



ORIGINAL ARTICLE

Open Access



Examining the vintage effect in hedonic pricing using spatially varying coefficients models: a case study of single-family houses in the Canton of Zurich

Jakob A. Dambon^{1,2*} , Stefan S. Fahrländer³, Saira Karlen³, Manuel Lehner³, Jaron Schlesinger³, Fabio Sigrist²  and Anna Zimmermann³

Abstract

This article examines the spatially varying effect of age on single-family house (SFH) prices. Age has been shown to be a key driver for house depreciation and is usually associated with a negative price effect. In practice, however, there exist deviations from this behavior which are referred to as vintage effects. We estimate a spatially varying coefficients (SVC) model to investigate the spatial structures of vintage effects on SFH pricing. For SFHs in the Canton of Zurich, Switzerland, we find substantial spatial variation in the age effect. In particular, we find a local, strong vintage effect primarily in urban areas compared to pure depreciative age effects in rural locations. Using cross validation, we assess the potential improvement in predictive performance by incorporating spatially varying vintage effects in hedonic models. We find a substantial improvement in out-of-sample predictive performance of SVC models over classical spatial hedonic models.

Keywords: Gaussian process, Spatial statistics, Real estate, Mass appraisal

JEL Classification: C31, C53, R31, R32

1 Introduction

Hedonic real estate models contain several predictor variables, and age is a key explanatory variable. The marginal effect of the building age on house prices has been well-studied. It has been found that the age effect is non-linear (Clapp & Giaccotto, 1998; Goodman & Thibodeau, 1995). In particular, Case et al. (2004) report a “plausible quadratic form” for the building age. This behavior is a result of two main features of the age as an independent variable: (1) In general, older buildings depreciate due to deterioration; (2) “however, beyond some point, only those houses with the best locations and the highest

construction quality survive.” (Case et al., 2004, p. 171). The quadratic appearance of the age effect has also been observed by Fahrländer (2006) and linked to the building material and architectural style. Studies investigating this particular type of behavior, i.e., a deviation from a pure depreciative effect once a particular age has been reached, are referencing to it as a *vintage effect* (Clapp & Giaccotto, 1998; Goodman & Thibodeau, 1995; Rubin, 1993).

Over the last two decades, there emerged a special focus on location specific effects due to newly available modeling methodologies. There are numerous publications which show a clear indication of spatially varying covariate effects within hedonic pricing models. For instance, when applying additive mixed regression models on rents in Vienna (Austria), Brunauer et al.

*Correspondence: jakob.dambon@gmail.com

¹ Department of Mathematics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Full list of author information is available at the end of the article

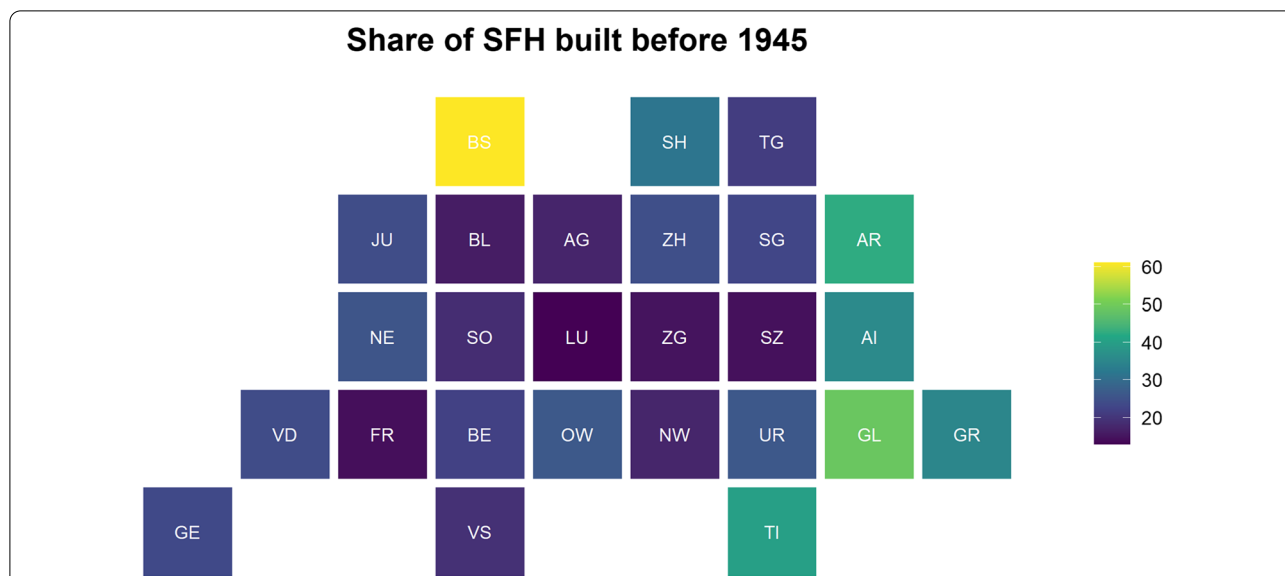


Fig. 1 Share of SFHs built before 1945 by canton, as of 2019. The respective shares are given in percent. *Data:* Federal Statistical Office FSO (2021)

(2010) find “substantial spatial variation” of covariate effects between the districts of Vienna.

Existing methods to model such spatially varying coefficients (SVC) are Bayesian processes (Gelfand et al., 2003) and geographically weighted regression (Fotheringham et al., 2002). Applications of these methods consistently show the existence of non-stationary coefficients, e.g., Baton Rouge (LA, United States, see Gelfand et al., 2003), in Toronto (ON, Canada, see Wheeler et al., 2014), Singapore (Cao et al., 2019; van Eggermond et al., 2011), and Shenzhen (China, see Geng et al., 2011).

The goal of this paper is to unify both frameworks, i.e., vintage effects and SVC modeling, to investigate a possible spatially varying vintage effect. In particular, we want to examine if such non-stationary vintage effect exists and, in a second step, see if we can improve the quality of hedonic models in price prediction. Our study is motivated by previous work on spatially varying relationships between house prices and age. For instance, one of the first observations of spatial differences in the age effects can be found in Malpezzi et al. (1987). They compared individual hedonic models for 59 metropolitan areas in the United States and concluded that “[s]everal metropolitan areas exhibited significant deviations from the average depreciation patterns.” (Malpezzi et al., 1987, p. 382). More recent evidence for such behavior is presented in Brunauer et al. (2010) as well as Dambon et al. (2021a) who found pronounced spatially varying effects on the rents and the prices of apartments, respectively.

In this paper, we model spatially varying vintage effects for single-family houses (SFHs) in the Canton of Zurich (ZH, Switzerland). We select the Canton of Zurich as

our area of analysis for several reasons. Firstly, the Canton of Zurich is sufficiently large and contains urban as well as rural areas. This is relevant in that our working hypothesis is that, on average, age has a negative effect on SFH price, but that spatial deviations in the form of vintage effects might occur in metropolitan and urban areas. One hypothesis is that such spatial deviations are driven by unobserved attributes such as architectural style and build quality of the SFH. New research also suggests that redevelopment options might also have an impact (Clapp & Salavei, 2010; Munneke & Womack, 2016). Hence, the Canton of Zurich with its above average rate of SFHs in urban areas is of particular interest. Further reasons for choosing the Canton of Zurich as the area of analysis are that the age structure of the Canton of Zurich is very similar to that of Switzerland as a whole and that some information on the full census of SFH transactions in the Canton of Zurich in the chosen study period is available. The latter is valuable in that it allows to check the representativity of our data.

Our analysis is economically relevant as a sizeable portion of the SFHs in the Canton of Zurich and in Switzerland in general are old. More specifically, a quarter of all SFHs in Switzerland were built before 1945 and a third were built before 1960. Very similar proportions are found for the Canton of Zurich (see Fig. 1 below and Table 6 in the “Appendix”). Given the existence of a vintage effect, accounting for such effects could therefore yield more accurate predictions for a sizeable portion of SFH transactions.

To verify our hypothesis on spatially varying vintage effects, we will use a new methodology introduced by

Dambon et al. (2021a) to model spatially varying coefficients using Gaussian processes (GP). In the next section, we first introduce and then extend the definition on SVC models and GP-based SVC models. In Sect. 3, we present the real estate data and justify the model. The model results are presented in Sect. 4. In Sect. 5 we assess predictive performance of the SVC model and compare it to a standard hedonic model. We conclude with a discussion of our results in Sect. 6.

2 Spatially varying coefficient models

Spatially varying coefficient models are a generalization of classical linear regression models, where we allow the regression coefficients to vary over space. That is, the effect of a covariate $x^{(j)}$ denoted by the coefficient β_j can depend on a geographic location s , which we assume to be two-dimensional. SVC models can be applied to spatial points data sets, where for each of the n observations of the response variable $y := (y_1, \dots, y_n)^T \in \mathbb{R}^n$ and p covariates $x^{(j)} := (x_1^{(j)}, \dots, x_n^{(j)})^T \in \mathbb{R}^n, j = 1, \dots, p$, every observation has an associated location s_i . In summary, SVC models are defined as

$$y_i = \beta_1(s_i)x_i^{(1)} + \dots + \beta_p(s_i)x_i^{(p)} + \epsilon_i, \tag{1}$$

where $i = 1, \dots, n$ indexes the observations with their corresponding locations s_i and ϵ_i is a classical $N(0, \tau^2)$ iid error term with $\tau^2 > 0$.

If one assumes that not all coefficients should contain spatial structures, one can define mixed SVC models. Let q with $1 \leq q \leq p$ be the number of covariates for which we want to model SVCs. Without loss of generality, we define the mixed SVC model as

$$y_i = \beta_1(s_i)x_i^{(1)} + \dots + \beta_q(s_i)x_i^{(q)} + \beta_{q+1}x_i^{(q+1)} + \dots + \beta_px_i^{(p)} + \epsilon_i. \tag{2}$$

From now on, we assume that the first coefficient $j = 1$ always models an intercept. In the special case when $q = 1$, we have the *classical geostatistical model* that is also used in most hedonic models. The exact assumptions for the coefficients $\beta_j(\cdot), j = 1, \dots, q$, and how they are estimated, have yet to be defined. The literature on how to do so for both the classical geostatistical and SVC models is extensive. For geostatistical models, see Cressie (2011) and Heaton et al. (2019) for an overview. For SVC models, see Dambon et al. (2021a), Wheeler and Calder (2007), and Wheeler and Waller (2009) for comparisons.

2.1 Gaussian process-based SVC models

We specify the SVC model such that each coefficient is defined by a Gaussian process (Rasmussen & Williams, 2006). Gaussian processes are well-studied and

Table 1 Parametrizations of two Gaussian processes

	Mean μ	Range ρ	Variance σ^2
Parametrization 1	2	0.25	1
Parametrization 2	-1	0.10	2

widely used tools to model dependency structures with applications including—but not limited to—spatial statistics (Banerjee et al., 2008; Datta et al., 2016; Gelfand & Schliep, 2016), econometrics (Wu et al., 2014), and time series modeling (Roberts et al., 2013). They are infinite dimensional stochastic processes that are defined similarly to a finite-dimensional normal distribution. We assume the GP to be jointly independent as well as independent of the error term $\epsilon := (\epsilon_1, \dots, \epsilon_n)^T \sim N_n(\mathbf{0}_n, \tau^2 \mathbf{I}_n)$. For n observations $s := (s_1, \dots, s_n)^T$, they are given by

$$\beta_j(s) \sim N_n(\mu_j \cdot \mathbf{1}_n, \Sigma^{(j)}), \tag{3}$$

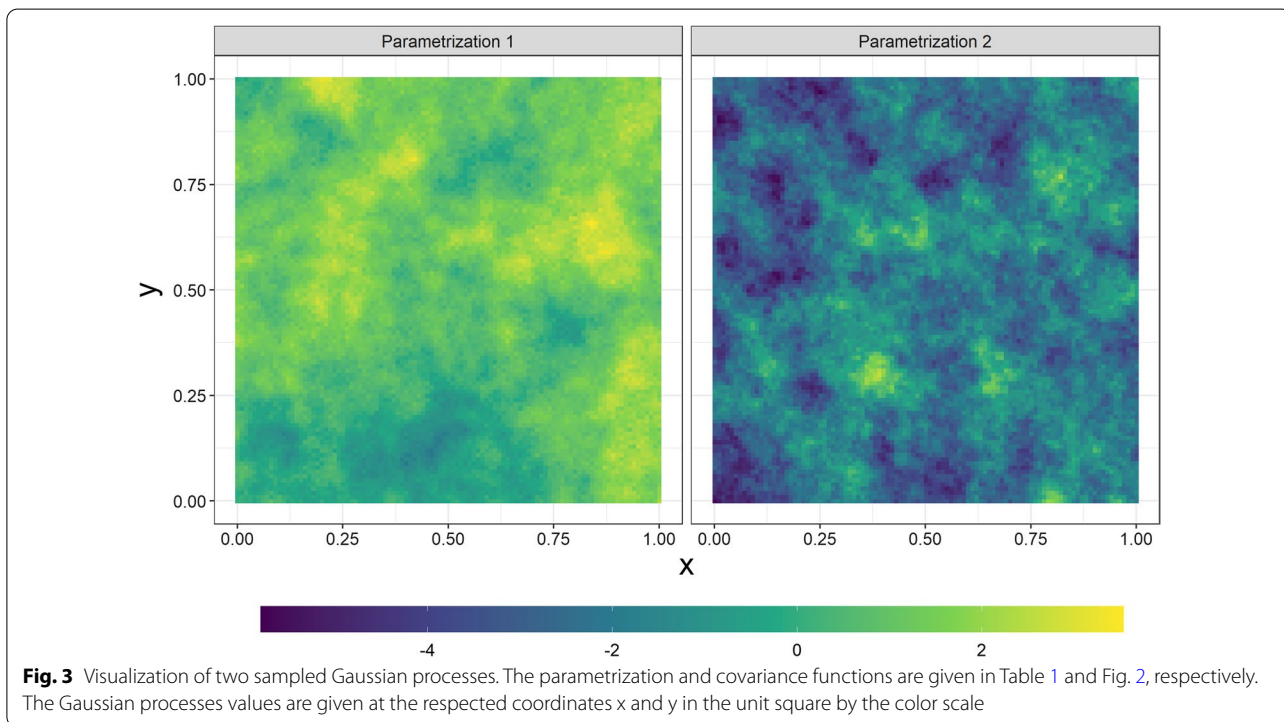
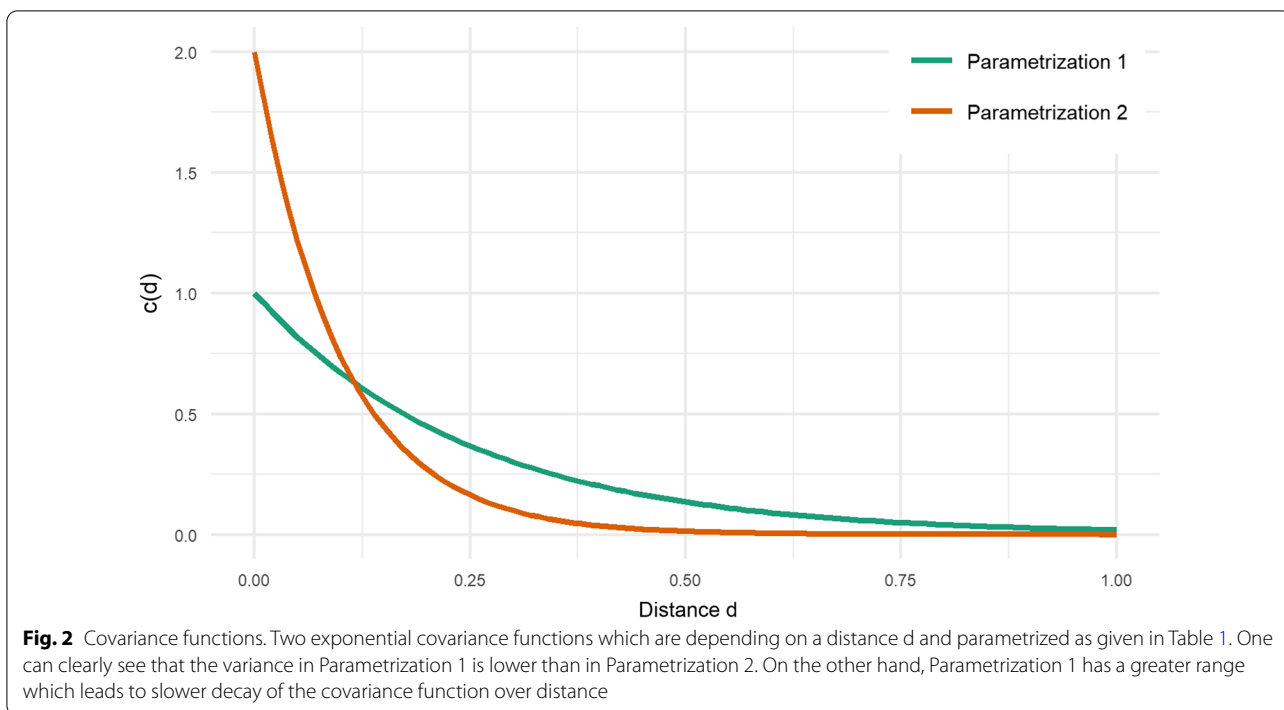
for $j = 1, \dots, q$. We assume a constant mean μ_j and a covariance matrix $\Sigma^{(j)}$, which is defined by a covariance function $c^{(j)}$ and the corresponding observation locations s . The observation locations are being used to model the dependency between observations by computing the distances. In spatial statistics, one usually assumes that closer observations share higher dependency than observations which are far apart.¹ We use the Euclidean distance denoted by $\|\cdot\|$ which yields pair-wise distances $d_{kl} := \|s_k - s_l\|$ between all observations, $1 \leq k, l \leq n$. Here, we assume to have exponential covariance functions $c^{(j)}(d) = \sigma_j^2 \cdot \exp(-d/\rho_j), d \geq 0$, parametrized by variances $\sigma_j^2 \geq 0$ and ranges $\rho_j > 0$. The former parameter defines the extent of variation within an SVC $\beta_j(s)$ and the latter defines the decay of spatial dependency with distance. The covariance function is then applied to the distances, which yields the following corresponding covariance matrix

$$\left(\Sigma^{(j)}\right)_{kl} := c^{(j)}(d_{kl}) = \sigma_j^2 \cdot \exp(-\|s_k - s_l\|/\rho_j).$$

2.1.1 Example of two sampled Gaussian processes

In this section, we illustrate the interpretation of the parameters for a GP with the help of two samples. Both are defined by their corresponding parameters given in Table 1. Under the assumption of an exponential

¹ This statement is also referred to as the *first law of geography* according to Waldo R. Tobler: “Everything is related to everything else, but near things are more related than distant things.” (Tobler, 1970).



covariance function, these parameters, more specifically, the ranges and variances, define the covariance functions given in Fig. 2. With the given covariance functions as well as the mean parameters, we sample the GPs on a

regular 101×101 from the unit square. The sampled GPs are given in Fig. 3.

The influence of each of the corresponding 3 parameters, i.e., the mean μ , the range ρ , and the variance σ^2 ,

can be directly seen from the individual visualized samples in Fig. 3. First, we note that the values of each parametrization are scattered around their individual means. The greater range of parametrization 1 relative to parametrization 2 expresses itself by larger color patches in Fig. 3. The greater variance of parametrization 2 leads to a wider range of values in the simulation which manifests itself by a wider color range in the visualization.

2.2 Maximum likelihood estimation of GP-based SVC models

We give a brief summary of a maximum likelihood estimation (MLE) approach for SVC models as introduced in Dambon et al. (2021a). Additionally, we extend the framework such that not only full GP-based SVC models as given in (1), but also mixed GP-based SVC models as given in (2) can be estimated.

With a data matrix X , where the entry $(X)_{ij} := x_i^{(j)}$ is the i th observation of the j th covariate, a mean vector $\mu := (\mu_1, \dots, \mu_p)^T \in \mathbb{R}^p$, the element-wise matrix product, and using the independence assumptions from above, the distribution of the response is given by

$$Y \sim N_n \left(X\mu, \sum_{j=1}^q \Sigma^{(j)} \odot \mathbf{x}^{(j)} (\mathbf{x}^{(j)})^T + \tau^2 I_n \right). \quad (4)$$

The differences between the response’s distribution as above and as given in Dambon et al. (2021a) are twofold. The first q entries of the mean vector μ are the means of the GP as defined in (3), while the further entries are the coefficients $\beta_{q+1}, \dots, \beta_p$. For simplicity, we identify them with μ_{q+1}, \dots, μ_p , respectively. The second difference is the sum building the covariance matrix. Since only covariates $j = 1, \dots, q$ are defined to have SVCs, only q covariance matrices and the respective covariates enter the sum.

The model is thus fully parametrized by the covariance parameters $\theta := (\rho_1, \sigma_1^2, \dots, \rho_q, \sigma_q^2, \tau^2)^T \in (\mathbb{R}_{>0} \times \mathbb{R}_{\geq 0})^q \times \mathbb{R}_{\geq 0}$ and the mean parameters $\mu \in \mathbb{R}^p$. We define $\omega := (\theta^T, \mu^T)^T$ as our parameter of interest which we estimate by maximizing the log-likelihood of (4). Since there exists no analytical solution, we must turn to numeric optimization. Once the estimate $\hat{\omega}$ is found, one can use it to predict the SVCs for (new) locations s' using the conditional distribution, i.e., one obtains $\hat{\beta}_j(s')$, $j = 1, \dots, p$. The estimator and predictor are implemented in the statistical software R (R Core Team, 2020)

and can be used via the package *varycoef* (Dambon et al., 2021b).

3 Data and model

3.1 Data

The analysis is based on transaction data for SFHs in the Canton of Zurich. The data is provided by Fahrländer Partner Raumentwicklung (FPRE), Zurich (Switzerland) and was collected by Swiss banks and insurance companies in their day-to-day business. It covers a time span of 6 consecutive quarters ranging from the 3rd quarter of 2018 to 4th quarter of 2019 and consists of 1578 observations.² Comparing the total number of transactions between the full census (approximately 3392 observations) and our data set at hand, we cover approximately 47% of the transactions of SFHs in the Canton of Zurich for the given period (Statistisches Amt des Kantons Zürich, 2021a, 2021b). The median transaction price of a SFH in our dataset is 1,390,000, which is comparable to the median transaction price in the full survey, which was 1,200,000 in 2018 and 1,250,000 in 2019 (Statistisches Amt des Kantons Zürich, 2021a). An overview of the data alongside some summary statistics is given in Table 2(a) and (b).

Due to Swiss banking secrecy, the exact geographic locations of the SFH cannot be disclosed. Here, FPRE works with a fine grid of cells that divides the Canton of Zurich into a total of 563 cells. The true SFH locations in our data are given by representative centroids of the cells, c.f. Table 2(c) and Fig. 4. The centroid’s location is provided in the LV03 coordinate reference system (Federal Office of Topography swisstopo, 1900). The cell’s resolution is higher in densely populated areas and the cells were defined by real estate experts to account for differences on sub-ZIP-code level. The high resolution of the cells allows us to differentiate between districts of municipalities, for instance the proximity of a cell to a lake or city center. The median cell size is 3.576 km², with the total range of areas extending from 0.246 to 18.809 km². In total, we observe data at 268 distinct cells. Additionally, each cell is labeled with a location type, see Table 2(c), which will turn out helpful when analyzing our findings in Sect. 4.

3.2 Model

The model has the natural logarithm of the transaction price as the response variable. Further, we standardize the age using the following transformation,

² A total of two observations were removed from the data set, for which real estate experts from Fahrländer Partner Raumentwicklung (FPRE) assume that they were incorrectly classified as arm’s length transactions.

Table 2 Description and summary statistics of underlying data set

<i>(a) Continuous variables</i>					
Variable	Description	Min	Median	Max	SD
Price	Adjusted transaction price in Swiss Francs excluding parking and special factors	400,000	1,390,000	10,500,000	938,813
Age	Age at the time of purchase	− 1 ^a	36	99 ^b	27
Volume	Building volume in m ³ (SIA Zürich, 2003)	300	780	3134	297
Plot size	Plot size in m ²	103	471	3101	343
Renov	Need for renovation (difference between actual and theoretical building condition, higher meaning better; h.m.b.)	0.00	0.00	4.00	0.91
Standard	Standard; h.m.b	2.00	3.00	5.00	0.67
Micro	Micro-location; h.m.b	2.00	3.50	5.00	0.64

<i>(b) Categorical variables, reference level in italics</i>			
Variable	Description	Levels	Observations
Year quarter	Transaction year and quarter	<i>20,183: 3rd Quarter of 2018</i>	365
		20,184: 4th Quarter of 2018	168
		20,191: 1st Quarter of 2019	265
		20,192: 2nd Quarter of 2019	275
		20,193: 3rd Quarter of 2019	303
		20,194: 4th Quarter of 2019	202
SFH type	Type of SFH	<i>1: Detached</i>	736
		2: Semi-detached	532
		3: Row house	310
Energy	Energy standard	<i>1: Insulated shell</i>	1538
		2: Enhanced energy efficiency	40

<i>(c) Observation locations</i>		
Coordinates	Description	Range
Easting LV03x	Coordinates in the LV03 coordinate reference system (Federal Office of Topography swisstopo, 1900) in m	200 × 10 ³ – 800 × 10 ³
Northing LV03y		100 × 10 ³ – 400 × 10 ³

Variable	Description	Levels	Observations
FPRE type	Type of cell	1: Top-locations	446
		2: Urban agglomerations	728
		3: Other agglomerations	265
		4: Rural areas	139

^a Negative age values are given if the date of transaction is prior to the year of construction

^b The value 99 is given for SFHs with an age of 99 years or older

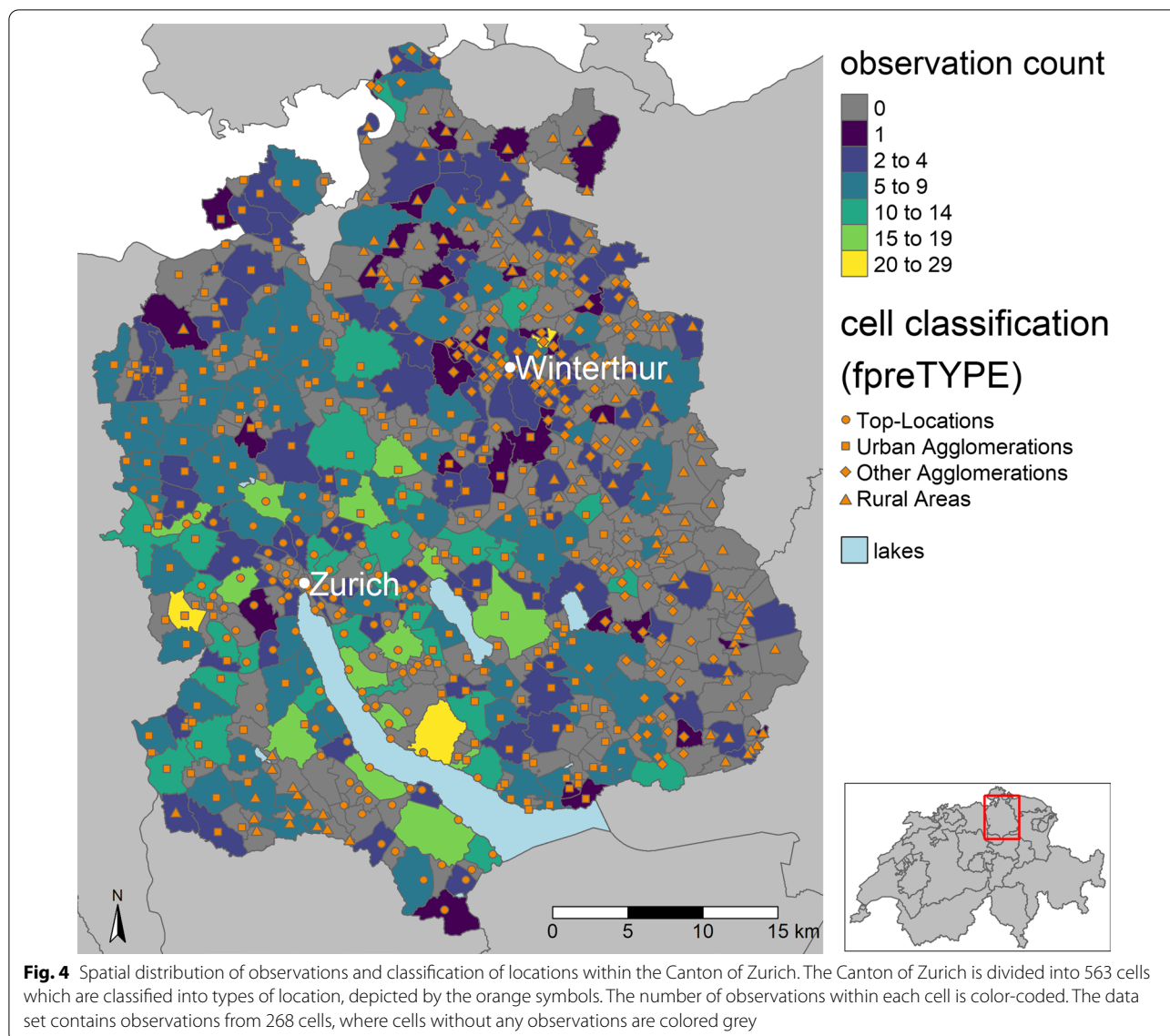
$$Z.age = \frac{age}{s_{age}}$$

where s_{age} is the empirical standard deviation of age of all observations.

The advantage of working with $Z.age$ rather than the actual age is a numerical stable optimization process of the maximum likelihood estimation. As mentioned above, we expect a quadratic effect (Case et al., 2004; Clapp & Giaccotto, 1998; Fahrländer, 2006; Goodman & Thibodeau, 1995), which is why we also include the covariate $Z.age^2$. As we expect spatial variation in these

coefficients, we use SVCs for these variables, c.f. first line in (5). The $plotsize$ as well as the $volume$ enter the model under a natural logarithm transformation. The rest of the continuous covariates $renov$, $standard$ and $micro$ are included without further transformation. Thus, all continuous covariates have approximately the same standard deviations which results in a well-behaved numeric optimization procedure for estimating the model.

The categorical variables $yearquarter$, $energy$ and $SFHtype$ and the error term complete our model which can be formulated as:



$$\begin{aligned}
 y_i = \log price_i &= \beta_1(s_i) + \beta_2(s_i) \cdot Z.age_i + \beta_3(s_i) \cdot Z.age_i^2 \\
 &+ \beta_4 \cdot \log volume_i + \beta_5 \cdot \log plotsize_i + \beta_6 \cdot renov_i \\
 &+ \beta_7 \cdot standard_i + \beta_8 \cdot micro_i \\
 &+ \beta_9 \cdot yearquarter_i + \beta_{10} \cdot SFHtype_i + \beta_{11} \cdot energy_i + \epsilon_i.
 \end{aligned}
 \tag{5}$$

Comparing the general mixed SVC model (2) and our explicit hedonic model (5) we note that we have $q = 3$ and $p = 16$ including the intercept and all factor levels deviating from the reference levels. The model is therefore fully parametrized by

$$\begin{aligned}
 \omega &= (\theta^T, \mu^T)^T \\
 &= (\rho_1, \sigma_1^2, \rho_2, \sigma_2^2, \rho_3, \sigma_3^2, \tau^2, \mu_1, \dots, \mu_{16})^T \in \mathbb{R}^{23}.
 \end{aligned}$$

We will use a numeric optimization over the profile likelihood. Thus, we must optimize over the covariance parameters θ and the mean parameters μ are determined implicitly by calculating the generalized least square estimate.

3.3 Observation locations

As the LV03 coordinates for the centroid's locations $s_i = (LV03x_i, LV03y_i)^T$ cover a fairly large range, we standardized them to kilometers using the following formula:

$$\begin{pmatrix} Z.LV03x_i \\ Z.LV03y_i \end{pmatrix} := 10^{-3} \cdot \left(\begin{pmatrix} LV03x_i \\ LV03y_i \end{pmatrix} - \begin{pmatrix} 600000 \\ 200000 \end{pmatrix} \right)$$

Table 3 Mean and covariance estimates $\hat{\omega}_{MLE}$ of the SVC model (5)

Covariates	Mean $\hat{\mu}_j$	Sign. level	Range $\hat{\rho}_j$	Variance $\hat{\sigma}_j^2$ and $\hat{\tau}^2$	Sign. level
Intercept	9.1555 (0.1571)	***	13.5511 (4.1157)	0.0413 (0.0116)	***
<i>Z.age</i>	-0.1495 (0.0254)	***	18.9370 (41.6414)	0.0009 (n.a.)	n.a
<i>Z.age</i> ²	0.0114 (0.0062)	.	1.7153 (0.8673)	0.0003 (0.0003)	n.s
log (<i>volume</i>)	0.5132 (0.0222)	***			
log (<i>plotsize</i>)	0.1720 (0.0129)	***			
<i>renov</i>	0.0274 (0.0059)	***			
<i>standard</i>	0.0911 (0.0082)	***			
<i>micro</i>	0.0385 (0.0075)	***			
<i>yearquarter</i> 20184	0.0040 (0.0160)	n.s			
<i>yearquarter</i> 20191	0.0235 (0.0137)	.			
<i>yearquarter</i> 20192	0.0327 (0.0135)	*			
<i>yearquarter</i> 20193	0.0357 (0.0135)	**			
<i>yearquarter</i> 20194	0.0279 (0.0149)	.			
<i>SFHtype</i> 2	-0.0062 (0.0120)	n.s			
<i>SFHtype</i> 3	-0.0274 (0.0165)	.			
<i>energy</i> 2	0.0119 (0.0281)	n.s			
Error term				0.0242 (0.0010)	

The corresponding estimates' standard errors are given in parenthesis. In most cases, the standard errors can be approximated and computed by the Hessian from the numeric optimization. For the mean and the variance estimates, we use a two-sided Z- and a Wald-test to test whether $\hat{\mu}_j \neq 0$ and $\hat{\sigma}_j^2 > 0$, respectively. This is only possible if the standard error is available.

Significance levels: " $p < 0.1$; " * $p < 0.05$; " ** $p < 0.01$; " *** $p < 0.001$; 'n.s.' not statistically significant, 'n.a.' not available

Again, this ensures a well-behaved numeric optimization while remaining interpretable as the ranges ρ_j now act as a scaling factor on the kilometer distances.

4 Results

4.1 Parameter estimates

We first look at the ML estimates $\hat{\omega}_{MLE}$, which are given in Table 3. Here, we find that the mean estimates match our expectations. In particular, the vintage-related covariates, i.e., *Z.age* and *Z.age*², show the following:

1. The mean effect for *Z.age* is negative, as one would expect, and statistically significant at the 0.1% level. This can be interpreted in the sense that on average the value of a single-family home typically decreases with age.
2. The quadratic effect for *Z.age*² shows a relatively small, positive mean effect, which is statistically significant at the 10% level. A larger quadratic mean age effect would correspond to an emphasized vintage effect.

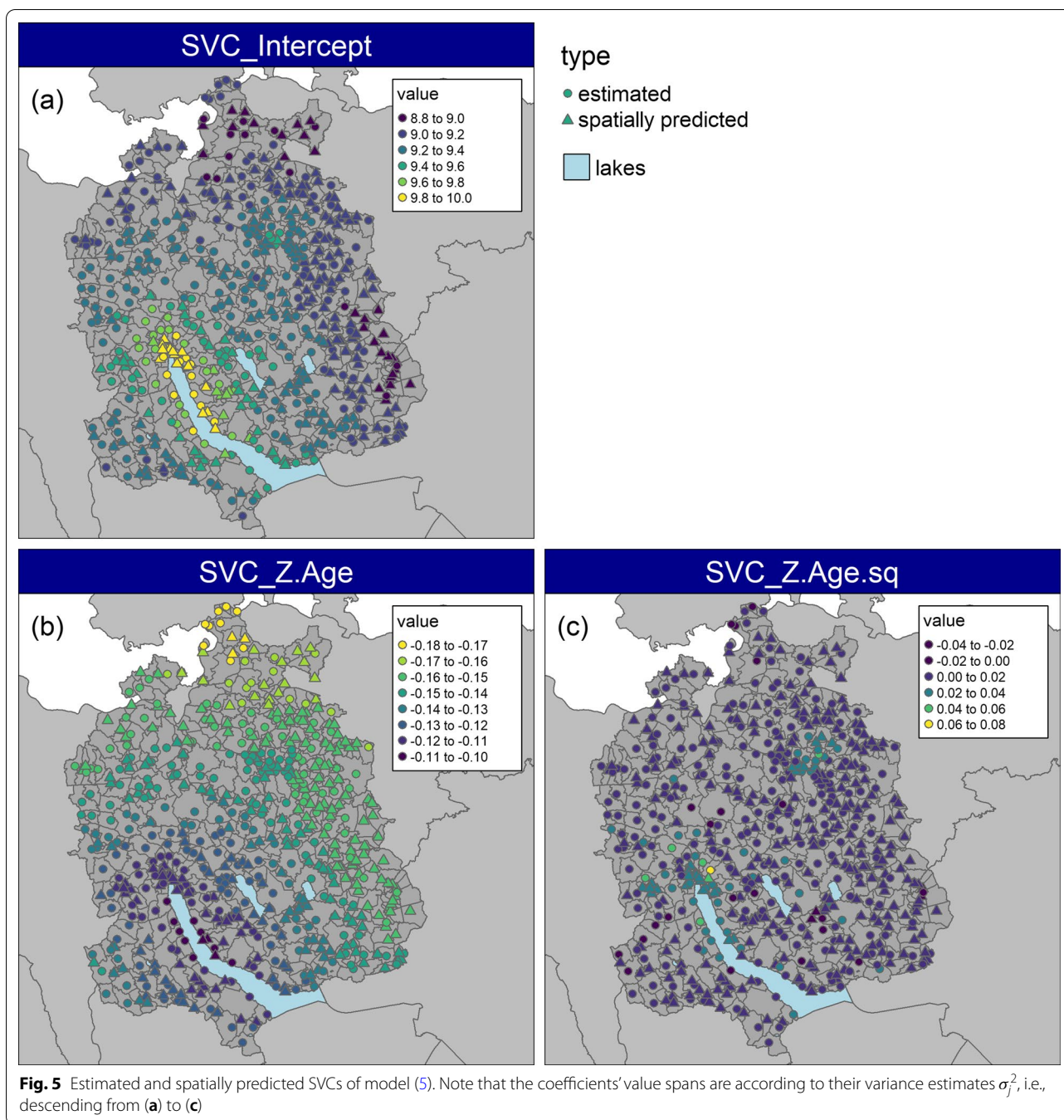
All other mean effects have plausible signs, too. Namely, all other coefficients of continuous covariates are positive and statistically significant at reference levels. As for the categorical covariates, we observe some temporal price volatility for the transaction year and quarter,

a premium for stand-alone, detached SFHs compared to other SFHs for the type of SFH and a premium for houses with enhanced energy efficiency for the energy standard.

The estimates for ranges of the Gaussian processes show that the range for the intercept and *Z.age* are considerably larger than the one for *Z.age*². This will be expressed in larger spatial structures for the SVCs modeling the intercept and the linear age effect compared to the SVC modeling the quadratic age effect. The small range of *Z.age*² on the other hand indicates that the SVCs corresponding to the quadratic age effect will behave much more selective in their deviations from the mean. Finally, we analyze the estimated variances of the spatially varying coefficients. The intercept's variance is the largest and highly statistically significant. As for the linear and quadratic age effects, the range of the coefficients' values is smaller, see Table 3 and Fig. 5.

4.2 Visualization and interpretation of SVCs

In Fig. 5 we visualize fitted and predicted SVCs. Specifically, the figure shows the estimated SVCs for the observation locations the model has been trained on as well as for the spatial predictions for all other cell's centroid where we did not have any observations. The quality of these coincides with the previous parameter estimates' interpretations from Sect. 4.1 and real estate experts' knowledge.



For the intercept's SVC, c.f. Fig. 5a, which also can be interpreted as a mean price level, we can see that the highest values are achieved close to the city of Zurich and Lake Zurich, with a local peak in the city of Winterthur. As expected, the lowest values can be found towards the northern and eastern borders of ZH, which are rural areas.

A similar pattern as for the intercepts' SVC can be observed for the *Z.age* SVC, c.f. Fig. 5b. In

absolute values, the linear age effect is smallest in the city of Zurich and in the area surrounding Lake Zurich, while it increases towards the northern and eastern regions of the Canton of Zurich. This indicates that the depreciation of SFH prices by age is higher in rural areas. For the *Z.age*² SVC, c.f. Fig. 5c, small scale deviations from the mean effect can be observed around Lake Zurich and the city of Winterthur. This hints at the presence of a vintage effect in the metropolitan areas of Zurich and

Table 4 Summary statistics of SVCs

	Intercept $\hat{\beta}_1(\cdot)$	$Z.age$ $\hat{\beta}_2(\cdot)$	$Z.age^2$ $\hat{\beta}_3(\cdot)$
<i>(a) Estimated</i>			
Minimum	8.894	-0.172	-0.024
Mean	9.315	-0.139	0.013
Maximum	9.905	-0.107	0.067
<i>(b) Spatially predicted</i>			
Minimum	8.919	-0.172	-0.003
Mean	9.257	-0.142	0.012
Maximum	9.897	-0.108	0.044

Winterthur. The resulting coefficient values for the three SVCs are summarized in Table 4.

The individual interpretation of both panels b and c in Fig. 5 is cumbersome and inadequate as the fitted SVCs originate from the same covariate. As we are simultaneously modeling a linear and quadratic effect, one could therefore interpret the results as spatially varying paraboids. Using the SVCs $\hat{\beta}_2(\cdot)$ and $\hat{\beta}_3(\cdot)$ for all observation locations s_{train} within the training data, we back-transform the estimated effects to receive the marginal effect $me(s_{train}, age)$ for the $age \in [-1, 99]$:

$$me(s_{train}, age) := \hat{\beta}_2(s_{train}) \cdot \frac{age}{s_{age}} + \hat{\beta}_3(s_{train}) \cdot \left(\frac{age}{s_{age}}\right)^2. \tag{6}$$

This is what we visualize in Fig. 6. The grey lines are the marginal effects $me(s_{train}, age)$, grouped in panels by FPRE type of location and filtered such that (i) there are at least 5 observations per location s_{train} and (ii) cropped to the span of observed years of construction at the corresponding location. This is to ensure that we have sufficient data backing up the results and that we do not extrapolate to unobserved building age. The red line is obtained by aggregating all marginal effects by type of location, i.e.,

$$me_{\kappa}(age) = \frac{1}{|S_{\kappa}|} \sum_{s \in S_{\kappa}} me(s, age), \tag{7}$$

where $S_{\kappa}, \kappa \in \{1, 2, 3, 4\}$ are the sets of all observations s in respective type of location and $age \in [-1, 99]$. We observe a pure depreciation for a majority of SFHs with $age < 25$. This holds not only for the aggregated age effects, but also for most individual age effects per cell. It is at this point ($age > 25$) that the marginal age effects start to differ. Aggregated on the type of location, we see a strong vintage effect for top locations while all other types of locations have aggregated marginal age effects of pure depreciative nature.

Looking at the individual cell's marginal age effects, one can observe some variety within each location type. Overall, it motivates the usage of spatially varying random effects, since the type locations cannot account for the geographical variety. At top locations a vintage effect is present such that some of the oldest SFHs have the same marginal

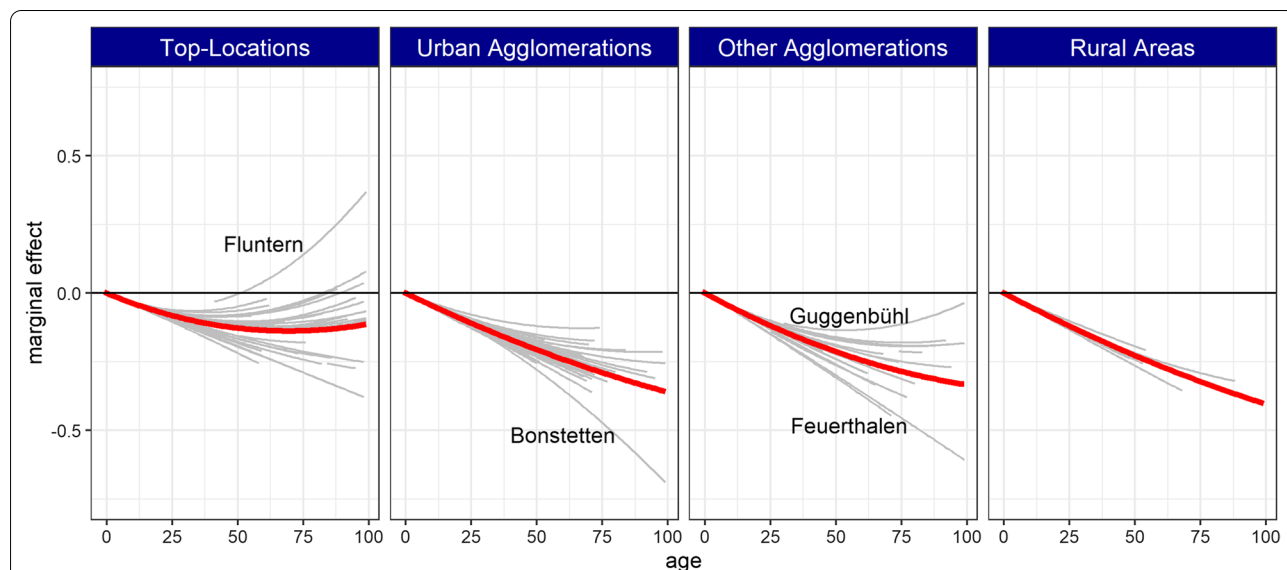


Fig. 6 Back-transformed, aggregated effect of age. The grey curves correspond to the marginal effects as defined in (6) and the red lines are the aggregated marginal effects as defined in (7). The most extreme effects are displayed with their respective cell names, i.e., Fluntern (a district of the city of Zurich), Bonstetten (a suburb to the West of the city of Zurich), Guggenbühl (a district of the city of Winterthur) and Feuerthalen (a district bordering the city of Schaffhausen)

effect for age as just recently built SFHs. Here, one location (Fluntern, a district of Zurich) stands out as it gives a premium of 0.3 log-points for a SFH built in the 1920s compared to a recently built SFH, *ceteris paribus*. A variety of different marginal age effects per cell is also present in urban and other agglomerations. Here we see various rates of price depreciation and in some cases also a vintage effect. For instance, Guggenbühl, a district of Winterthur, shows a rather constant age effect while Feuerthalen and Bonstetten show a steep depreciation with age. Upon further investigation of the latter two, we note that the estimates in Bonstetten are primarily driven by two observations of very old SFHs with low transaction prices. Regarding the district of Feuerthalen it is worth mentioning that that Feuerthalen is located in the very north of the Canton of Zurich bordering the city of Schaffhausen. Since we do not have data on the city of Schaffhausen in our dataset, margin effects could come into play here, and therefore also the estimated marginal effects for Feuerthalen should be treated with caution. An in-depth analysis for all of these individual age effects is out of scope for this work. However, these insights underline the broad variety location-specific effects that is hard to account for by regular fixed effects, say, like interactions between the covariates of age and regionalization factors.

We conclude this section by noting that we observe spatially varying age effects which clearly deviate from a pure depreciation. These effects are locally pronounced and mostly appear at top locations, which backs our hypothesis of spatially varying vintage effects and is in line with both initial citations taken from Case et al. (2004) and Malpezzi et al. (1987).

5 Predictive performance

We now assess the implications of our findings on predictive performance. As suggested in Sect. 4, there appears to be a spatially varying age effect that deviates from a linear depreciation with age. Now, we investigate if one can exploit this to enhance classical hedonic models to increase predictive performance.

We validate and compare our findings to a classical hedonic model with only the mean price, i.e., the intercept depending on spatial location s . Thus, we use a geostatistical model similarly defined as the SVC model in (5) but with $q = 1$:

$$\begin{aligned}
 y_i = \log price_i = & \beta_1(s_i) + \beta_2 \cdot Z.age_i + \beta_3 \cdot Z.age_i^2 \\
 & + \beta_4 \cdot \log volume_i + \beta_5 \cdot \log plotsize_i \\
 & + \beta_6 \cdot renov_i + \beta_7 \cdot standard_i + \beta_8 \cdot micro_i \\
 & + \beta_9 \cdot yearquarter_i + \beta_{10} \cdot SFHtype_i + \beta_{11} \cdot energy_i + \epsilon_i.
 \end{aligned}
 \tag{8}$$

To compare the two models (5) and (8), we conducted a tenfold cross validation that accounts for the temporal structure of the data. The first 5 quarters of transaction

Table 5 Summary of tenfold cross-validation

Type	Median RMSE(m, f)		Improvement (%)
	Geostatistical model (8)	SVC model (5)	
In-sample	0.157	0.145	7.7
Out-of-sample	0.208	0.179	13.9

Median in-sample (9) and out-of-sample (10) RMSE of the geostatistical model (8) and SVC model (5) with their respective percentage improvement

data were exclusively used as training data. We randomly divided the observations from the last quarter, i.e., the 4th quarter of 2019, into 10 sets $V_f, f = 1, \dots, 10$, of 20 or 21 observations each. In each fold f , the observations V_f were withheld from training to provide a validation set. Therefore, for all folds f the training data denoted $T_f = \{1, \dots, 1578\} \setminus V_f$ consists of 1557 or 1558 observations while the validation set V_f consists of 21 or 20 observations from the 4th quarter of 2019. In such a way, we account for the temporal structure of the data.

The root mean square error (RMSE) is chosen as a measure of comparison and computed for in-sample estimates and out-of-sample predictions. Let $\hat{y}_i(m, f)$ denote the estimate or prediction of y_i by model m in fold f , i.e., if $i \in V_f$ we have an out-of-sample prediction and if $i \in T_f$ we have an in-sample estimate. For each model m and fold f , we define the RMSE of in-sample estimates and out- of-sample predictions as:

$$RMSE(m, f) = \sqrt{\frac{1}{|T_f|} \sum_{i \in T_f} (y_i - \hat{y}_i(m, f))^2}, \tag{9}$$

$$RMSE(m, f) = \sqrt{\frac{1}{|V_f|} \sum_{i \in V_f} (y_i - \hat{y}_i(m, f))^2}, \tag{10}$$

respectively. We report the respective medians over all folds as well as the percentage improvement in Table 5. The in-sample performance of the geostatistical model is improved by 7.7% by using the SVC model. This is to be expected, as the SVC model (5) offers more flexibility and the geostatistical model (8) is a true sub-model of the SVC model. For the out-of-sample predictions we find that the absolute error values are higher compared to their in-sample counterparts, which again is to be expected. In addition, we observe an 13.9% improvement in price prediction. We thus conclude that accounting for spatially varying age effects using SVC models as discussed in the previous section, translates into more accurate out-of-sample predictions.

Table 6 Share of SFHs built before a specified date by canton, as of 2019. *Data:* Federal Statistical Office FSO (2021)

Area (country; canton)	Canton code	Before 1945	Before 1960	Before 1970
Switzerland	–	23	34	44
Aargau	AG	17	29	38
Appenzell Ausserrhoden	AR	42	49	59
Appenzell Innerrhoden	AI	36	42	51
Basel Landschaft	BL	16	28	40
Basel Stadt	BS	61	81	86
Bern	BE	22	35	46
Fribourg	FR	14	20	27
Geneva	GE	23	32	42
Glarus	GL	49	59	66
Graubünden	GR	35	43	53
Jura	JU	24	36	47
Lucerne	LU	13	20	28
Neuchâtel	NE	26	38	47
Nidwalden	NW	17	27	39
Obwalden	OW	26	35	48
Schaffhausen	SH	32	41	46
Schwyz	SZ	14	22	32
Solothurn	SO	19	32	42
St. Gallen	SG	23	32	43
Thurgau	TG	21	29	36
Ticino	TI	40	57	65
Uri	UR	26	36	45
Valais	VS	19	27	37
Vaud	VD	24	33	43
Zug	ZG	15	23	32
Zurich	ZH	25	36	44

The respective shares are given in percent

6 Conclusion

To the best of our knowledge, the presented work is the first of its kind to investigate a spatially varying age effect for SFHs. While we find a purely depreciative age effect for some locations in the Canton of Zurich, there appears to be a substantial price premium for older SFHs, primarily at top locations. The existence of a not purely depreciative age effect is in line with the scientific literature and the assumptions of real estate experts. Further, even if there is no vintage effect present, we observe various grades of age depreciation by location. In this context, we consider it likely that age acts as a proxy for unmeasured covariates that directly have an impact on prices, such as quality of built or architectural style (e.g., room height, architectural details) of the object as has been suggested by the existing literature (Case et al., 2004; Goodman & Thibodeau, 1995). Our findings are also in line with a relatively new concept of redevelopment options as suggested by Clapp and Salavei (2010),

where further analyses by Munneke and Womack (2016) also showed substantial spatial variation of such redevelopment options. However, both referenced works analyze data from the United States. It would be interesting to see if these concepts transfer to hedonic models based on Swiss SFHs. However, it is out of scope for this work.

Finally, our assessment of the predictive performance in a cross validation yields more accurate predictions from an SVC model with spatially varying age coefficients than a classical geostatistical model.

Overall, our analysis suggests a spatially varying vintage or at least location specific age effect. Further research on the topic based on data from different regions or with higher resolution would be desirable.

Appendix

Table 6 provides a summary of the SFHs' age structure mentioned in the Introduction.

Abbreviations

FPRE: Fahrländer Partner Raumentwicklung; GP: Gaussian process; h.m.b.: Higher meaning better; ML(E): Maximum likelihood (estimation); RMSE: Root mean square error; SFH: Single-family house; SVC: Spatially varying coefficient; ZH: Canton of Zurich.

Acknowledgements

We would like to thank the anonymous reviewers and the editor for their valuable feedback, which improved the quality of this manuscript.

Authors' contributions

JD and FS contributed the statistical fundamentals serving as a basis for this paper. The model estimates and other calculations were carried out by SK, with support from JD. An analysis of the data and results was performed by JD and SK. The interpretation of the results was performed to a large extent by JD, with support from ML, AZ and JS. JD was responsible for writing the paper, with selective contributions from SK. FS, ML, AZ and SF have revised the paper and have initiated a number of changes to the paper. All authors read and approved the final manuscript.

Funding

This study was jointly funded by Innosuisse (the Swiss Innovation Agency) and Fahrländer Partner Raumentwicklung (FPRE) in the framework of a project on space-time machine learning models for valuation and prediction of real estate objects (Innosuisse Project Number 28408.1 PFES-ES). The design of this study, the collection, analysis and interpretation of the data and the writing of the manuscript were not influenced by the funding body.

Availability of data and materials

The data used for this analysis is subject to Swiss banking secrecy and can therefore neither be made available publicly nor on request.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Mathematics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland. ²Institute of Financial Services Zug (IFZ), Lucerne University of Applied Sciences and Arts, Suurstoffi 1, 6343 Rotkreuz, Switzerland. ³Fahrländer Partner Raumentwicklung, Seebahnstrasse 89, 8003 Zurich, Switzerland.

Received: 31 August 2021 Accepted: 2 December 2021

Published online: 06 January 2022

References

- Banerjee, S., Gelfand, A. E., Finley, A. O., & Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B*, 70(4), 825–848.
- Brunauer, W. A., Lang, S., Wechselberger, P., & Bienert, S. (2010). Additive hedonic regression models with spatial scaling factors: An application for rents in Vienna. *Journal of Real Estate Finance and Economics*, 41(4), 390–411.
- Cao, K., Diao, M., & Wu, B. (2019). A big data-based geographically weighted regression model for public housing prices: A case study in Singapore. *Annals of the American Association of Geographers*, 109(1), 173–186.
- Case, B., Clapp, J. M., Dubin, R., & Rodriguez, M. (2004). Modeling spatial and temporal house price patterns: A comparison of four models. *Journal of Real Estate Finance and Economics*, 29(2), 167–191.
- Clapp, J. M., & Giaccotto, C. (1998). Residential hedonic models: A rational expectations approach to age effects. *Journal of Urban Economics*, 44, 415–437.
- Clapp, J. M., & Salavei, K. (2010). Hedonic pricing with redevelopment options: A new approach to estimating depreciation effects. *Journal of Urban Economics*, 67(3), 362–377.
- Cressie, N. (2011). *Statistics for spatio-temporal data*. Wiley.
- Dambon, J. A., Sigrist, F., & Furrer, R. (2021a). Maximum likelihood estimation of spatially varying coefficient models for large data with an application to real estate price prediction. *Spatial Statistics*. <https://doi.org/10.1016/j.spa.2020.100470>
- Dambon, J. A., Sigrist, F., & Furrer, R. (2021b). *varycoef: An R package for Gaussian process-based spatially varying coefficient models*. Retrieved from 1 Aug 2021. <https://arxiv.org/abs/2106.02364>.
- Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical nearest neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514), 800–812.
- Fahrländer, S. S. (2006). Semiparametric construction of spatial generalized hedonic models for private properties. *Swiss Journal of Economics and Statistics*, 142(4), 501–528.
- Federal Office of Topography swisstopo. (1900). *LV03*. Retrieved from 1 Aug 2021. <https://www.swisstopo.admin.ch/en/knowledge-facts/surveying-geodesy/reference-frames/local/lv03.html>
- Federal Statistical Office FSO. (2021). *Gebäude nach Kanton, Gebäudekategorie, Bauperiode und Jahr*. Retrieved from STAT-TAB: 1 Aug 2021. https://www.pxweb.bfs.admin.ch/pxweb/en/px-x-0902010000_101/-/px-x-0902010000_101.px/table/tableViewLayout2/
- Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Wiley.
- Gelfand, A. E., Kim, H.-J., Sirmans, C. F., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462), 387–396.
- Gelfand, A. E., & Schliep, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, 18(Part A), 86–104.
- Geng, J., Cao, K., Yu, L., & Tang, Y. (2011). Geographically Weighted Regression Model (GWR) based spatial analysis of house price in Shenzhen. In *2011 19th international conference on geoinformatics* (pp. 1–5).
- Goodman, A. C., & Thibodeau, T. G. (1995). Age-related heteroskedasticity in hedonic house price equations. *Journal of Housing Research*, 6(1), 25–42.
- Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., & Zammit-Mangion, A. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3), 398–425.
- Malpezzi, S., Ozanne, L., & Thibodeau, T. G. (1987). Microeconomic estimates of housing depreciation. *Land Economics*, 63(4), 372–385.
- Munneke, H. J., & Womack, K. S. (2016). Valuing the redevelopment option component of urban land values. *Real Estate Economics*. <https://doi.org/10.1111/1540-6229.12192>
- R Core Team. (2020). *R: A language and environment for statistical computing*. Retrieved from <http://www.R-project.org/>.
- Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press.
- Roberts, S., Osborne, M., Ebdon, M., Reece, S., Gibson, N., & Aigrain, S. (2013). Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A*, 371, 20110550.
- Rubin, G. M. (1993). Is housing age a commodity? Hedonic price estimates of unit age. *Journal of Housing Research*, 4(1), 165–184.
- SIA Zürich. (2003). *SIA 416, Flächen und Volumen von Gebäuden*. SIA Zürich.
- Statistisches Amt des Kantons Zürich. (2021a). *Freihandverkäufe von Immobilien*. Retrieved from 1 Aug 2021. <https://www.zh.ch/de/politik-staat/statistik-daten/datenkatalog.html#/details/77@statistisches-amt-kanton-zueri-ch>
- Statistisches Amt des Kantons Zürich. (2021b). *Quartalsbericht Handänderungsstatistik (Daten)*. Retrieved from 1 Aug 2021. <https://www.zh.ch/de/planen-bauen/raumplanung/immobilienmarkt.html#2055805364>
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit Region. *Economic Geography*, 46(Supplement), 234–240.
- van Eggermond, M., Lehner, M., & Erath, A. (2011). *Modeling Hedonic Prices in Singapore*. Retrieved from https://www.researchgate.net/publication/266868391_MODELING_HEDONIC_PRICES_IN_SINGAPORE
- Wheeler, D. C., & Calder, C. A. (2007). An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, 9(2), 145–166.
- Wheeler, D. C., Páez, A., Spinney, J., & Waller, L. A. (2014). Bayesian hedonic price analysis. *Papers in Regional Science*, 93(3), 663–683.
- Wheeler, D. C., & Waller, L. A. (2009). Comparing spatially varying coefficient models: A case study examining violent crime rates and their

relationships to alcohol outlets and illegal drug arrests. *Journal of Geographical Systems*, 11(1), 1–22.

Wu, Y., Hernández-Lobato, J. M., & Ghahramani, Z. (2014). Gaussian process volatility model. In *Advances in neural information processing systems 27 (NIPS 2014)*.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
