AStA Advances in Statistical Analysis (2024) 108:159–183
https://doi.org/10.1007/s10182-023-00476-w

ORIGINAL PAPER

A spatial semiparametric M-quantile regression for hedonic price modelling

Francesco Schirripa Spagnolo[1] · **Riccardo Borgoni**[2] · **Antonella Carcagnì**[2] · **Alessandra Michelangeli**[2] · **Nicola Salvati**[1]

Received: 11 July 2022 / Accepted: 11 March 2023 / Published online: 30 March 2023

Abstract

This paper proposes an M-quantile regression approach to address the heterogeneity of the housing market in a modern European city. We show how M-quantile modelling is a rich and flexible tool for empirical market price data analysis, allowing us to obtain a robust estimation of the hedonic price function whilst accounting for different sources of heterogeneity in market prices. The suggested methodology can generally be used to analyse nonlinear interactions between prices and predictors. In particular, we develop a spatial semiparametric M-quantile model to capture both the potential nonlinear effects of the cultural environment on pricing and spatial trends. In both cases, nonlinearity is introduced into the model using appropriate bases functions. We show how the implicit price associated with the variable that measures cultural amenities can be determined in this semiparametric framework. Our findings show that the effect of several housing attributes and urban amenities differs significantly across the response distribution, suggesting that buyers of lower-priced properties behave differently than buyers of higher-priced properties.

Keywords Robust method · Hedonic regression · Penalised splines · Spatial data

✉ Francesco Schirripa Spagnolo
francesco.schirripa@unipi.it

Riccardo Borgoni
riccardo.borgoni@unimib.it

Antonella Carcagnì
antonella.carcagni@unimib.it

Alessandra Michelangeli
alessandra.michelangeli@unimib.it

Nicola Salvati
nicola.salvati@unipi.it

[1] Università di Pisa, Pisa, Italy

[2] Università Degli Studi di Milano-Bicocca, Milan, Italy

1 Introduction

In the hedonic approach, the price of a house is interpreted as a market evaluation of the particular package of characteristics embodied in it using a hedonic price function (Gravel et al. 2006). The heterogeneity of preferences amongst households, however, implies that the estimated implicit prices of housing characteristics for different points of the house price distribution may be not constant. To date, some researchers have used quantile regression (Koenker and Bassett 1978) to capture consumer heterogeneity in housing demand or identify housing submarkets (Amédée-Manesme et al. 2017).

This paper is the first attempt to use the M-quantile approach to model the distribution of housing prices conditional on housing attributes and estimate the implicit prices of these attributes at different points of the housing price distribution. We focus on the Milan apartment market. Similar to other large cities, the apartment market in Milan is characterised by complex dynamics reflecting the heterogeneity of such a large city in terms of neighbourhood, building and population features. The Milan residential market has already been investigated in several fairly recent studies (Michelangeli and Zanardi 2009; Brambilla et al. 2013; Borgoni et al. 2018b, 2019). In particular, in one of these studies (Borgoni et al. 2018b), the hedonic approach was used to estimate the effect of culture, public transport, education and environmental conditions on the average housing market value in the city of Milan.

In this paper, we implement a hedonic framework and we apply a statistical model based on the M-quantile regression to obtain a robust estimation of the hedonic price function accounting for the heterogeneities of the price market mentioned above whilst preserving the efficiency of the regression parameters (Alfò et al. 2017).

M-quantile regression (Breckling and Chambers 1988) is a robustified 'quantile-like' approach based on an influence function that can be used to grasp the differential effect of covariates at different levels of the conditional distribution of the response variable. This approach also allows a different set of regressors at different levels of the response function to be specified and, as will be made clear later in this paper, encompasses a wide variety of models, ranging from expectile regression (Newey and Powell 1987), to ordinary multiple regression and quantile regression (Koenker and Bassett 1978), hence providing a very rich and flexible tool for empirical market price data analysis. In fact, M-quantile regression can also be seen as a combination of quantile and expectile regression aiming at combining the robustness properties of quantiles with the efficiency properties of expectiles (Alfò et al. 2017). Hence, the suggested approach is able to provide robust estimators of the parameters of interest whilst preserving their efficiency.

In addition, M-quantile regression is a methodology nonparametric in nature and permits one to avoid the ubiquitous log transformation of the response variable typically adopted in the usual regression analysis of house prices. As will be discussed in more detail at the end of the paper, the logarithmic transformation approach has major disadvantages when one is interested in estimating the implicit prices of amenities, since it is necessary to back transform the log-prices on the original scale, thus introducing bias in estimates.

Since its introduction, M-quantile regression has been developed in several directions. Chambers and Tzavidis (2006) suggested that this approach can be an alternative to the mixed effect model in the small area estimation and Chambers et al. (2016) applied the M-quantile regression for binary data in this context. To account for the hierarchical structure of many datasets, Tzavidis et al. (2016) and Borgoni et al. (2018a) extended the M-quantile regression approach to two- and three-level random effect models, respectively, and Schirripa Spagnolo et al. (2020) included sampling weights in the M-quantile random-effects regression estimation procedure. Alfò et al. (2017) developed a finite mixture of quantile and M-quantile regression models for heterogeneous and/or dependent/clustered data. A semiparametric specification of M-quantile regression has been obtained by including univariate and bivariate spline components in the linear predictor to capture nonlinearities or to account for spatial trends (Pratesi et al. 2009; Dreassi et al. 2014).

In this paper, we also include parametric and semiparametric components in the model to account for the nonlinear effects of some predictors on price formation (Brunauer et al. 2013). As mentioned above, the spatial component is fundamental in determining house prices. Hereafter, we mainly investigated the spatial variability of prices induced by spatial trends that are modelled in a flexible manner using appropriate basis functions.

Moreover, we propose a method to determine the implicit price associated with the attribute modelled by the semiparametric component. The implicit price corresponds to the partial derivative of the hedonic price function for every quantile of interest. We show how to calculate this derivative when a semiparametric component is included in the model.

Our empirical findings show that several housing attributes vary greatly across the response distribution suggesting that buyers of lower-priced properties behave differently than buyers of higher-priced properties.

The remainder of this paper is organised into six sections. Section 2 presents an overview of the theoretical framework of hedonic price modelling. In Sect. 3, the dataset employed for the analysis is described. The statistical methodologies applied in this paper are discussed in detail in Sect. 4. Section 5 presents the empirical results. Conclusions are summarised in the last section of the paper.

2 Hedonic price modelling

In hedonic price theory, housing is viewed as a bundle of utility-bearing characteristics that are usually divided into housing-specific attributes and (dis)amenities, i.e. local-specific characteristics with a (negative) positive impact on household utility. Accordingly, in the hedonic approach, the price of a house is interpreted as a market evaluation of the particular package of characteristics embodied in it using a hedonic price function.

Rosen (1974) first developed a partial equilibrium model, where the supply of housing units is supposed to be fixed. This implies that housing prices are entirely demand driven. The theoretical framework shows that when an individual chooses a housing unit to buy, he implicitly decides the best combination of housing-specific attributes and local amenities according to his preferences and budget constraints.

At equilibrium, households, who are price takers, select the preferred housing unit by equalising their marginal evaluation of each housing characteristic to the hedonic price implicitly determined by the housing market.

2.1 Theoretical framework

Mathematically, the hedonic price corresponds to the first derivative of the hedonic price function with respect to a given characteristic. In the case of a representative consumer in the housing market, or equivalently, assuming that all households are equal; the implicit price associated with a characteristic is the consumer's marginal willingness to pay for an additional amount of that characteristic at the consumer's optimal choice. For a given housing unit, let $\boldsymbol{x} \in \mathbb{R}^H$ be the vector of the unit's characteristics considered as normal goods. The representative consumer preferences are represented by an increasing and strictly concave utility function $U(\boldsymbol{x}, w)$, where $w \in \mathbb{R}_+$ is the composite good assumed to be the numéraire. Let $P(\boldsymbol{x})$ be the equilibrium price schedule associated with the housing unit with attributes $\boldsymbol{x}$. The optimal bundle corresponds to the solution of the following problem:

$$\underset{(X,w)\in R_+^{H+1}}{max} U(\boldsymbol{x},\ w) \qquad \text{s.t.}\ \ m \geq P(\boldsymbol{x}) + w,$$

where m represents the consumer's monetary resources. First order conditions for the internal solution $(\boldsymbol{x}^*, w^*)$ imply the following set of equations:

$$\frac{\partial P(\boldsymbol{x}^*)}{\partial x_h} = \frac{U(\boldsymbol{x}^*, w^*)_{x_h}}{U(\boldsymbol{x}^*,\ w^*)_w}, \qquad \forall\, h = 1, \ldots, H,$$

where $P(\boldsymbol{x}^*) = m - \boldsymbol{x}^*$, $U(\cdot)_{x_h}$ is the consumer's marginal utility associated with the unit's characteristic x_h, and $U(\cdot)_w$ is the marginal utility associated with the numéraire. At the optimum, the marginal substitution rate between x_h and the numéraire is equal to the marginal willingness to pay for an additional amount of x_h.

In empirical applications, the classical estimation of the hedonic price function by ordinary least squares fits the representative agent framework since the estimated implicit price is a measure, on average, of the impact of each characteristic on housing prices (Zietz et al. 2008). However, when the representative consumer assumption is removed to analyse the market with heterogeneous households, it is likely that housing attributes are valued quite differently across the conditional price distribution. As we briefly review in the next section, several studies have provided empirical evidence that confirms these differences.

A further aspect that is worth mentioning is that the theoretical model sketched in this section naturally leads to a nonlinear hedonic price structure (Malpezzi 2002). This means that the marginal willingness to pay for a given characteristic of the house is not constant (see Fritsch et al. 2016 and references therein). As shown by Freeman (1993), the curvature of the hedonic price function could be convex, concave or linear and it is generally accepted that the hedonic price function is nonlinear (Kostov 2009). In our empirical application, we address this issue by using splines and polynomial terms for modelling nonlinearities.

2.2 The use of QR-based models for the house price function

As mentioned in the previous section, quantile regression has been used for a long time in modelling house prices to capture consumer heterogeneity in housing demand or identify housing submarkets (Amédée-Manesme et al. 2017). The heterogeneity of household preferences implies that the estimated implicit prices of housing characteristics for different points of the house price distribution may not be constant. Quantile regression allows one to determine the extent to which housing characteristics are valued differently across the distribution of housing prices (Mak et al. 2010). The variation across the price distribution is referred as vertical market segmentation. The price surface has irregularly distributed spatial subcentres, which is referred to horizontal housing market segmentation (Fritsch et al. 2016).

Bayer et al. (2004) found that the marginal willingness to pay for desirable housing characteristics and neighbourhood amenities increases with income and that the housing preferences of poor and wealthy households differ (Leung and Tsang 2012). Uematsu et al. (2013) employed a quantile regression approach to investigate the potentially heterogeneous impact of natural amenities on farmland values in the USA. Including regional dummies in the model specification allow for the estimation of the differences in farmland values across regions. Chasco and Le Gallo (2015) and Chasco and Sánchez (2015) evaluated the impact of air pollution and urban noise. Huang (2018) focused on schools, whilst Diao et al. (2018) examined at the effect of rail infrastructure. Diao et al. (2018) estimated the effect of public transport on housing prices. The findings of such studies show important variations in the willingness to pay for better conditions in these amenities. Waltl (2019) combined penalised quantile regression models with the hedonic imputation approach to construct house price indices.

The spatial dimension plays a central role when one wants to address house price dynamics. The spatial component has been introduced only recently in quantile modelling of house prices (Trzpiot 2012; McMillen 2012). An increasing number of studies have used spatial econometrics to control for spatial dependence and spatial heterogeneity (Wan et al. 2017). For example, Kostov (2009) applied a spatial lag quantile regression to a hedonic land prices model. This allows for varying effects of the hedonic characteristics and varying degrees of spatial lag autocorrelation. McMillen (2012, 2015) used a conditionally parametric quantile model accounting for local variation in an overall spatial trend. The advantage of this model is that it is computationally feasible for quite larger datasets. Moreover, the author showed a series of graphs that make easy to illustrate the effects of discrete changes in the explanatory variables on the distribution of the dependent variable. Fritsch et al. (2016) incorporated a semiparametric approach into the quantile regression framework to flexibly account for nonlinear covariate effects when studying the rental housing market in the German city of Regensburg. Wan et al. (2017) and Tomal and Helbich (2022) proposed space varying coefficient quantile regressions to examine the heterogeneity of the marginal effects of attributes across the distribution of housing prices. This approach which allows the coefficients to vary with a variable not included in the linear predictor, permits nonlinear interactions between this effect modifier and the other covariates, providing a flexible tool to investigate price heterogeneity. In the latter paper, a spatial autoregressive geographically weighted quantile regression was proposed to explore housing rent determinants in Amsterdam and Warsaw, showing that housing rent determinants vary over space and the price distribution.

3 Data description

The data come from different sources and are combined into a unique dataset to take advantage of the geocoding of any single data source.

Housing market data are from the Real Estate Observatory (Osservatorio del Mercato Immobiliare). The dataset is composed of 4000 individual housing transactions in Milan that occurred between 2004 and 2010. In addition to housing market values, the dataset provides information about the main characteristics of the house units recorded in the sample and discussed below. Housing units in the sample are spatially identified by their civic address. Each civic address is geocoded by its UTM coordinates using Java script to retrieve this information from Google Maps geographical databases. This allows us to add geocoded data on urban amenities to the housing transaction dataset. Urban amenity variables are taken from the open data portal of the municipality of Milan and the Regional Environmental Protection Agency (ARPA) of the Lombardy region. In particular, we consider the availability of public transport, education, cultural activities and related infrastructures and the presence of abandoned areas. Finally, to control for the effect of the financial crisis of 2008 on the housing market, a binary variable that identifies all the transactions that occurred before and after this year has also been defined. More specifically, this variable is equal to 1 if the housing unit has been sold in the postcrisis sample period (2009 onwards), and it is equal to 0 otherwise. The list, description and definition of the variables used in the empirical analysis are given in Table 1. Descriptive statistics for the price and explanatory variables are provided in Table 2. This dataset description is completed with a few additional comments on some of the variables hereinafter.

3.1 Housing-specific characteristics

We consider the following housing-specific attributes: the total floor area, floor level, presence of a second bathroom or more, presence of an elevator, whether the housing unit has an independent heating system, the presence of a garage and the age of the building. Regarding the floor level and the building's age, we adopt the same coding suggested by Michelangeli and Zanardi (2009), namely, the floor level is divided into three levels (the house is the on ground floor or first floor; the house is on the second floor or third floor; the house is on the fourth floor or higher); the building's age is divided into two levels according to whether the unit was built before or after 1950.

Table 1 Variables description

Variable	Value
Housing-specific characteristics	
Annual market value	Annual market value in Euros
Total floor area	Positive real values in square metres
Lift	1: at least one elevator; 0: otherwise
Parking area	1: the house has a parking place or garage; 0: otherwise
Bathroom	1: two or more bathrooms; 0: otherwise
Floor	Coded on three levels: house is located on the ground floor or first floor (low floor); house is located on the second floor or third floor (medium floor); house is located on the fourth floor or higher (high floor)
Heating system	1: autonomous heating system; 0: otherwise
Age of the building	1: if the unit is in a property building constructed before 1950; 0: otherwise
Urban amenities	
Cultural Catalyst	Positive real values
Metro	1: if the distance from the nearest metro station is not larger than 680 m that represents the first quartile of the sampling distances; 0: otherwise. Hence, this variable identifies those houses that are closer to the metro line.
University	1: if the distance from the nearest university site is not larger than 441 m which represents the first quartile of the sampling distances; 0: otherwise. Hence, this variable identifies those houses that are closer to universities.
Abandoned area	1: if the distance from the house is 200 m; 0 otherwise
Year	1: sold in the postcrisis sample period (2009-2010); 0 otherwise

3.2 Urban amenities

Culture. Cultural amenities and related infrastructure are measured by the Cultural Catalyst developed by Borgoni et al. (2018b). It is a composite indicator of the following cultural amenities: theatres, museums, libraries and auditoria. The Cultural Catalyst is obtained in two steps: in the first step, an accessibility index for the four cultural amenities is constructed according to the following equation:

$$\tilde{v}_j = \sum_{s=1}^{S_j} w_{js} e^{(-\gamma d_{hjs})},$$

where $\tilde{v}_j$ is the variable measuring the accessibility to amenity v_j; S_j is the total number of locations of amenity v_j, w_{js} is the weight associated with v_j constantly set equal to 1 (i.e. $\tilde{v}_j$ is a weighted total of the amenity in the study areas); $e^{(-\gamma d_{hjs})}$ is a distance-decay function; d_{hjs} is the Euclidean distance (in metres) between housing unit h and site s where amenity v_j is located; the parameter γ is selected via cross-validation, as suggested by Borgoni et al. (2018b). In the second step, a principal

Table 2 Housing prices (in euros) summary statistics by dwelling characteristics

Variables	*N*	Mean	Sd	Min	Q25	Median	Q75	Max
Floor								
Low Floor	1312	12092	13666	3600	5760	7821	12836	129509
Medium Floor	1215	10379	9492	3617	5456	6859	11343	119922
High Floor	1419	12064	13117	3604	5696	7656	12674	128603
Elevator								
No Elevator	708	7723	6903	3604	5035	6105	7805	126734
Elevator	3238	12392	13085	3600	5875	8077	13408	129509
Heating system								
Centralised	3470	11494	11974	3600	5615	7443	12342	128603
Autonomous	476	11999	14733	3604	5849	7756	12244	129509
Parking area								
No parking	3909	11490	12254	3600	5633	7428	12257	129509
Parking	37	18364	18273	4068	8535	12515	19832	98375
Bathroom								
1-Bathroom	2859	8460	7901	3600	5193	6371	8750	128603
Bathroom>1	1087	19695	17240	3873	9468	13804	23479	129509
Age of the building								
Built after 1950	3031	10344	9841	3600	5567	7217	11170	127619
Built before 1950	915	15565	17751	3604	5919	8734	18614	129509
Abandoned area								
> 200 m	3471	11730	12710	3600	5672	7510	12409	129509
< 200 m	475	10274	9094	3611	5362	7079	11465	74266
University								
Far	2959	10200	11489	3604	5470	6929	10498	129509
Near	987	15617	13823	3600	6275	10617	19583	93364
Metro								
Far	2959	11297	12268	3600	5592	7293	11889	129509
Near	987	12327	12521	3618	5790	7924	13615	120024
Year								
Pre crises	2775	11605	12125	3604	5772	7762	12445	129509
Post crises	1171	11434	12835	3600	5402	6694	11893	120464

component (PC) analysis is computed via a single-value decomposition of the correlation matrix of the four accessibility variables. Only the largest eigenvalue is found to be significantly larger than 1; hence, the first PC defines the Cultural Catalyst, and the first PC scores are the sampling values of the index.

Public transport Accessibility to the dwelling is accounted for by the distance to the nearest metro station from each housing unit. A georeferenced map of the metro stations is available from the open data portal of the municipality of Milan. A binary accessibility index is calculated for each housing unit according to whether this distance is larger than the third quartile of all the distances associated with each dwelling in the sample.

Education As in Brambilla et al. (2013); Garretsen and Marlet (2017), universities are considered a proxy for education.[1] There are 710 university sites in the municipality of Milan, spread across the municipality area, belonging to seven main institutions and four academies of arts and design. It is assumed that a potential dweller considers the proximity to a specific higher education institution that he is interested in rather than to a variety of different institutions when making a residence choice; thus, we construct a proximity index to capture this effect. More specifically, we first geolocate all the university sites and then, calculate the distance between each sample unit and its nearest university site. Finally, to better identify those sites more exposed to the effect of university proximity from the others, we derive a binary index. A value of one if taken if such distance is in the bottom 25% of all the distances calculated for the sample units and 0 otherwise.

Urban slum areas It is expected that the presence of ruined or degraded buildings or slum areas may negatively impact on the prices of nearby houses. Information on the location of abandoned buildings and areas (private, productive or natural) is available from the open data portal of the municipality of Milan, which provides a shape file reporting the UTM coordinates of those sites. To calculate a dismissed area index, the Euclidean distance from each house in the sample and each abandoned site is calculated, and a dummy variable indicating whether at least one ruined site is present within a distance of 200 ms from each sample unit is constructed to account for the potential impact of neighbouring degradation.

3.3 Preliminary analysis

Table 2 shows some summary statistics of the variables described above, and it also describes how the house price conditional distribution changes according to their levels. For instance, the difference between the first sample quartile of the prices of housing units with one bathroom and the first sample quartile of housing units with two or more bathrooms is €4,305, whereas this difference is €14,729 in the third quartile. In percentage terms, the first sample quartile of housing prices with two or more bathrooms is 83% higher than the same quartile of one-bathroom houses; this spread increases to 168% in the third quartile. Looking at the parking area, the difference between the first quartile of housing prices of units with and without a parking area is €2,902, and it increases to €7,575 in the third sample quartile. This suggests that having two or more bathrooms or a parking area has a much larger impact for high-valued houses than for less expensive houses. Similar patterns are found for other variables in Table 2.

From this preliminary analysis, a quantile-like approach seems more appropriate than ordinary multiple regression to account for possible variations in the implicit prices along the house price distribution.

[1] Unfortunately, we do not have information on other variables for the quality of education, such as the percentage of pupils moving up to a higher class or parameters for classroom and/or building facilities.

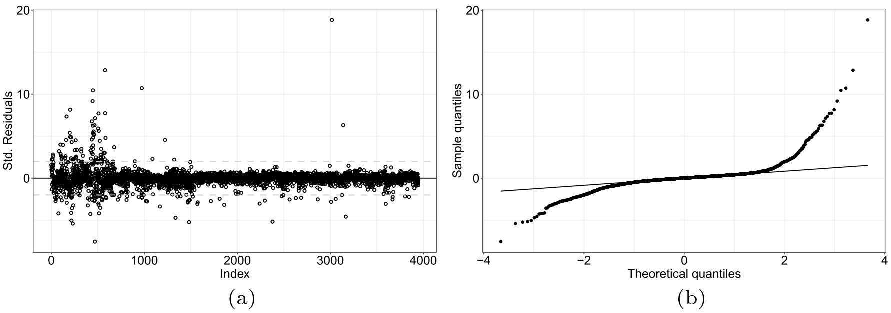


Fig. 1 Semiparametric linear model diagnostic: standardised residual plot (**a**) and standardised residual Normal probability plot (**b**)

To assess the need for using a robust approach, the standard semiparametric linear model below has been preliminary estimated:

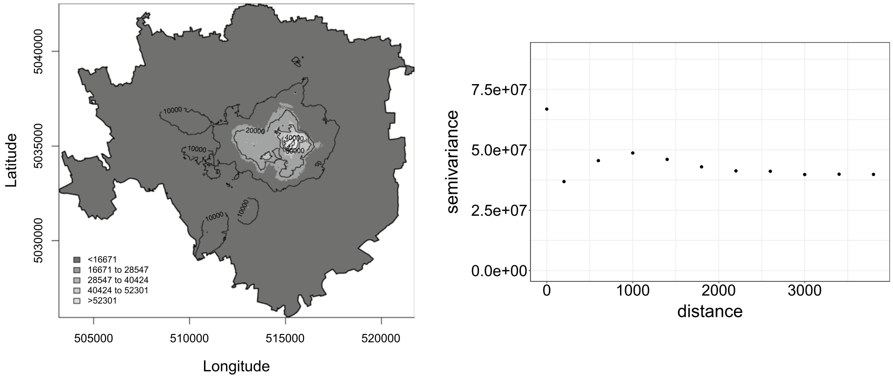
$$\mathrm{E}\big(y|\boldsymbol{x}, t, \mathbf{s}_i\big) = \boldsymbol{x}'\boldsymbol{\beta} + f_1(t) + f_2(\mathbf{s}_i). \tag{1}$$

$f_1(\cdot)$ and $f_2(\cdot)$ are nonlinear functions represented by spline terms that will be discussed in detail in Sect. 4; $\mathbf{s}_i \in \mathbb{R}^2$ are the geographical coordinates of unit i; $\boldsymbol{x}$ contains the set of variables listed in Table 1. The actual specification of each variable included in the additive predictor is clarified ins Sect. 5. t is the Cultural Catalyst that is expected to impact Y nonlinearly.

We show in Fig. 1(a) the plot and in Fig. 1(b) the normal probability plot of standardised residuals of the model above. These two plots indicate that the normality assumption adopted in standard semiparametric modelling does not hold. This is confirmed by the Shapiro test, for which the null hypothesis of normality is rejected (p-*value* $\simeq 0$). Moreover, looking at the plots, outliers are easily detectable. The proportion of outliers, i.e. standardised residuals greater than ±2, is approximately 4%.

To evaluate the spatial dependence in the data, we report the spatial pattern of the house price Cultural Catalyst obtained by smoothing the observed values from the sampling locations via inverse distance weighted interpolation in Fig. 2(a). The map shows that a well-defined spatial structure and larger values are expected to occur towards the city centre. Figure 2(b) shows that the empirical variogram of the residuals of the semiparametric linear model described above. The variogram generally appears to be constant when considered at different distances (a situation known as a *pure nugget* in geostatistics), suggesting that the residuals do not show any spatial dependence once the effect of regionalised variables as well as the impact of the spatial trend has been taken into account.

We also consider the residuals obtained from a regression model where prices are taken on the log scale. Additionally, in this case, the normality assumption does not hold (the value of the Shapiro test is equal to $W = 0.672$ with p-*value* $\simeq 0$) and the percentage of outlying observations remains substantially unchanged. This suggests that the log transformation, ubiquitously adopted in hedonic price analysis, is not appropriate to compensate for the presence of outlying observations or the lack of Gaussianity of the price data.



(a) (b)

Fig. 2 Spatial pattern of the house price (**a**) and semivariogram of residuals of the semiparametric linear model (**b**)

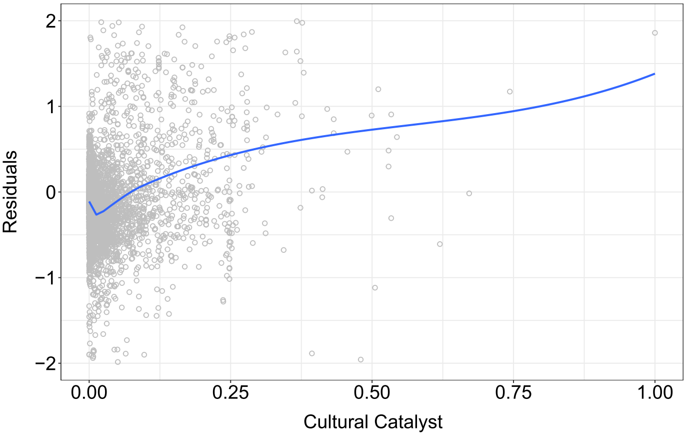


Fig. 3 Residuals of the SSPMQ model where the Cultural Catalyst was removed from the linear predictor plotted versus the Cultural Catalyst

Robust estimation has been suggested in several papers (see Huggins 1993; Huggins and Loesch 1998) to address the non-normality of the dependent variable. This is achieved using a loss function in the log-likelihood that increases with the regression residuals at a slower rate than the squared loss function. As will be shown in the next section, the M-quantile approach provides a robust and efficient estimator of the hedonic price function without assuming any specific probabilistic model for the data at hand.

Finally, we consider the potential nonlinear impact of the Cultural Catalyst. The need to include nonlinear effects in house price modelling has been discussed previously (see Sect. 2). Figure 3 shows the plot of the residuals of the M-quantile model of order of 0.5 (see Sect. 4 below for details) where the Cultural Catalyst has been removed from the linear predictor. The model residuals are scattered towards the Cultural Catalyst values in the plot, showing that a nonlinear pattern remains with respect to their variability. This suggests adding the spline term in the linear predictor, as discussed above.

4 The spatial semiparametric M-quantile model

M-quantile regression (Breckling and Chambers 1988) is a 'quantile-like' generalisation of regression based on influence functions (M-regression) and can be used to understand the differential effect of a covariate at different levels of the conditional distribution of the response variable. The M-quantile of order q for the conditional density of Y given the set of covariates $\boldsymbol{x}$,$f(y|\boldsymbol{x})$, is defined as the solution $MQ_Y(q|\boldsymbol{x},\psi)$ of the integral equation $\int \psi_q\left[y - MQ_y(q|\boldsymbol{x},\psi)\right]f(y|\boldsymbol{x})dy = 0$ where ψ_q denotes an asymmetric influence function. Given $\boldsymbol{x}$, the linear M-quantile regression model is defined by $MQ_y(q|\boldsymbol{x},\psi)) = \boldsymbol{x}'\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ represents a vector of unknown parameters. The set $\boldsymbol{x}$ includes a range of variables representing housing-specific characteristics and urban amenities described in detail in Sect. 3. Throughout the paper, the influence function is obtained as the derivative of the Huber loss function $\rho_q(r)$ (Huber 2011), which is defined as follows:

$$\rho_q\left(r_{iq}\right) = \begin{cases} 2c\left|r_{iq}\right| - c^2\{qI(y > 0) + (1-q)I(y \leq 0)\} & |r_{iq}| > c \\ r_{iq}^2\{qI(y > 0) + (1-q)I(y \leq 0)\} & \left|r_{iq}\right| \leq c \end{cases} \tag{2}$$

where $I(A)$ is the indicator function of set A, and c is an appropriate tuning constant. Conventionally, in M-regression, the tuning constant is suitably selected to provide a trade-off between robustness and efficiency. Huber (2011) suggested that 'good choices are in the range between 1 and 2'. The default value for c is 1.345, which guarantees 95% efficiency of the estimators under normality and still offers protection against outliers. This value is also used in the rest of the paper. Note that different sets of regressors can be included in the linear predictor at different M-quantiles and that a wide range of models can be obtained by modifying the influence function and/or the tuning constant. For instance, using a square loss function, the linear expectile regression model is obtained if $q \neq 0.5$ (Newey and Powell 1987), whereas setting $q = 0.5$ produces the standard linear regression model. Defining the loss function to be the absolute value function described by Koenker and Bassett (1978) gives the linear quantile regression model. Hence, the approach suggested in this paper provides a very flexible tool for analysing housing market prices. We include a semiparametric component for the cultural catalysis in the model to account for its potential nonlinear effect, which is expected, as discussed by Borgoni et al. (2018a). We also include a smooth bivariate function to capture the spatial trends of the data.

The spatial semiparametric model at the M-quantile q (SSPMQ hereafter) is now given as follows:

$$MQ_Y\big(\boldsymbol{x}, t, \mathbf{s}_i;\psi\big) = \boldsymbol{x}'\boldsymbol{\beta}_q + f_{1q}(t) + f_{2q}(\mathbf{s}_i), \tag{3}$$

where $f_{1q}(\cdot)$ and $f_{2q}(\cdot)$ represent two unknown arbitrary smooth functions. $\mathbf{s}_i \in \mathbb{R}^2$ represents that the geographical coordinates of unit i and t are the Cultural Catalyst.

In the rest of this paper, $f_{1q}(\cdot)$ is a penalised spline that relies on a set of univariate quadratic basis functions, i.e.

$$f_{1q}(t) = \sum_{j=1}^{K_1} b_{1j}(t)\theta_{1jq} \tag{4}$$

where $(b_{1j}(t), j = 1, \dots K_1)$ and θ_{1jq} are the basis function and a M-quantile specific spline coefficients set, respectively. In vector form, the spline is written as

$$f_{1q}(t) = \boldsymbol{x}_t'\boldsymbol{\beta}_{1q} + \boldsymbol{z}'_1\boldsymbol{\gamma}_{1q}, \tag{5}$$

where $\boldsymbol{x}_t' = \left[1, t, t^2\right]$, $z_1 = \left[\left(t - k_j\right)_+^2 \ : j = 1, \dots, K_1\right]$ with $(x^2)_+$ denotes the function $x^2 I\{x > 0\}$, $I\{x > 0\}$ being the indicator function of the set $x > 0$, k_j is the j-th knot of the spline and K_1 is the number of spline knots.

In Equation (3), the function

$$f_{2q}(\mathbf{s}) = \sum_{j=1}^{K_2} b_{2j}(\mathbf{s})\theta_{2jq} \tag{6}$$

is a M-quantile specific bivariate thin plate spline that accounts for the spatial trends in prices; $(b_{2j}(\mathbf{s}), j = 1, \dots K_2)$ and θ_{2jq} are the bivariate basis function and an M-quantile specific spline coefficients set, respectively. In vector form, the spline is specified as follows:

$$f_{2q}(\mathbf{s}) = \boldsymbol{x}_{\mathbf{s}}'\boldsymbol{\beta}_{2q} + z_2'\boldsymbol{\gamma}_{2q}, \tag{7}$$

where $\boldsymbol{x}_{\mathbf{s}}' = \left[1, s_1, s_2\right]$; K_2 is the number of spline knots; z_2' is a row of the $n \times K_2$ spline matrix $\mathbf{Z}_{sp}$, and $\boldsymbol{\gamma}_{2q}$ is a K_2-column vector of M-quantile specific spline coefficients. The bivariate spline matrix is defined (Opsomer et al. 2008) by:

$$\mathbf{Z}_{sp} = \left[C(\mathbf{s}_i - \mathbf{k}_j)\right]_{1 \le j \le K_2}^{1 \le i \le n} \left[C(\mathbf{k}_j - \mathbf{k}_k)\right]_{1 \le j,k \le K_2}^{-1/2} \tag{8}$$

where $\mathbf{k}_j$ and $\mathbf{k}_k$, $j, k = 1, \dots, K_2$, are two-dimensional vectors representing the cartographic coordinates of knots j and k; $\mathbf{s}_i$ is a two-dimensional vector representing the cartographic coordinates of sampling location i; $C(\mathbf{s}) = \|\mathbf{s}\|_2^2 \log \|\mathbf{s}\|_2$, where $\mathbf{s} \in \mathbb{R}^2$; $\|\mathbf{s}\|_2$ is the Euclidean norm of $\mathbf{s}$ in $\mathbb{R}^2$.

In matrix notation, the spline terms in Equation (4) and (6) are $\mathbf{f}_{hq} = \mathbf{B_h}\boldsymbol{\theta}_{hq}$ $h = 1;2,$ where $\mathbf{f_{1q}} = [f_{1q}(t_1) \dots f_{1q}(t_n)]^T$ and $\mathbf{f_{2q}} = [f_{2q}(\mathbf{s}_1) \dots f_{2q}(\mathbf{s}_n)]^T$; $\boldsymbol{\theta}_{hq}^T = (\boldsymbol{\beta}_{hq}^T, \boldsymbol{\gamma}_{hq}^T)$ is the q^{th} M-quantile specific vector of coefficients used in the linear combination, and $\mathbf{B_h}$ is the spline basis regression matrix.

The smooth terms $f_{hq}(t)$, $h = 1;2$ in Equation (3) introduce an identification problem (Wood 2017). To address this problem, we define a column centred matrix: $\tilde{\mathbf{B}}_h = \mathbf{B_h} - \mathbf{1}\mathbf{1}^T\mathbf{B_h}/n$, calculate $\tilde{\mathbf{f}}_{hq} = \tilde{\mathbf{B}}_h\boldsymbol{\theta}_{hq}$ and use $\tilde{\mathbf{f}}_{hq}$ in the semiparametric linear predictor. To simplify the notation, we use $\mathbf{f_{hq}}$ instead of $\tilde{\mathbf{f}}_{hq}$ in the rest of the paper. The nodes of the splines are determined by the cluster separation method `clara`, which is implemented in the `R` software (R Core Team 2020) and applied to the sample values of the geographical coordinates.

The SSPMQ model is estimated via the penalised least squares by solving the following estimation equations (Pratesi et al. 2009):

$$\sum_{i=1}^{n} \psi_q(y_i - \mathbf{z}_i^T\boldsymbol{\eta}_q)\mathbf{z}_i^T + \lambda_{1q}\mathbf{D}_1\boldsymbol{\eta}_q + \lambda_{2q}\mathbf{D}_2\boldsymbol{\eta}_q = \mathbf{0}, \tag{9}$$

where $\boldsymbol{\eta}_q = (\boldsymbol{\beta}_q^T, \boldsymbol{\theta}_{1q}^T, \boldsymbol{\theta}_{2q}^T)^T$, $\mathbf{z}_i^T = (\boldsymbol{x}_i, \mathbf{b}_1(t_i), \mathbf{b}_2(\mathbf{s}_i))$, $\mathbf{D}_1$ and $\mathbf{D}_2$ are two penalty matrices and λ_{1q}, and λ_{2q} are the smoothing parameters estimated via external cross-validation. In particular, the Generalised Cross-Validation (GCV) to be minimised to obtain $\Lambda_q = (\lambda_{1q}, \lambda_{2q})$ is:

$$GCV(\Lambda_q) = \frac{||(\mathbf{I} - \boldsymbol{S}_{\Lambda_q})\mathbf{y}||^2}{(1 - n^{-1}\delta\text{tr}(\boldsymbol{S}_{\Lambda_q}))^2}, \tag{10}$$

where $\boldsymbol{S}_{\Lambda_q}$ is a smoother-type matrix associated with $MQ_Y\left(\boldsymbol{x}, t, \mathbf{s}_i;\psi\right)$, and δ is a penalisation term for the additional degrees of freedom given by the trace of the smoother matrix (Pratesi et al. 2009).

The estimation procedure is as follows:

1. Select an initial value of $\boldsymbol{\eta}_q$.
2. At each iteration step r, calculate the residuals $e_{iq}^{r-1} = y_i - \mathbf{z}_i^T\boldsymbol{\eta}_q$ and the associated weights $\alpha_{iq}^{r-1} = \psi_q(e_{iq}^{r-1})/e_{iq}^{r-1}$.
3. Optimise the GCV(Λ_q) over a fine grid of values of Λ_q to obtain $\Lambda_q^\star = (\lambda_{1q}^\star, \lambda_{2q}^\star)$.
4. Calculate the new weighted penalised least squares estimates as follows:

$$\hat{\boldsymbol{\eta}}_q^T = [\mathbf{Z}\mathbf{A}^{r-1}\mathbf{Z}^T + \lambda_{1q}^\star\mathbf{D}_1 + \lambda_{2q}^\star\mathbf{D}_2]^{-1}\mathbf{Z}^T\mathbf{A}^{r-1}\mathbf{y}, \tag{11}$$

   where $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1,\dots,n}$ and $\mathbf{A}^{r-1}$ is a diagonal matrix of weights with diagonal element α_{iq}^{r-1}.
5. Iterate steps 1-4 until convergence.

This procedure above is implemented in `R` software (R Core Team 2020), and the `R` functions are available from the authors upon request.

From Equation (11) and using simple algebraic manipulation, the variance-covariance matrix of the estimated coefficients of the semiparametric M-quantile regression model may be estimated by:

$$\mathrm{var}(\hat{\boldsymbol{\eta}}_q) = (\mathbf{Z}^T\mathbf{A}\mathbf{Z} + \lambda_{1q}\mathbf{D}_1 + \lambda_{2q}\mathbf{D}_2)^{-1}\mathbf{Z}^T\mathbf{A}\mathbf{Z}(\mathbf{Z}^T\mathbf{A}\mathbf{Z} + \lambda_{1q}\mathbf{D}_1 + \lambda_{2q}\mathbf{D}_2)^{-1}\sigma^2. \tag{12}$$

An estimate of this variance-covariance matrix can be obtained by plugging in the sample estimates of $\boldsymbol{\eta}_q$, of the error variance σ^2 and the final values of the smoothing parameters. An estimate of σ^2 can be obtained by using the minimum absolute deviation method and is given by $\hat{\sigma}^2 = (\mathrm{median}(|y_i - \mathbf{z}_i^T\hat{\boldsymbol{\eta}}_q|)/0.6745)^2$ (Chambers and Tzavidis 2006). The asymptotic theory of the M-quantile coefficients estimators for the nonparametric M-quantile has been discussed in Pratesi et al. (2009). Bianchi et al. (2018) showed that the M-quantile estimates can be obtained via maximum likelihood estimation using the Generalised Asymmetric Least Informative distributed error terms and the authors adapted the usual testing procedures to the M-quantile regression.

The variance in Equation (12) can be used to assess the statistical significance of the spline term by calculating a pointwise variability band around the curve and checking whether it includes the horizontal axis. The variability band is constructed using an approach similar to suggested by Ruppert et al. (2003) for ordinary semiparametric spline regression. Moving from the estimated version of Equation (5) and using simple algebra, we determine the variance-covariance matrix of the spline term as follows:

$$\widehat{\mathrm{var}}(\hat{f}_{1q}(t)) = (\boldsymbol{x}_t, z_1)\mathrm{var}(\hat{\boldsymbol{\theta}}_{1q})(\boldsymbol{x}_t, z_1)^T. \tag{13}$$

The pointwise variability band is given by $\hat{f}_{1q}(t) \pm 2 \cdot \sqrt{\widehat{\mathrm{var}}(\hat{f}_{1q}(t))}$.

5 Modelling housing prices in Milan using semiparametric M-quantile regression

This section presents the regression results. The specification of the SSPMQ model described in Sect. 4 includes housing-specific covariates as well as urban amenities discussed in Sect. 3.[2]

Table 3 shows the estimated coefficients for each quantile, when the penalization term $\delta = 3$ in Equation (10).[3] We set the number of knots K_1 equal to 20 for f_{1q} and $K_2 = 40$ for f_{2q}; the knots are located in the plane using the `clara` algorithm implemented in `R` software. We test the impact of using a different number of knots and have found that the results tend to be very stable. Note that a general rule of thumb is to place one knot every 4 or 5 observations. However, for large datasets, this can lead to an excessive number of knots (and therefore parameters) making the computational burden extremely heavy. Therefore, a maximum number of allowable knots may be recommended. In any case, the number of knots does not seem crucial

[2] To improve the numerical stability of the estimates, the Cultural Catalyst has been scaled between 0 and 1 in all the statistical analyses presented in the paper.

[3] We consider different values of the penalisation term, and the estimated coefficients found to be stable. The results are available upon request.

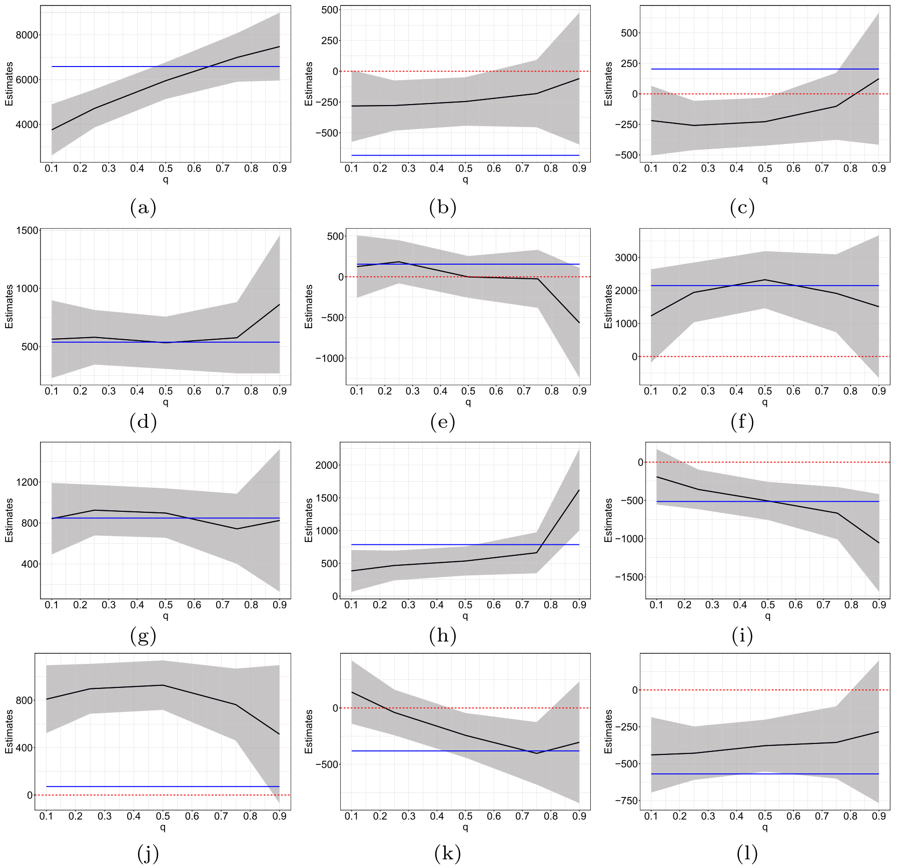
Table 3 Results of SSPMQ ($q = 0.10, 0.25, 0.50, 0.75, 0.90$) estimated using 20 knots for f_{1q} and 40 knots for f_{2q} - $\delta = 3$†

Variable	$\beta_{0.10}$	$\beta_{0.25}$	$\beta_{0.50}$	$\beta_{0.75}$	$\beta_{0.90}$
Housing-specific characteristics					
Intercept	3755.36***	4712.36***	5951.43***	6987.73***	7475.13***
	(582.01)	(435.49)	(418.10)	(553.04)	(773.76)
Floor low	– 282.06*	– 278.25***	– 245.67**	– 181.39	– 59.53
	(148.31)	(103.94)	(100.43)	(139.40)	(273.26)
Floor high	– 218.79	– 258.78***	– 227.98***	– 101.96	124.35
	(145.36)	(102.90)	(100.03)	(139.94)	(276.12)
Lift	562.95***	579.63***	532.38***	575.58***	862.57***
	(170.38)	(119.65)	(114.56)	(156.44)	(302.79)
Heating System	125.97	184.53	– 1.57	– 25.19	– 567.16
	(195.78)	(135.17)	(30.98)	(181.81)	(345.35)
Parking area	1227.97*	1940.22***	2325.26***	1913.61***	1508.37
	(721.31)	(462.08)	(441.55)	(602.36)	(1103.27)
Bathroom	841.78***	924.98***	896.58***	742.37***	824.39**
	(178.24)	(126.16)	(123.13)	(174.75)	(355.76)
Housing Area	13793.34***	2229.88	– 18234.58***	– 38649.82***	– 59740.72***
	(3749.06)	(2575.02)	(2402.27)	(3718.76)	(7237.17)
Housing Area^2	26441.33**	88902.99***	201195.88***	312160.71***	450974.91***
	(12661.30)	(8845.19)	(8073.78)	(13545.17)	(25960.66)
Housing Area^3	59418.29***	21609.05***	– 70502.26***	– 146289.17***	– 281400.47***
	(11060.53)	(7778.60)	(6841.90)	(13132.12)	(24340.79)
Age	385.45**	467.61***	536.03***	662.14***	1622.88***
	(162.86)	(115.82)	(112.99)	(159.36)	(316.73)
Urban amenities					
Abandoned area	– 191.62	– 356.68***	– 506.70***	– 667.02***	– 1056.59***
	(184.95)	(132.17)	(127.00)	(174.07)	(325.83)
University	808.94***	896.77***	927.38***	763.42***	513.88*
	(146.14)	(107.38)	(106.50)	(154.85)	(297.03)
Metro	140.45	– 39.91	– 242.94**	– 402.89***	– 304.77
	(143.41)	(103.14)	(101.00)	(141.32)	(275.28)
Year	– 440.49***	– 428.92***	– 377.85***	– 356.10***	– 283.69
	(130.27)	(92.56)	(89.97)	(125.08)	(246.91)

† Point estimates with standard errors in parentheses and the associated p-value: $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

for penalised regression splines once a certain minimum value is reached (Pratesi et al. 2009; Ruppert et al. 2003).

All covariates act in an a priori predictable manner. Most of them are statistically significant and are valued differently at different points of the conditional distribution of house prices. This is also clearly displayed in Fig. 4, which shows the effect



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 4 Plot of regression coefficients with their confidence intervals of all the covariates included in the SSPMQ model: Intercept (**a**), Floor low (**b**), Floor high (**c**), Lift (**d**), Heating system (**e**), Parking Area (**f**), Bathroom (**g**), Building's age (**h**), Abandoned area (**i**), University (**j**), Metro (**k**) and Year (**l**). The plots are drawn using the estimates obtained for the five M-quantiles considered in table 3 and interpolate them linearly. The blue lines represent the regression coefficients obtained fitted a standard semiparametric linear model on the mean

of each variable at different M-quantiles. Grey bands represent the 95% confidence interval of the parameter of interest. We have also added a blue line to each graph to represent the regression coefficients obtained by fitting the standard semiparametric additive model on the mean specified in Equation (1). This allows us to assess whether and for which variables there is a heterogeneous effect on the distribution of response.

Some housing-specific attributes have a different impact at different levels of the price distribution, suggesting that their value is different at different points of the housing price distribution. The floor at which the house is located has been found to be statistically significant only for very low and medium-value houses. A non-constant effect of floor level along the price distribution has also been found by

Amédée-Manesme et al. (2017) in Paris: the higher the price category is, the lower the premium assigned to a given floor level. In contrast, our analysis suggests that a medium floor (second or third floor) is valued lower than a high floor (fourth floor or above) or a low floor (ground or first floor) at the lowest M-quantiles. The presence of an elevator has a positive and significant effect that remains quite stable across M-quantiles. In line with Michelangeli and Zanardi (2009), we find that the age of the building has a positive effect on housing prices. Moreover, this effect is more pronounced at the top of the housing price distribution. This can be explained by the fact that the oldest buildings tend to be located in the most elegant districts of the city centre; moreover, many of them are interesting from an architectural point of view and/or have beautiful gardens inside the court. Therefore, it is reasonable for the price differential to be higher for very high-valued houses. The presence of a heating system does not significantly affect the price distribution. The same results have also been found by Brambilla et al. (2013) using a standard model on the mean. Other housing-specific attributes, such as the presence of an elevator and more than one bathroom in the housing unit, have a positive and significant effect that remains quite stable across M-quantiles.

Some patterns of heterogeneity have also been found for the environmental characteristics. Neighbouring degradation (measured by the “abandoned area” variable) contributes negatively to the house price. In particular, it more importantly affects the price of average and high-valued houses. In contrast, it has been found not statistically significant for very low-valued houses ($q = 0.1$). This is surprising, “environmental quality is very much like leisure time: as people become wealthier, they demand more of it, mostly because they can better afford it” (Boudreaux 2008). the proximity to a university follows an inverted U-curve that is higher at the centre of the outcome distribution. This result appears reasonable: typically, luxury properties are less likely to be of interest to students, who largely represent the demand for housing close to universities. Quite surprisingly, we have found that proximity to a metro station negatively significantly affect the housing price at $q = 0.5$ and $q = 0.75$. This result reveals that the nuisance and congestion created by the station are not always compensated by the benefit arising from direct access to public transport. The global economic crisis in 2008 had a significant detrimental effect on the Milan housing market only for low- or average-valued houses. In contrast, it did not significantly affect the price of high-valued houses that were actually less impacted by the negative shock of the crisis. Finally, we note that for many of the variables considered in our model the impact at different M-quantiles differs from the impact they have on the mean (blue line in the graphs), which often also lies outside the 95% confidence interval. This fosters the idea that there exists a remarkable heterogeneity in housing demand and prices due to the structural and environmental characteristics of the property.

The estimated effect of housing size at the five M-quantiles $q = 0.1, 0.25, 0.5, 0.75$ is reported in Fig. 5. The curves in Fig. 5 have been constructed considering a reference housing unit located at the barycentre of the municipality with a value of the Cultural Catalyst to 0.6, sold in 2008 or later, with two or more bathrooms, a heating system that is not autonomous and a parking area or a garage. This reference unit

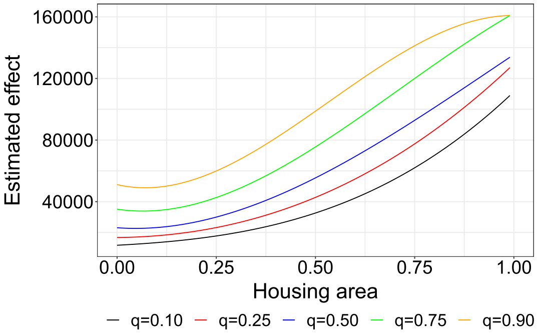


Fig. 5 Estimated housing area effect. Housing area has been standardised in the unit interval

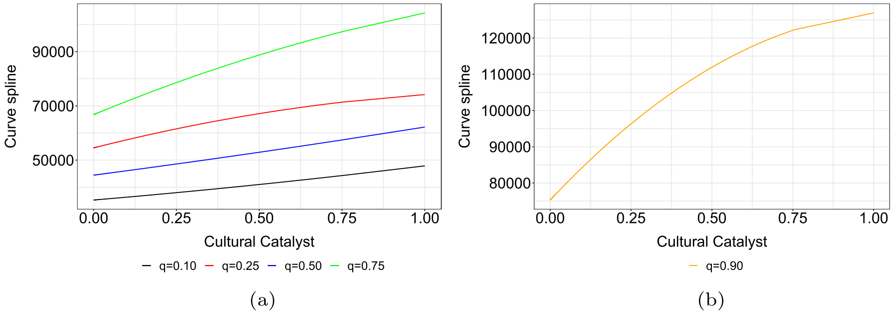


Fig. 6 **a** Spline effect (f_{1q}) of the Cultural Catalyst at M-quantile $q = 0.10$ (black), $q = 0.25$ (red), $q = 0.50$ (blue), $q = 0.75$ (green). **b** Spline effect (f_{1q}) at $q = 0.90$

has been assumed to be located on a medium floor (second or third floor), in a building built after 1950 with at least one elevator, far from an abandoned site, the metro station and a university site. The plot clearly shows that the impact of housing area on prices tends to be similar at different M-quantiles. However, the estimated curve for the higher M-quantile becomes concave for larger houses suggesting that buyers of more luxurious units attribute a progressively lower value to the house size.

The spline effect (f_{1q}) of the Cultural Catalyst at the five M-quantiles $q = 0.1, 0.25, 0.5, 0.75, 0.9$ is depicted in Fig. 6. The curves in panels (a) and (b) have been constructed considering the same reference housing unit used for Fig. 5 whilst setting the value of (standardised) housing area as large as 0.5. It is worth mentioning that the additive nature of the model implies that the specific housing unit considered as reference does not influence the spline shape with the exception of its intercept. Figure 6(a) shows a nearly linear or slightly concave shape for the considered M-quantiles. However, a clear concave shape is observed for $q = 0.9$. This seems to suggest that, similar to environmental quality (Boudreaux 2008), people are willing to pay more for culture when they become richer, probably because they can better afford it. However, this higher willingness to pay for culture increases at a decreasing rate.

Fig. 7 Spline effect of the Cultural Catalyst at M-quantiles $q = 0.25;0.5;0.75$ and variability bands

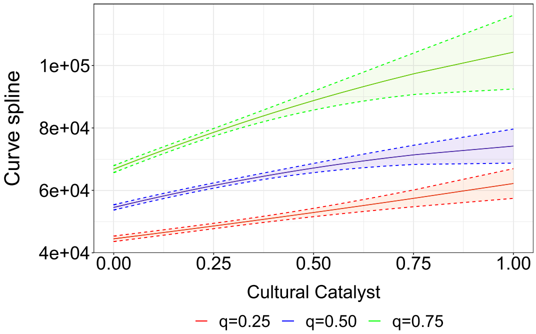
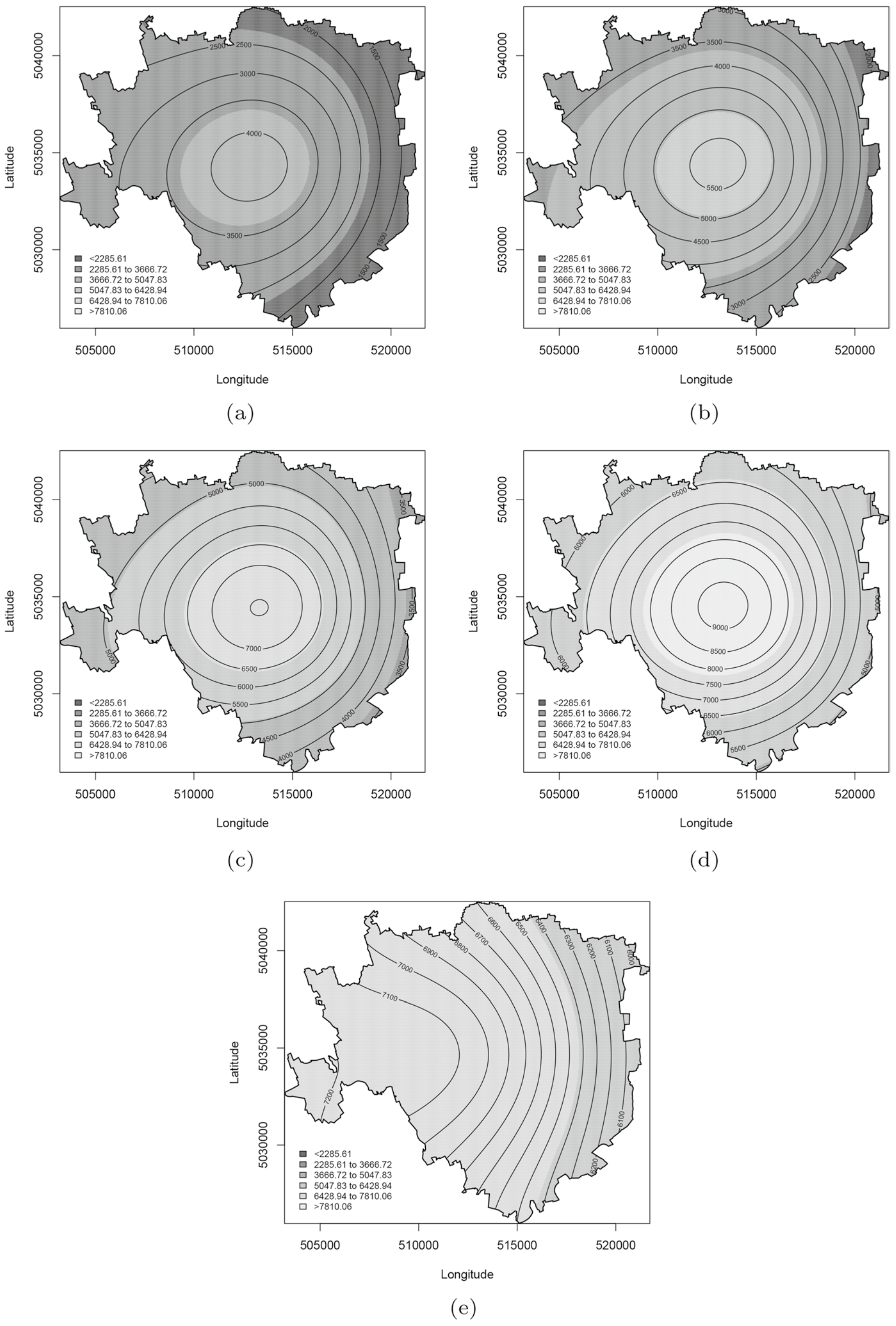


Figure 7 displays the spline effect of the Cultural Catalyst at $q = 0.25, 0.50, 075$ with the variability band of the curve, as defined in Sect. 4, represented by the coloured area. This band is largely above zero suggesting a significant positive effect of cultural amenities on housing prices at all the considered M-quantiles. Furthermore, envelops at different M-quantiles are clearly separated, suggesting that the effect of the Cultural Catalyst is significantly different at different levels of the price distribution.

Figure 8 shows the spatial dynamics of prices estimated by the bivariate thin plate component at the five considered M-quantiles. It is not surprising that higher prices tend to concentrate in the central area of the city and decrease when moving to the outskirts regardless of whether the house is a low, medium or high-valued unit.

To evaluate the goodness of fit of our model, we calculate the pseudo-R^2_ρ goodness-of-fit measure proposed by Bianchi et al. (2018): $R^2_\rho(q) = 1 - \frac{\sum_{i=1}^n \rho_q(e_{iq})}{\sum_{i=1}^n \rho_q(\tilde{e}_{iq})}$ for several M-quantiles. In the previous equation, e_{iq} are the scaled residuals under the full model, $\tilde{e}_{iq}$ are the scaled residuals under the null model (i.e. the model in which all the coefficients except for the intercept are set to zero), q is the M-quantile order, and ρ is the loss function defined in Equation (2). For all the considered M-quantile models, namely for $q = 0.1, 0.25, 0.5, 0.75$ and 0.9, we find that the M-quantile regression performs reasonably well (with $R^2_\rho(q)$ ranging from 40% to 75%) and that the R^2_ρ increases with the quantile order. The pseudo-$R^2_\rho(q)$ values associated with the M-quantiles models considered above are larger than the pseudo-$R^2_\rho(q)$ of the corresponding quantile regressions (the results of the latter models have not been reported in detail here), suggesting that the proposed approach is more appropriate than quantile regression for the data at hand.

Finally, it is worth showing how the implicit price associated with the Cultural Catalyst may be determined for any desired M-quantile. Let t be the value of the Cultural Catalyst and let $P_q(t)$ be the q-th M-quantile of the price distribution as a function of t, assuming that all of the other covariates in Equation (3) are fixed. To calculate the implicit price at any value t, it is necessary to estimate the derivative of $P_q(t)$ at t for every M-quantile of interest (see Sect. 2). By differentiating Equation (3) with respect to t, we obtain:



(a)

(b)

(c)

(d)

(e)

Fig. 8 Spatial spline effects at the five considered M-quantiles: **a** 0.10; **b** 0.25; **c** 0.50; **d** 0.75 and **e** 0.90

$$P'_q(t) = f'_{1q}(t)$$

where $f'_{1q}(t)$ is the derivative of the spline transformation at t.

From Equation (5), we find that an estimate of $f'_{1q}(t)$ can be obtained as follows:

$$\hat{f}'_{q1}(t) = \dot{\boldsymbol{x}}'_t \widehat{\boldsymbol{\beta}}_{1q} + \dot{\boldsymbol{z}}'_{sp} \widehat{\boldsymbol{\gamma}}_{1q} \tag{14}$$

where $\dot{\boldsymbol{x}}'_t = [0,\ 1,\ 2t]$, $\dot{\boldsymbol{z}}'_{sp} = \left[2\left(t - k_j\right)_+ \ : j = 1, \cdots, K_1\right]$ and, as above, k_j is the j-th knot of the spline and $\widehat{\boldsymbol{\beta}}_{1q}$ and $\widehat{\boldsymbol{\gamma}}_{1q}$ are the estimates of the regression coefficients associated with the spline basis.

Using Equation (14), the implicit price has been calculated for M-quantiles 0.10, 0.25, 0.5, 0.75 and 0.90 at the median sample value of the Cultural Catalyst for the reference housing unit described before. Assuming that the amount of t changes by a small quantity dt, say 0.01, the increase in the price per square metre is approximately €38.8 for $q = 0.10$, €49.1 for $q = 0.25$, €63.6 for $q = 0.50$, €76.3 for $q = 0.75$ and €86.2 for $q = 0.90$, suggesting that households with low-priced properties behave differently than households with high-priced housing in terms of the marginal willingness to pay for culture. The latter households attribute a greater value to a marginal increase in cultural amenities.

6 Conclusions

The aim of this paper was to employ an M-quantile approach to examine how the effect of housing characteristics may vary across the conditional distribution of house prices, preserving the robustness and efficiency of the estimators of the regression parameters. More specifically, a semiparametric M-quantile regression was proposed for the residential housing market of Milan. We included a spline component in the model to estimate the potential nonlinear effect of the Cultural Catalyst, an index for cultural amenities. We also considered other urban amenities, such as the presence of abandoned sites, metro stations and university sites. Our findings suggest that several housing attributes differ significantly across the response distribution, supporting the choice of estimating the conditional M-quantile functions in addition to the conditional mean (Liao and Wang 2012). High-income residents are more concerned about the environmental quality and are willing to pay a higher price for an improvement of the context where the unit is located. Similarly, cultural amenities have a stronger positive effect on high-valued houses. At the top of the distribution of prices, the impact of cultural amenities increases more than linearly as its quantity increases. These results suggest that people tend to demand more cultural amenities and environmental quality as they become wealthier, mostly because they can better afford them. The proximity to university sites has been found to significantly increase the price of housing for low and average value houses where the effect on price distribution is fairly stable, whereas the impact is negligible on high-value units. The latter units, in fact, are presumably less interesting for students, who largely represent the demand for housing near a university site. Some specific attributes of the house, such as its size and the presence of a lift or of a parking area are much more valuable for high-value units although their impact has also been found to be statistically significant at a lower level of the price distribution.

The nonparametric nature of the proposed approach permits one to avoid the ubiquitous log transformation of the prices in regression analysis. The log transformation has important drawbacks when, as in the present paper, one is interested in predicting the value of implicit prices of amenities and housing-specific attributes, since it is necessary to back transform the log-prices on the raw scale. The simple exponentiation of the predicted log price provides naive estimates of the implicit prices biased downwards. The first way to adjust the bias is to assume a log-normal distribution of the residuals. This assumption is often difficult to test and indeed it is rarely tested in this strand of literature. A second method is to use transformation bias correction discussed by Chambers and Clark (2012). On the one hand, this correction does not require any particular distribution model. On the other hand, it requires calibrating the naive estimates of the implicit prices by a data-based factor that reduces but does not eliminate the bias of the final estimates. We also note that log transformation is often used to mitigate the influence of extreme raw scale values. In our case, this does not occur since the percentage of outliers remains basically the same on the log scale. Moreover, log transformed data are susceptible to 'small' outliers (i.e. values close to zero). This again may induce an increased variability of parameter estimates on the log scale, and hence, it further justifies the M-quantile approach that downweights outliers. More generally the M-quantile approach has permitted housing prices to be modelled in a natural manner, avoiding strong assumptions and preserving the statistical efficiency of the estimators of the regression coefficients. Perhaps, this last point represents the main advantage of this approach. There is a sort of balance between robustness and efficiency of estimators through the tuning constant of the influence function (see Sect. 4). Moreover, the option to select several continuous influence functions in the M-quantile regression—in contrast to the absolute value function in the quantile regression—offers the opportunity to obtain additional computational stability.

Finally, the methodology employed in this paper has proven to be extremely flexible. We showed how it can be straightforwardly coupled with semiparametric specifications that allow one to take into account effects that are potentially nonlinear or that follow spatial dynamics. We also believe that this methodology has a potential wide range of applications in the residential housing market, from identifying housing submarkets to designing tax systems or financing local public goods, such as culture and better environment conditions.

Funding Open access funding provided by Università di Pisa within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Alfò, M., Salvati, N., Ranallli, M.G.: Finite mixtures of quantile and M-quantile regression models. Stat. Comput. **27**(2), 547–570 (2017)

Amédée-Manesme, C.O., Baroni, M., Barthélémy, F., et al.: Market heterogeneity and the determinants of Paris apartment prices: a quantile regression approach. Urban Stud. **54**(14), 3260–3280 (2017)

Bayer, P., McMillan, R., Rueben, K.: An equilibrium model of sorting in an urban housing market. Technical report, National Bureau of Economic Research (2004)

Bianchi, A., Fabrizi, E., Salvati, N., et al.: Estimation and testing in M-quantile regression with applications to small area estimation. Int. Stat. Rev. **86**(3), 541–570 (2018)

Borgoni, R., Del Bianco, P., Salvati, N., et al.: Modelling the distribution of health-related quality of life of advanced melanoma patients in a longitudinal multi-centre clinical trial using M-quantile random effects regression. Stat. Methods Med. Res. **27**(2), 549–563 (2018)

Borgoni, R., Michelangeli, A., Pontarollo, N.: The value of culture to urban housing markets. Reg. Stud. **52**(12), 1672–1683 (2018)

Borgoni, R., Degli Antoni, G., Faillo, M., et al.: Natives, immigrants and social cohesion: intra-city analysis combining the hedonic approach and a framed field experiment. Int. Rev. Appl. Econ. **33**(5), 697–711 (2019)

Boudreaux, D.: Globalization. Greenwood Press, Westport (2008)

Brambilla, M., Michelangeli, A., Peluso, E.: Equity in the city: on measuring urban (ine) quality of life. Urban Stud. **50**(16), 3205–3224 (2013)

Breckling, J., Chambers, R.: M-quantiles. Biometrika **75**(4), 761–771 (1988)

Brunauer, W., Lang, S., Umlauf, N.: Modelling house prices using multilevel structured additive regression. Stat. Modell. **13**(2), 95–123 (2013)

Chambers, R., Clark, R.: An Introduction to Model-Based Survey Sampling with Applications. Oxford University Press, Oxford (2012)

Chambers, R., Tzavidis, N.: M-quantile models for small area estimation. Biometrika **93**(2), 255–268 (2006)

Chambers, R., Salvati, N., Tzavidis, N.: Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. J. R. Stat. Soc. Ser. A **179**(2), 453–479 (2016)

Chasco, C., Le Gallo, J.: Heterogeneity in perceptions of noise and air pollution: a spatial quantile approach on the city of Madrid. Spat. Econ. Anal. **10**(3), 317–343 (2015)

Chasco, C., Sánchez, B.: Valuation of environmental pollution in the city of madrid: an application with hedonic models and spatial quantile regression. Rev. d'Econo. Reg. Urbaine **1**, 343–370 (2015)

Diao, M., McMillen, D.P., Sing, T.F.: A quantile regression analysis of housing price distributions near MRT stations. Tech. rep., Annual Conference Real Estate and Urban Economics (2018)

Dreassi, E., Ranalli, M.G., Salvati, N.: Semiparametric M-quantile regression for count data. Stat. Methods Med. Res. **23**(6), 591–610 (2014)

Freeman, M.: The Measurement of Environmental and Resource Values: Theory and Method. Resources for the Future, Washington (1993)

Fritsch, M., Haupt, H., Ng, P.T.: Urban house price surfaces near a world heritage site: modeling conditional price and spatial heterogeneity. Reg. Sci. Urban Econ. **60**, 260–275 (2016)

Garretsen, H., Marlet, G.: Amenities and the attraction of Dutch cities. Reg. Stud. **51**(5), 724–736 (2017)

Gravel, N., Michelangeli, A., Trannoy, A.: Measuring the social value of local public goods: an empirical analysis within Paris metropolitan area. Appl. Econ. **38**(16), 1945–1961 (2006)

Huang, P.: Impact of distance to school on housing price: evidence from a quantile regression. Empir. Econ. Lett. **17**(2), 149–156 (2018)

Huber, P.J.: Robust statistics. Springer, Berlin (2011)

Huggins, R.: On the robust analysis of variance components models for pedigree data. Aust. J. Stat. **35**(1), 43–57 (1993)

Huggins, R., Loesch, D.: On the analysis of mixed longitudinal growth data. Biometrics **54**(2), 583–595 (1998)

Koenker, R., Bassett, G., Jr.: Regression quantiles. Econometrica **46**(1), 33–50 (1978)

Kostov, P.: A spatial quantile regression hedonic model of agricultural land prices. Spat. Econ. Anal. **4**(1), 53–72 (2009)

Leung, T.C., Tsang, K.P.: Love thy neighbor: income distribution and housing preferences. J. Hous. Econ. **21**(4), 322–335 (2012)

Liao, W.C., Wang, X.: Hedonic house prices and spatial quantile regression. J. Hous. Econ. **21**(1), 16–27 (2012)

Mak, S., Choy, L., Ho, W.: Quantile regression estimates of Hong Kong real estate prices. Urban Stud. **47**(11), 2461–2472 (2010)

Malpezzi, S.: Hedonic pricing models: a selective and applied review. In: O'Sullivan, T., Kenneth, G. (eds.) Housing Economics and Public Policy, pp. 67–89. John Wiley & Sons, Hoboken (2002)

McMillen, D.P.: Quantile Regression for Spatial Data. Springer Science & Business Media, Berlin (2012)

McMillen, D.: Conditionally parametric quantile regression for spatial data: an analysis of land values in early nineteenth century Chicago. Reg. Sci. Urban Econ. **55**, 28–38 (2015)

Michelangeli, A., Zanardi, A.: Hedonic-based price indexes for the housing market in Italian cities: theory and estimation. Polit. Econ. **25**(2), 109–146 (2009)

Newey, W.K., Powell, J.L.: Asymmetric least squares estimation and testing. J. Econom. Soc. **55**(4), 819–847 (1987)

Opsomer, J., Claeskens, G., Ranalli, M., et al.: Nonparametric small area estimation using penalized spline regression. J. R. Stat. Soc. Ser. B **70**(1), 265–283 (2008)

Pratesi, M., Ranalli, M.G., Salvati, N.: Nonparametric M-quantile regression using penalised splines. J. Nonparametric Stat. **21**(3), 287–304 (2009)

R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

Rosen, S.: Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. **82**(1), 34–55 (1974)

Ruppert, D., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, Cambridge (2003)

Schirripa Spagnolo, F., Salvati, N., D'Agostino, A., et al.: The use of sampling weights in M-quantile random-effects regression: an application to programme for international student assessment mathematics scores. J. R. Stat. Soc. Ser. C (Appl. Stat.) **69**(4), 991–1012 (2020)

Tomal, M., Helbich, M.: A spatial autoregressive geographically weighted quantile regression to explore housing rent determinants in Amsterdam and Warsaw. Urban Anal. City Sci. Environ. Plan. B (2022)

Trzpiot, G.: Spatial quantile regression. Comp. Econ. Res. Central East. Eur. **15**(4), 265–279 (2012)

Tzavidis, N., Salvati, N., Schmid, T., et al.: Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Cohort Study children in England using M-quantile random-effects regression. J. R. Stat. Soc. Ser. A **179**(2), 427–452 (2016)

Uematsu, H., Khanal, A.R., Mishra, A.K.: The impact of natural amenity on farmland values: a quantile regression approach. Land Use Policy **33**, 151–160 (2013)

Waltl, S.R.: Variation across price segments and locations: a comprehensive quantile regression analysis of the Sydney housing market. Real Estate Econ. **47**(3), 723–756 (2019)

Wan, A.T., Xie, S., Zhou, Y.: A varying coefficient approach to estimating hedonic housing price functions and their quantiles. J. Appl. Stat. **44**(11), 1979–1999 (2017)

Wood, S.N.: Generalized additive models: an introduction with R. Chapman and Hall/CRC, Routledge (2017)

Zietz, J., Zietz, E.N., Sirmans, G.S.: Determinants of house prices: a quantile regression approach. J. Real Estate Finance Econ. **37**(4), 317–333 (2008)