectively predicted by assigning the middle point of the corresponding quarter\(^3\) as the date of the transaction. The point estimation and the lower and upper limits of the Bayesian credibility interval of the index numbers are calculated from the median and the 5% and 95% quantiles, respectively, from samples (13). For comparative purposes, we also show the estimations given by the Ministry of Housing for the city of Zaragoza and the heteroscedastic hedonic model.

Both models give an estimation of around 5.16% for the quarterly growth (22.30% annual); however, the 90% confidence bands are narrower for the STAR model, showing better accuracy. The Ministry of Housing provides a quarterly increase of about 4.57% (19.57% annual). The differences are not statistically significant due to the great uncertainty consequence of the small sample size, as well as to the high leptokurtosis of the predictive distribution of the errors.

**Conclusion**

In this paper we propose a Bayesian statistical methodology to carry out a retrospective analysis of the spatio-temporal evolution of the dwelling price. We use a semi-parametric Bayesian approach based on hedonic models with heteroscedastic spatio-temporal autoregressive (STAR) errors with neighborhood effects, similar to the model proposed in Beamonte, Gargallo, and Salvador (2010b). The methodology allows the elaboration of retrospective predictions, as well as the construction of a retrospective index number that shows the evolution of the dwelling price level in a specific geographical area (quarter, district, city, etc.). This procedure does not depend on asymptotic results and incorporates the uncertainty associated to the estimation of the model parameters.

The methodology is applied to the analysis of the price evolution between the last quarters of 2002 and 2004 in a specific area of Zaragoza. The results reveal the existence of a systematic upward temporal linear trend of around 5.16% (22.30% annual). This estimation essentially coincides with the values provided by the Ministry of Housing for this period.

In this model, only geographical dependences have been considered to select the spatial neighbors; nevertheless, it is reasonable to expect the existence of other spatial dependences linked to structural characteristics and location (distance to schools, commercial areas, etc.). The incorporation of these aspects in the mean
function of the model would improve the results provided by STAR model, making the computation and interpretation of the results easier (Case, Clapp, Dubin, and Rodriguez, 2004). We are working on the inclusion of that information in the construction of the spatial dependence matrices, $S$.

Finally, it would be interesting to combine the process of retrospective valuation proposed in this paper with the SPAR method of elaboration of indexes, which remains as a topic for future work. In addition, we are planning to incorporate more recent data in order to analyze the impact of the economic crisis on the dwelling price evolution.

## Appendix

**Proofs of Propositions 1 and 3**

**Proof of Proposition 1**

Let $Y_i^\rightarrow = (y_j : j > i)$. From (1) we have that:

\[
[y|Y, Z, z, X, x, \theta] \\
\propto [y|Y_i^\rightarrow, Z_i^\rightarrow, z, X_i^\rightarrow, x, \theta][Y_i^\leftarrow, Y_i^\rightarrow, y, Z, z, X, x, \theta] \\
\propto \exp \left[ -\frac{\tau}{2} \tau_i(y - \mu_i)^2 \right] \exp \left[ -\frac{\tau}{2} \sum_{j>i} \tau_j(y_j - \mu_j)^2 \right] \\
= \exp \left[ -\frac{\tau}{2} \left\{ \tau_i(y - \mu_i)^2 + \sum_{j>i} \tau_j(y_j - \mu_j - \alpha_i - f_{ji}y)^2 \right\} \right] \\
\propto \exp \left[ -\frac{\tau}{2} \left\{ y^2 \left( \tau_i + \sum_{j>i} \tau_j f_{ji}^2 \right) \\
- 2y \left( \tau_i \mu_i + \sum_{j>i} \tau_j f_{ji}(y_j - \alpha_i) \right) \right\} \right] \\
\propto \exp \left[ -\frac{1}{2} s^2(\theta)(y - m(\theta))^2 \right],
\]

and, therefore, the thesis is followed. ■

**Proof of Proposition 3**

Let $(Y_k)_i^{<} = \{ y_{k,j} : j < i_k \}$ and $(Y_k)_i^{>} = \{ y_{k,j} : j > i_k + R - 1 \}$.

Let $(Z_k)_i^{<} = (z_{k,j} : j < i_k)$ and $(X_k)_i^{<} = (x_{k,j} : j < i_k)$.
From (1) we have that:

$$
[y^{basket}_k | Y_{k-1}, Z_k, X_k, \theta] 
\propto [y^{basket}_k | (Y_{k-1})_i^u, Z_k, X_k, \theta] [y^{basket}_k]$$

\[= \exp \left[ -\frac{\tau}{2} (y^{basket}_k - \mu_k)'D(y^{basket}_k - \mu_k) \right] \exp \left[ -\frac{\tau}{2} \sum_{u \geq i + R - 1} \tau_f (y_j - \mu_j)^2 \right]
= \exp \left[ -\frac{\tau}{2} \left( (y^{basket}_k - \mu_k)'D(y^{basket}_k - \mu_k)
+ \sum_{u > i + R - 1} \tau_u (y^{basket}_k f_u, f'_u, d^{basket}_k - 2y^{basket}_k f_u (y_u - a_u)) \right) \right]
\propto \exp \left[ -\frac{\tau}{2} (y^{basket}_k - m_k(Y_{k-1}, Z_k, X_k, \theta))'S_k(Y_{k-1}, Z_k, X_k, \theta) \right.
\left. (y^{basket}_k - m_k(Y_{k-1}, Z_k, X_k, \theta)) \right],
\]

and, therefore, the thesis is followed. ■

Retrospective Comparison of Models Procedure

This describes a leave-one-out cross-validation procedure to compare the retrospective predictive behavior of the models. We have considered transactions \(\{y_i = \log(p_i); i = n_0 + 1, \ldots, n\}\) where \((\max\{m_{\text{FT-net}}, m_{\text{SNet-S}}\} < n_0 < n)\) is the number of observations to which we condition the likelihood function of the STAR model in order to diminish the influence of the oldest transactions in the estimation of the parameters. In the empirical example, we have taken the observations corresponding to the last quarter of 2002 \((n_0 = 57)\).

Let \(M\) be the considered model. For each observation \(y_i\) we calculate the leave-one-out posterior distribution \([\theta | Y_{-(i)}, Z_{-(i)}, X_{-(i)}, M]\) where \(Y_{-(i)} = Y - \{y_i\}\), \(X_{-(i)} = X - \{x_i\}\) and \(Z_{-(i)} = Z - \{z_i\}\). To that aim, we use an importance sampling algorithm with the weighted sample:

\[
\{(w^{(l)}, \theta^{(l)}) ; l = 1, \ldots, n_{\text{sample}}\}
\]
being \( w^{(\ell)} \propto [\theta^{(\ell)}|Y_{-i}, X_{-i}, Z_{-i}, M]/[\theta^{(\ell)}|Y, X, Z, M]; \ell = 1, \ldots, n_{\text{sample}} \), where \( \{\theta^{(\ell)}; \ell = 1, \ldots, n_{\text{sample}}\} \) is the sample (3).

The comparison of models is carried out by means of a set of criteria that evaluate the outsampling predictive behavior of the model using point forecasts (RMSE, MAD), predictive intervals (COV, WIDTH), and densities (LPRED and %LPRED). The considered criteria are the following:

Root Mean Square Error:

\[
\text{RMSE}(M) = \sqrt{\frac{\sum_{i=m_0+1}^{n} (y_i - E[y_i|Y_{-i}, X, Z, M])^2}{n-n_0}}.
\]

Mean Absolute Deviation:

\[
\text{MAD}(M) = \frac{\sum_{i=m_0+1}^{n} |y_i - E[y_i|Y_{-i}, X, Z, M]|}{n-n_0},
\]

with:

\[
E[y_i|Y_{-i}, X, Z, M] = \frac{\sum_{\ell=1}^{n_{\text{sample}}} w^{(\ell)} m_i(\theta^{(\ell)})}{\sum_{\ell=1}^{n_{\text{sample}}} w^{(\ell)}} \quad \text{(A.2)}
\]

and where \( m_i(\theta^{(\ell)}) \) and \( s_i^2(\theta^{(\ell)}) \) are calculated as in Proposition 1.

The two criteria above evaluate the outsampling predictive behavior of the model using a point forecast. Both quantify, approximately, the mean relative prediction error of the dwelling price in such a way that, the lower is the value, the better is the predictive behavior of model \( M \). The MAD criterion is more robust to outliers and, for this reason, gives a more reliable evaluation of the predictive behavior of the model.

To evaluate the predictive behavior of the model using Bayesian intervals of a given credibility level, \( 1 - \alpha \) (0.5 < \( \alpha \) < 1), we calculate their empirical coverage, \( \text{COV}_{1-\alpha}(M) \), and their median width, \( \text{WIDTH}_{\alpha}(M) \).
Empirical Coverage:

\[
\text{COV}_{1-\alpha}(\mathbf{M}) = 100 \left( \frac{\sum_{i=n_0+1}^{n} I_{\{y_i(\alpha/2),y_i(1-\alpha/2)\}(y_i)} - n - n_0}{n - n_0} \right),
\]

Median Width:

\[
\text{WIDTH}_{1-\alpha}(\mathbf{M}) = \text{median} \left\{ y_i \left(1 - \frac{\alpha}{2}\right) - y_i \left(\frac{\alpha}{2}\right) ; \ i = n_0 + 1, \ldots, n \right\}, \quad (A.3)
\]

where \(y_i(\alpha)\) denotes the \(\alpha\)-quantile of \(y_i|\mathbf{Y}_{(i)}\), \(\mathbf{X}, \mathbf{Z}, \mathbf{M}\) obtained by Monte Carlo methods.

The empirical coverage measures its success rates and it is expected that \(\text{COV}_{1-\alpha}(\mathbf{M}) \approx 100(1 - \alpha)\) if the behavior is adequate. This assumption is tested by means of the standard statistical test \(z = \sqrt{n - n_0} \text{COV}_{1-\alpha}(\mathbf{M}) - 100(1 - \alpha)/100\sqrt{\alpha(1 - \alpha)}\) whose \(p\)-value is given by \(2\Pr(Z > |z|)\), where \(Z \sim \mathcal{N}(0,1)\). The median width quantifies the accuracy of the predictive intervals.

Logarithm of the leave-one-out cross-validation predictive density:

\[
\text{LPRED}(\mathbf{M}) = \frac{\sum_{i=n_0+1}^{n} \text{LPRED}_i(\mathbf{M})}{n - n_0}.
\]

Percentage of observations with larger logarithm of the leave-one-out cross-validation predictive density that a reference model \(\mathbf{M}_0\):

\[
\%\text{LPRED}(\mathbf{M}, \mathbf{M}_0) = 100 \left( \frac{\sum_{i=n_0+1}^{n} I_{\{\text{LPRED}(\mathbf{M}) > \text{LPRED}(\mathbf{M}_0)\}(i)} - n - n_0}{n - n_0} \right),
\]
where \( \text{LPRED}_i(M) = \log(\frac{y_i|Y_{-i}, X, Z, M)}{\text{LPRED}_i(M)}) \) with:

\[
y_{i|Y_{-i}, X, Z, M} \sim \frac{\sum_{l=1}^{\text{n_{sample}}} w^{(l)} N(m_l^{(l)}, s_l^{(l)})}{\sum_{l=1}^{\text{n_{sample}}} w^{(l)}}
\]

and where \( I_A(x) = 1 \) if \( x \in A \) and 0 otherwise, denotes the indicator function of the set \( A \).

LPRED evaluates the predictive behavior of the model by measuring its ability to explain, from an outsampling viewpoint, the log of the observed price. This criterion is based on the evaluation of the density function \([y_i|Y_{-i}, X, Z, M]\) that provides a global evaluation of \( y_i \) by taking into account the extent to which this observation is consistent with the uncertainty predicted by model \( M \). The larger the value, the better the ability of model \( M \) to explain the observed values \( \{y_i; i = n_0 + 1, ..., n\} \).

Moreover, LPRED criterion gives an aggregated evaluation of the predictive behavior and, hence, can be influenced by outliers due to the extreme negative value that the corresponding term \( \text{LPRED}_i(M) \) reaches in such situations. For this reason, we complete the comparison process with the use of \( \%\text{LPRED}(M, M_0) \) criterion that compares the behavior of model \( M \) with respect to a model of reference \( M_0 \), by assigning the same importance to each observation \( y_i \) in the evaluation process. When \( \%\text{LPRED}(M, M_0) > 50\% \) we conclude that model \( M \) predicts better than \( M_0 \) since it explains more observed values in terms of the LPRED criterion.

Endnotes

1 The price distribution usually presents right skewness; therefore, the dependent variable is transformed logarithmically to increase the degree of normality of the data.

2 When one or several transactions took place at the same date at which we want to predict the retrospective price of a dwelling, all of them will be considered to belong to its future.

3 The exact dates used for every quarter were February 14, May 15, August 15, and November 15.

References


---

A. Beamonte, University of Zaragoza, Zaragoza, Spain.
P. Gargallo, University of Zaragoza, Zaragoza, Spain.
M.J. Salvador, University of Zaragoza, Zaragoza, Spain or salvador@unizar.es.