Stabilizing Geo-Spatial Splines with Helper Points: How to Estimate Smooth Price Surfaces when there are Data Gaps

Norbert Pfeifer
University of Graz, Austria
norbert.pfeifer@uni-graz.at

Miriam Steurer
University of Graz, Austria
miriam.steurer@uni-graz.at

Paper prepared for the 37th IARIW General Conference

August 22-26, 2022

Poster Session

Time: Wednesday, August 24, 2022 [17:30-18:30 CEST]

Stabilizing Geo-Spatial Splines with Helper Points: How to Estimate Smooth Price Surfaces when there are Data Gaps

Norbert Pfeifer
norbert.pfeifer@uni-graz.at

Miriam Steurer
miriam.steurer@uni-graz.at

Department of Economics, University of Graz, Austria

Preliminary draft – please do not circulate

Version: 01.08.2022

Abstract

This paper examines how to overcome an essential disadvantage of polynomial spline behavior: overshooting of estimated spline functions in areas with poor data support. We introduce a new method that avoids the spline overshooting problem by placing helper points in data-gap areas before estimating the spline surface. We estimate helper point values via the Random Forest algorithm. Helper points force the algorithm to put a cost on deviating from reasonable local values in these areas. We show that our method can prevent spline overshooting where data are missing, can improve predictions in areas where data are scarce, but does not distort the spline surface in areas where data are plentiful. Our method also has a positive knock-on effect in that it reduces the need for high (global) penalisation values and thus improves the spline's response to changes in actual prices in regions with more data. Our method is particularly suited to the estimation of property price gradients, as property data are inherently unevenly distributed in space. We illustrate that our method can significantly improve the estimation of regional house price gradients using data for new apartment transactions in Vienna, Austria. To the best of our knowledge, our method is new - not only to the field of Real Estate Economics - but also to the spline literature.

Keywords: Penalized Regression Splines, Multilateral Splines, Random Forest, House Price Surface, Spatial Testing

1 Introduction

Estimates of how house price levels vary over geographic areas are useful to discover regional sub-centers (e.g., McMillen (2001)), to indicate the value of local amenities like public schools (e.g., Gibbons and Machin (2003)), and could be a useful input for quantitative spatial models (see Allen and Arkolakis (2014) or Ahlfeldt et al. (2015). These price/value surfaces can also replace local fixed effects in hedonic house price models, see e.g., Clapp (2004), Hill and Scholz (2018), Melser and Hill (2019), and Kholodilin et al. (2021).

The most poplular way to interpolate price/value surfaces in the real estate literature are locally-weighted regression and kernel methods see for example, McMillen (1996), Thorsnes and McMillen (1998), McMillen (2001), Gibbons and Machin (2003), and (Clapp, 2004)). Recently, penalized regression splines have been introduced for this task (see Craig and Ng (2001), Bao and Wan (2004), Hill and Scholz (2018), Melser and Hill (2019), and Kholodilin et al. (2021). In many ways, penalized regression splines are better suited to model three dimensional price/value surfaces than other methods. They are excellent for smoothing noisy data, are more flexible than other parametric or non-parametric data interpolation methods, and do not impose any symmetric or convex restrictions onto the surface (Liu et al., 2016). In contrast to locally-weighted and kernel regressions, they can handle irregularly shaped non-convex regions which often occur with housing data. In the context of real estate price/value surfaces this implies that splines are better able to handle sudden changes in property values – which can occur, for example, at school-district boundaries, rivers, or large roads. Furthermore, splines have a number of favorable axiomatic theoretical properties and are more efficient in their computation than other non-parametric techniques (see Schoenberg (1964), Wahba (1990), Wahba and Wang (2017), and Wood (2017), page 121).

However, penalized regression splines have one problem: they are bad at extrapolating into areas with little data support where they are prone to "overshoot" to unrealistic values. This behaviour is a consequence of how penalized regression splines are constructed, which is illustrated in Equation 1 for a univariate regression spline. The objective is to find the function $\hat{f}(x)$ that solves the following problem:

$$\min_{f(.)}\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx\}. \tag{1}$$

This expression has two parts: the first term is the squared deviation from the fitted function, which provides the goodness of fit, while the second term imposes a cost against overly "wiggly" functions by adding a penalization term based on the function's second derivative. We will discuss the make-up of penalized regression splines in more detail in section 2. For now, we like to draw attention to how the spline operates in data gap areas. In those areas, given that there are no data points, there is also no cost of deviating from "normal" values. Thus, the first term of Equation 1 will be zero regardless of the spline function's value, and the only guidance for splines is to minimize the overall "wiggliness". The natural consequence is the spline behavior illustrated in Figure 1, namely that the spline minimizes its second derivative by forming a "peak" or "trough" (overshooting). All other things equal, the larger the data gap, the larger the "overshooting" of the spline function.

Sparse data is the norm rather than the exception in Real Estate and Urban Economics. For example, if we aim to establish a price/value surface map for residential housing in a non-urban region we will find that real estate data is sparsely distributed throughout physical space. And even for urban areas, there will be multiple areas without data support for residential housing, for example, because the space is occupied by industrial land, a lake, or a public park. Let us consider the effect of a park on housing values and the shape of the estimated spline function: If the park is well-maintained, closeness to the park will increase property values, resulting in a rising price gradient when approaching the park (from any direction). We will thus find that our price surface tends to behave as illustrated in Figure 1: Overshooting of price estimates occurs and will be the more extreme, the larger the park (and the more attractive its vicinity).[1]

[1] In the case of local "bads" (e.g., a local garbage disposal site) the direction of the spike will tend to be negative as the

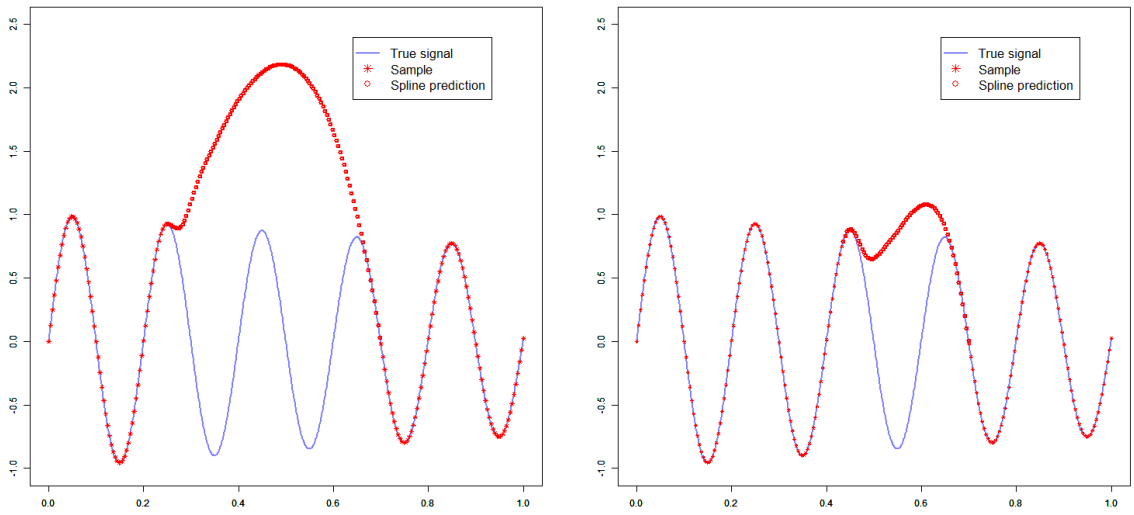


Figure 1: Spline behaviour by different length of data gaps
Note: the "wiggle" in Figure 1 (a) and (b) occurs because the spline is constructed with third-degree spline basis functions.

The existance of data gaps hinders the interpretation and further use of spline surfaces as inputs into hedonic price indices or for quantitative spatial models, since users cannot be sure whether price peaks (or troughs) are due to genuine price hot-spots (or low-spots) or the result of a data gap. Cutting areas without data support from the price/value surface either before or after spline estimation is one option to deal with this issue. For example, soap film splines (Wood et al., 2008) cut out data gaps before spline estimation.[2] However, "cutting-out" data gap areas leads to holes in the resulting price landscape which is not very satisfactory.

The main contribution of this paper is to provide a new two-stage method for penalized regression splines that overcomes this data-gap problem. In the first stage, we find reasonable estimates for all data gap locations via the Random Forest (RF) algorithm (Breiman, 2001), which is a powerful yet straightforward hierarchical clustering technique that uses decision trees as its base learners.[3] We then plant helper points, the values of which consist of these local price averages from the RF model, in areas with no (or little) data support. Finally, we estimate a price spline surface with original data plus helper points. The resulting price gradients provide the granularity of the original penalized spline estimations in areas where enough data are located and extrapolate them with local average price levels where data support is missing. Our method automatically prevents the overshooting problem, and as a result, the spline surfaces can be interpreted without fear of misinterpreting an overshooting peak for a local housing hot-spot.

We illustrate our method by estimating a geo-spatial spline surface that indicates the potential square meter price for new-built properties in Vienna, Austria, for 2020.[4] Given that Vienna is a historical, densely built up city, these new-built properties are sparsely populated throughout the city. Also we know intuitively (and from transaction data of existing properties) how the price surface of a city like Vienna is "supposed" to look like, which makes it easier to spot overshooting problems.

By adding helper points in areas without data support, we prevent spline overshooting and improve the accuracy of the price estimates. We test our method extensively and illustrate how the size of the data gap and the number of helper points influence the results. Also, we repeatedly test model behavior

price gradient towards such places will tend to decrease.

[2] A real estate application of soap film splines is provided by Kholodilin et al. (2021).

[3] The RF algorithm is well suited to this purpose, however a number of other cluster- and average valuation mechanisms could be used for this purpose.

[4] Data come from the firm `http://www.zt.co.at/ztneu/index.html`.

with geographically defined hold-out (testing) samples of various locations and sizes. Our price map indicates the potential value of location anywhere throughout the city of Vienna. These estimates are informative in their own right, but can subsequently also be used as components to other models (e.g., hedonic regression models or quantitative spatial models).

To the best of our knowledge, our method is new - not only to the field of Real Estate Economics - but also to the spline literature. Our method could, therefore, not only improve the estimation of regional house price gradients, it also has the potential to improve a wide variety of spline applications in other fields.

The structure of the remainder of the paper is as follows: In section 2 we provide a short summary of spline techniques, focusing on penalized regression splines. We illustrate the problem of splines to provide suitable values for areas without data support (i.e., the overshooting problem) in section subsection 2.3. We present the dataset for our spline estimations in section 3. We illustrate our method of estimating splines with helper points in section 4 and show how it can improve results. section 6 concludes the paper.

2 Some Background on Penalized Regression Splines

2.1 What is a penalized regression spline?

Mathematically, an M-th order spline is a piecewise $M-1$ degree polynomial with $M-2$ continuous derivatives at the knots (i.e., the points where the piecewise regressions meat) (Wakefield, 2013). In practice, one can think of splines as flexible bands (in 2-dimensions) or flexible sheets (in 3-dimensions) that can easily describe any shape in mathematical terms.[5] It was this flexibility to describe any shape that made splines popular with engineers in the automotive and airplane industries in the 1950s and 1960s, where they became instrumental in shifting geometric design away from free-hand drawings toward computer-assisted methods (Davis, 1996).[6] [7]

Academic interest in splines started in the 1970s with the publications of d. Boor (1978) and Wahba and Wold (1975). Towards the end of the 1980s, Hastie and Tibshirani introduced the concept of Generalized Additive Models (GAMs), which provide a framework of connecting splines (as well as other smoothing techniques) with the structure of the generalized linear model (see, e.g., Hastie and Tibshirani (1986) and Hastie and Tibshirani (1990). Their 1990 book introduced splines to the broader statistical community (Hastie et al., 2009). More recently, Wood introduced the statistical community to thin-plate splines (see e.g., Wood (2003) and Wood (2017), which was first published in 2006). There are some excellent textbooks on splines, such as the classic books by Wahba (1990), Wood (2017) and chapter 5 of (Hastie et al., 2009).Overviews are also provided in Greiner (2009), Wahba and Wang (2017), and in the chapters on splines in Wakefield (2013) and Shalizi (2013).

Splines can be constructed in different ways; however, the basic distinction is between regression splines and smoothing splines. While regression splines build a spline by adding information, smoothing splines

[5] The term *spline* was introduced into the field of mathematics by Schoenberg in his seminal work on B-splines (see Schoenberg (1946a) and Schoenberg (1946b)). Originally, the name refers to a type of flexible ruler used by East Anglian shipbuilders (d. Boor, 1978).

[6] Many of the big developments in spline theory go back to these engineers in the automotive and airline industry (especially de Casteljau at Citroen and Bezier at Renault), but these developments were generally not published until much later. In the middle decades of the 20th century, splines were not only revolutionizing graphic design, but they also became an important intermediary between human computers and producing machines (e.g., for cutting material with the help of computer-directed machines). See d. Boor (1978) and Davis (1996) for a discussion on the early history of splines.

[7] The practical implications of this flexibility were first discovered by engineers in the automotive and airplane industries in the 1950s and 1960s, where splines were instrumental in shifting geometric design away from free-hand drawings towards computer-assisted methods Davis (1996).

build splines by successively reducing (unnecessary) information. In their pure forms, neither regression splines nor smoothing splines are optimal for most "real-world" data applications.[8]

Thus, in practice, regression and smoothing spline approaches are combined into a "hybrid" version: a regression spline with a large number of knots (more than considered necessary to define f(x)) that also includes a second-derivative penalization term to smooth the "wiggliness" of the overall function. This method is usually referred to as either penalized spline or penalized regression spline.[9] The idea for this hybrid approach was first presented in Wahba (1980) and O'Sullivan (1986), and later popularised by Eilers and Marx (1996).

For practitioners, the popularity of penalized (regression) splines comes from their ability to model any data structure flexibly. Within statistical circles, their popularity increased when their close correspondence to Baysian and mixed models became better understood (Wahba (1978), Eilers and Marx (1996), Brumback et al. (1999), and Ruppert et al. (2003)).[10] Note that we will focus on penalized regression splines for the remainder of this paper.

The objective of a penalized regression spline is to solve the minimization problem stated in Equation 1, which is restated below. That is, finding the function $\hat{f}(x)$ that solves the following problem:

$$\min_{f(.)}\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx\}. \tag{2}$$

This expression has two parts: the first term is the squared deviation from the fitted function, which provides the goodness of fit (the same as any "normal" OLS expression). The second term imposes a cost against overly "wiggly" functions by adding a penalization term in the form of λ times the second derivative.

The $f(x)$ in Equation 1 is built from a combination of simple basis functions so that the resulting curve is continuous in value and first and second derivative (see, e.g., Wood (2017)). These basis functions can be tailored to the problem at hand, with cubic and thin-plate basis functions being the most popular options.[11] When we optimize the spline function, we find coefficients for each of these basis functions such that together they approximate the underlying data-generating function. The value of the spline function at point x is then simply the sum of these estimated basis functions at x: [12]

$$f(x) = \sum_{k=1}^{K} \beta_k b_k(x), \tag{3}$$

where the $b_k(x)$ represent the basis functions for vector x.

[8]The practical difficulty with pure regression splines lies in their need to have the optimal number and location of knots in advance of the spline estimation. While these pure regression splines are helpful for some graphic and computer-machine interaction applications, they are unsuitable for our purpose of estimating a price surface with unseen real estate price data (and most other data applications). Smoothing splines start with all data points as potential knots. They are estimated via least square regression to which a smoothing parameter (penalty term) is added which controls the trade-off between accuracy (fitting data points) and "wiggliness" (smoothness) of the function estimate. The practical difficulty with pure smoothing splines is one of computation complexity: starting with *all* possible knot combinations make pure smoothing splines computationally too complex to be useful in practice.

[9]Alternatively, the hybrid can be described as a smoothing spline with a reduced knot set. Sadly, there is no consistent terminology for this hybrid type in the spline literature.

[10]This correspondence implies that Bayesian analysis can be used for penalized regression splines, provides the non-parametric spline method with econometric credibility as well as the possibility to test its output with the help of likelihood ratio tests or Markov chain Monte Carlo techniques (see, e.g., Crainiceanu et al. (2005)).

[11]A multitude of other basis functions, such as B-splines d. Boor (1978), P-splines (based on work by O'Sullivan, 1986; and popularised by Eilers and Marx (1996) and Marx and Eilers (1998)), soapfilm splines Wood et al. (2008), truncated power series, cardinal splines, and many more, can be implemented via many different R-packages. A quick search on the internet provided us with over 100 spline packages in R with many alternative basis functions implemented.

[12]Typically, the value of basis functions is zero outside their knot range.

The smoothing/penalization parameter λ in equation Equation 1 is another essential part of the spline construction process as it determines the bias-variance trade-off. If data are plentiful, it can be shown that as long as the degree of the base functions is large enough (i.e., at least flexible enough to represent $f(x)$), then neither the exact choice of base function nor the placement of knots has a large influence on the overall model fit, (Wood, 2017) or (Wahba and Wang, 2017). Thus, by turning to one of the many R-packages on spline construction and tuning only the smoothing/penalisation parameter λ, it is easy for users to construct splines in practice.

Note that penalization is a global approach, even if the underlying spline function is built from locally estimated basis functions. Higher values of λ make changes in slope more costly and are therefore associated with less wiggly overall functions; at the limit, as λ goes to infinity, the smoothing function becomes a linear function. On the other hand, small λ values allow the curve to fit the data very closely, and when λ equals zero, the curve will go through every data point. The impact of varying λ values is illustrated below in Figure 2, which illustrates various spline estimates for the longitude value that goes through the Vienna city center.

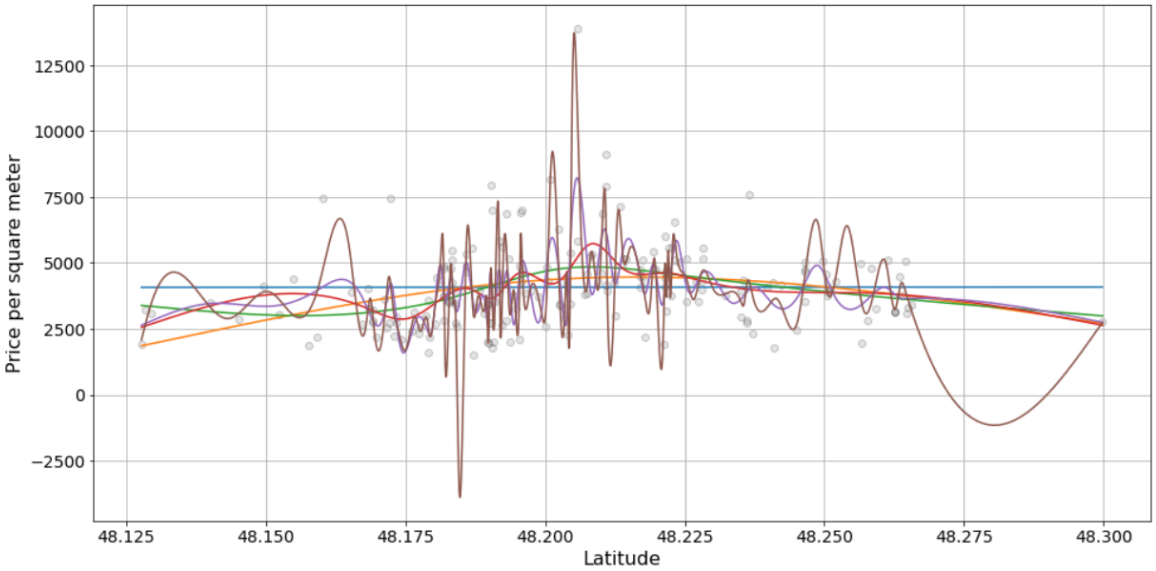


Figure 2: Spline interpolation depending on penalty term
All other things equal, the lower the penalisation term λ, the more extreme the spikes.

To choose the "best" fit smoothing spline (i.e., choose the optimal λ), Wahba and Wold (1975) developed the process of cross-validation, the process of "training" the model on (rotating/ alternating) subsets of the available input data and "testing" (evaluating) its performance on the unseen data points. Cross-validation quickly became a successful tool in its own right for choosing optimal parameter values (and to prevent overfitting) for non-parametric models in general. Today, many other methods, e.g., Bayesian inference-based methods Wood (2011), are also available to choose optimal λ values. In this paper, we use generalized cross-validation (GCV), a variant of Wahba and Wold's method (Wahba and Wold, 1975), to test the performance of our models in section 4.

2.2 Multivariate Splines

For geo-spatial splines, we typically want to model location values relative to both longitude and latitude dimensions. So we need spline techniques that can take at least two input variables. There are basically two ways to achieve this: The first way is to construct the spline out of basis functions that support multiple variables, in which case we talk about "multivariate splines". The second way is to construct multi-dimensional splines by forming tensor product smooths by estimating the spline surface as a tensor

product of unilateral spline functions. Both approaches have advantages and disadvantages, and various R-packages can implement both. For the remainder of the paper, we concentrate on multivariate splines.

In multivariate splines, to describe the price of land as a function of longitude and latitude, the three-dimensional price spline surface can be represented by a combination of three-dimensional basis functions that take *longitude* and *latitude* as inputs. A single (isotropic) penalization parameter λ is established and applied in each direction (i.e., with respect to each input variable). There are now more possibilities concerning which second derivatives are chosen in the penalization term, but otherwise, the estimation of such multivariate splines is like the uni-variate case of creating penalized regression splines (see above).

As in the univariate case, there are many possibilities concerning the type of basis function; however, in practice, thin-plate basis functions are the most popular choice for multivariate regression splines. Thin-plate splines were first introduced by Duchon (1977) and later popularized by Wood (see, e.g., Wood (2003), Wood (2017)). Thin-plate splines are are computationally efficient (Wood et al. (2008), Wood (2017)) and can be easily implemented via the mcgv R-package (Wood, 2011). A large benefit of thin-plate splines is that they avoid the problem of knot placement when implemented via the mcgv R-package (see, e.g., Wood et al. (2008), Wood (2017)) .[13]

2.3 Overshooting – a problem when data are sparse

Spline overshooting occurs due to (regional) data gaps and is positively related to the degree of the spline basis functions used, the size of the data gap, and the slope of the spline near the boundary of the data gap. It is negatively related to the severity of the penalization term. For a given penalization term λ, the higher the degree of the basis functions, the larger the data gap, and the steeper the gradient at the last data-point before (or after) the gap, the higher the overshooting becomes. Small gaps in knot values are generally handled quite well.

To understand the problem, we need to look at how splines handle the interpolation between data points. Figure 1 illustrates the problem. Coming up to the data gap, the spline has a positive slope at the last data-point before the gap (knot k_j), while the data behind the gap requires the slope to be negative. For example, we can observe this type of property price behavior in the vicinity of a well-maintained park. Closeness to the park will increase property values, resulting in a rising price gradient when approaching the park (from any direction). The interpolation of the spline function will produce an estimated spline similar to the one shown in Figure 1 below: It will create a spike in the price level. The larger the park (and the more attractive its vicinity), the bigger the spike.[14]

Price spikes (or troughs) due to data gaps can be prevented by increasing the penalization term λ. However, this comes at the cost of making the splines less responsive to local price differences in regions where data are plentiful, which limits their potential to indicate finer details of a city's price structure, such as the discontinuities in house price values that can exist along school- or other administrative boundaries, or across rivers and busy roads. Thus, high penalization terms undo the advantages splines have over locally-weighted and kernel regressions in terms of flexibly describing the price landscape.[15] We illustrate the impact of varying λ in Figure 2, in the Appendix.

[13] A clever trick of dimensionality reduction achieves this via an eigenvalue decomposition: The mgcv package initially takes every data point as potential knot value (i.e., each longitude/latitude combination); having such a large basis would be impossible to handle computationally, so the package performs an eigenvalue-decomposition and then uses the k largest eigenvalues as the new basis. Most of the information on the original (too large) basis is retained but in a much-condensed form.

[14] In the case of local "bads" (e.g., a local garbage disposal site) the direction of the spike will tend to be negative as the price gradient towards such places will tend to decrease.

[15] However, other smoothing procedures like locally-weighted regressions and kernel smoothing are also prone to over-smoothing.

One possibility to deal with the overshooting problem is to decrease the degree of the basis functions as done by Diewert and Shimizu (2019) who establish the price surface for Tokyo with linear interpolations following a technique by Colwell (1998). However, much of the detail possible with, say, cubic or tensor product basis functions is lost by this approach. Another possibility to deal with the overshooting problem is to "cut out" certain areas and produce splines only in the regions around them. This can be done by using soapfilm-splines (Wood et al., 2008). Kholodilin et al. (2021) proceed in this manner; in their paper on the historical housing market for St. Petersburg, they removed all waterways and defined soapfilm splines on the resulting topography. This works when we know in advance the areas where data will not be available (e.g., lakes, rivers, large parks, or airports), even though the process of creating the appropriate topography can be tedious. Figure 3 illustrates the large proportion of "transaction free" areas in the market for new-built apartments in Vienna during 2020.

3 Data

We estimate thin-plate spline surfaces illustrating square-meter price levels. Our dataset consists of prices and sizes for all new-built apartments for Vienna, Austria for 2019 and 2020. This transaction dataset was provided to us by the firm *ZTdatenforum* which transcribed the records from the official deed book, the "Grundbuch".[16]

For the main part of this paper we concentrate on the data for 2020, which consists of 2906 new-built apartments. We exclude untypically large or expensive properties (those with prices over 2 million Euros or areas above 350 square-meters) and end up with 2835 properties in 641 locations. Their mean square-meter price was 5347 Euros, with a standarad deviation of 2177. Figure 3 illustrates the geographical dispersion of these data throughout Vienna.[17]

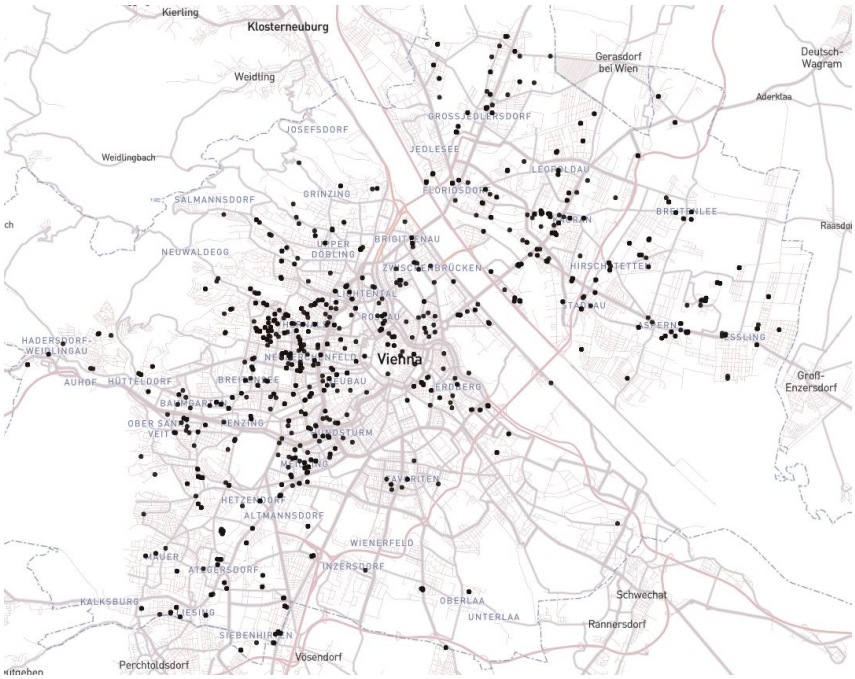


Figure 3: Transactions for new-built apartments in Vienna for 2020
The figures indicate the uneven distribution of new-built apartments throughout Vienna in 2020.

[16] http://www.zt.co.at/ztneu/index.html
[17] The locations of new-built properties transacted in 2019 as well as 2020 data are shown in Figure 10.

4 Method and Application

4.1 Spline surface with helper points

The principles of our method are simple: As regional gaps in the data structure are the main impediment to using splines to their full potential in urban economics, we eliminate these data gaps by adding helper points that represent the local price average into these gap areas and then proceed with the spline estimation as usual. These helper points introduce a cost for overshooting in the areas where such spline behaviour would otherwise be unpunished. As we add the helper points only in areas in which we have no transaction data, they do not interfere with the spline generating process in areas where data are plentiful, but they stabilize the spline function in areas where data gaps exist. The principle of our method is related to the practice of putting "clamps" on two-dimensional splines to stop them from overshooting at the ends."Clamped" splines are referred to as "natural" splines in the literature and are typically obtained by setting the second derivatives of the spline polynomials to 0 at the end knots (see e.g., Wood (2017)). However, our helper point method is more general than placing clamps at the endpoints of splines. Our method of placing helper points based on local average values into data gap areas is not limited to end points only and can – in contrast to clamps – be applied to higher dimensional spline surfaces.

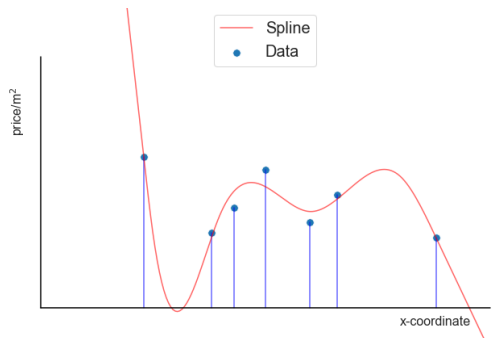
Here, we estimate helper points via the Random Forest (RF) algorithm (Breiman (2001)), a popular non-parametric decision-tree-based method that is widely available in R, Phyton, or Stata.[18] The Random Forest algorithm uses individual trees as base learners, which it then aggregates (see e.g. Hastie et al. (2009)).[19] Decision tree-based mechanisms are ideal for generating helper points because they simultaneously cluster adjacent transactions of similar price levels and provide the local price averages. Their particular strength is their ability to locate structural breaks in the output variable. Further benefits are that decision tree methods are also robust to outliers (as they fit the average rather than the extremes of the distribution) and can easily handle multi-dimensional data input.[20] An example of the type of data segmentation generated by decision-tree-based algorithms is shown in Figure A1 in Appendix A.

Figure 4 provides a graphical representation of our step-by-step process.

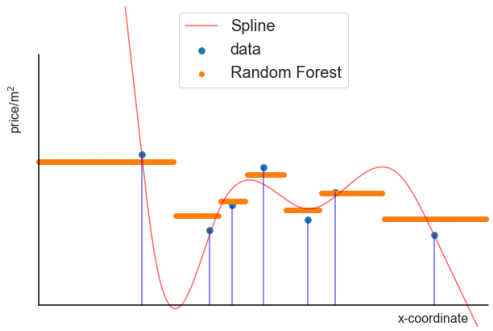
[18]The helper point method is not dependent on the RF algorithm. Other averaging techniques – such as the k-nearest neighbour algorithm or even the global average values – could be used instead.

[19]A single decision tree works as follows: at each step (round, branch of the decision tree), the mechanism tries to find the most significant structural break in the data (i.e., divide the dataset into two as distinct as possible price groups). After multiple rounds, the entire area will be divided into rectangular areas with similar price structures within each sub-area. The average square meter price per area can fill the missing data points that fall within that area.

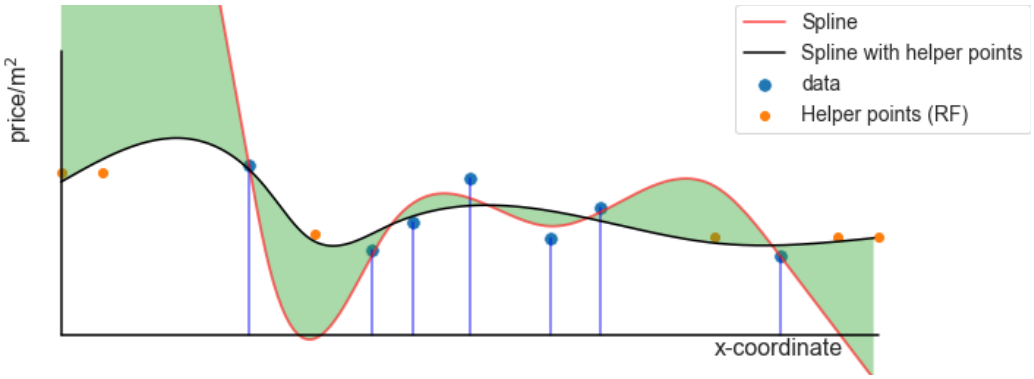
[20]This is not important in this current application. However, we are planning to extend the current paper to include multi-dimensional input.



(a) Spline fitting



(b) Fit a tree-based model (RF) to estimate helper point values



(c) Guided Spline Function (i.e., cubic regression spline with helper points)

Figure 4: Illustration of Guided Spline Function Approach

Note: Figure 4a depicts the square meter price levels of the observations in the dataset as well as the fit of a thin-plate spline surfaced based on these data. Figure 4b illustrates the output of a tree-based ML model providing an average price per square meter level for the entire area covered in the dataset. Where data density is low, we take values of these average price levels and include them as helper points (see the orange dots in Figure 4c) when constructing the spline surface. The shaded area in Figure 4(c) illustrates the difference between the spline with and without helper points.

4.1.1 Step-by-step procedure to construct spline with helper points

Steps 1 to 5 describe our spline construction with helper points. A condensed version of these steps in form of a pseudo algorithm is presented in Appendix B.

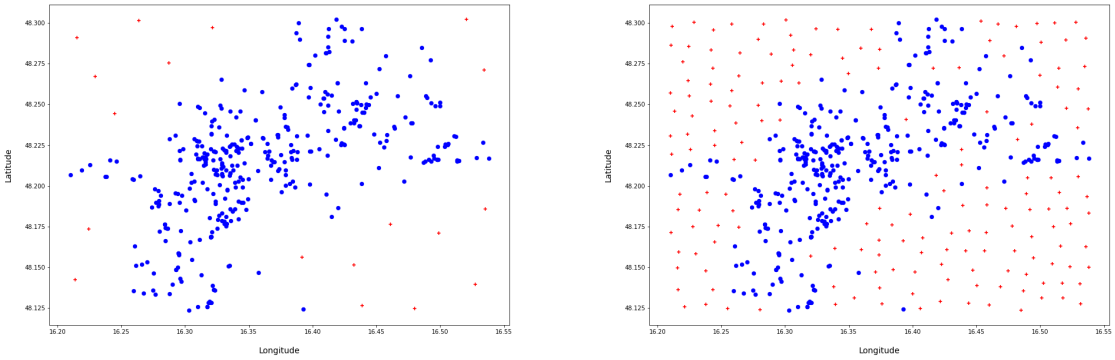
Step 1: Split data into training and test set
We first split the data into geographically connected training and test set areas based on longitude-latitude values.[21] We describe our testing procedure in more detail in subsection 5.2.

Step 2: Place helper points
We place helper points randomly, but inversely to the density of data points by defining a minimum distance between points. In this way, helper points are only placed where no property transaction data is available. Figure 5a and Figure 5b illustrates the output of this process for a minimum distance of 5km

[21] Special care needs to be taken when deciding how to define training and test sets for spatial analysis, as data are often autocorrelated, which leads to overfitting. Using random splits (for train/test sets or Cross-validation) tends to violate the IID assumption because the samples are not statistically independent of each other. As properties close to one another are often very similar – particularly in the case of new-built apartments, a random split may put property x in the training set set and its next-door neighbour in the test set, leading to overstated model accuracy. Grouping the data by area prevents this from happening.

and 1km respectively.

Figure 5: Placement of helper points



(a) Placement with 2.5km min distance (18 helper points)

(b) Placement with 1km min distance (173 helper points)

Step 3: construct helper point values
Next, we estimate a Random Forest model of square-meter prices dependent on only longitude and latitude values. The R package *RandomForest* by Liaw and Wiener (2002) is by far the most popular way to implement basic Random Forest models, easy to use and very robust.[22] We on purpose take a Random Forest algorithm off the shelf without any hyper-parameter tuning.[23] Once we estimated the Random Forest model, we use it to predict square-meter prices at the helper point locations chosen in step 2.

Step 4: Estimate spline surface with original data and with helper points
This step comprises the estimation of the spline surface with helper points using the thin-plate spline methodology of Wood (2003) using the R-package *mcgv* (Wood, 2011).[24] The input data consist of all training data observations plus helper point values.

Step 5: Estimate accuracy of spline surface
Finally, we use the test-set to estimate the goodness of fit of the established spline surface. We describe the testing procedure and provide accuracy results in subsection 5.2. Figure 6a and Figure 6b provide a two-dimensional illustration of the estimated 3-dimensional spline surfaces – once with and once without helper points. Another presentation of the same output – this time as a contour plot of the estimated price levels – is presented in Figure 7a and Figure 7b.

5 Results

5.1 Estimated spline surfaces with and without helper points

Figures 6a and 6b compare the splines calculated without and with helper-points for new-built apartments in Vienna in 2020. The inclusion of the helper points into the dataset leads to a less wiggly price contour. The difference between the two outputs is particularly visible at the boundaries: while the "original" spline shows unnaturally high price estimates for these boundary regions, the spline with helper points

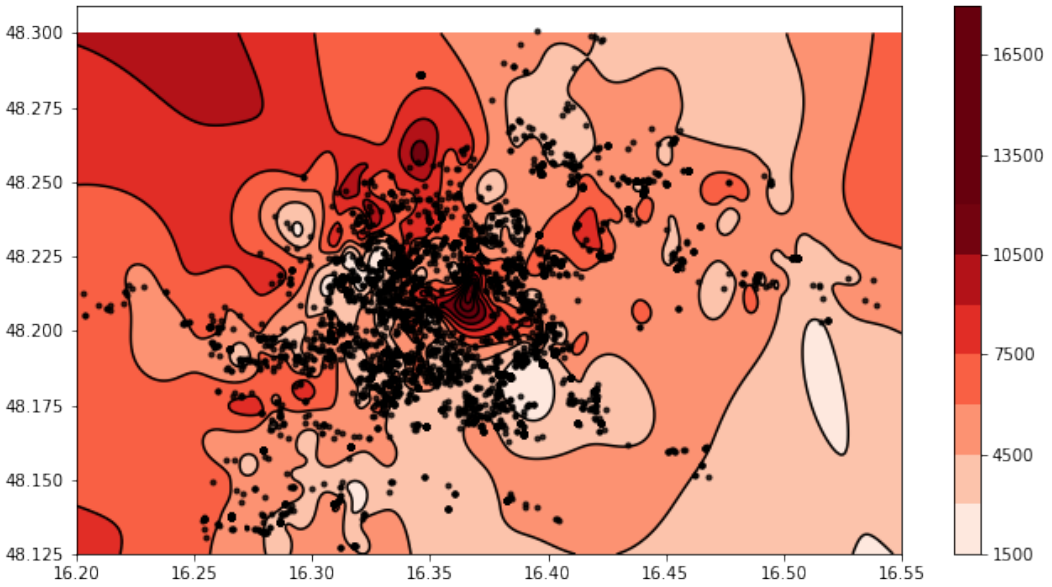
[22] It is also available in Python via the scikit-learn RandomForest implementation.
[23] We take the default settings without CV and n_estimator=500.
[24] We use version 1.8-33 of *mcgv*. See https://cran.r-project.org/web/packages/mgcv/citation.html for more information on the the R-package.

provides much more realistic values for these regions where data are scarce.

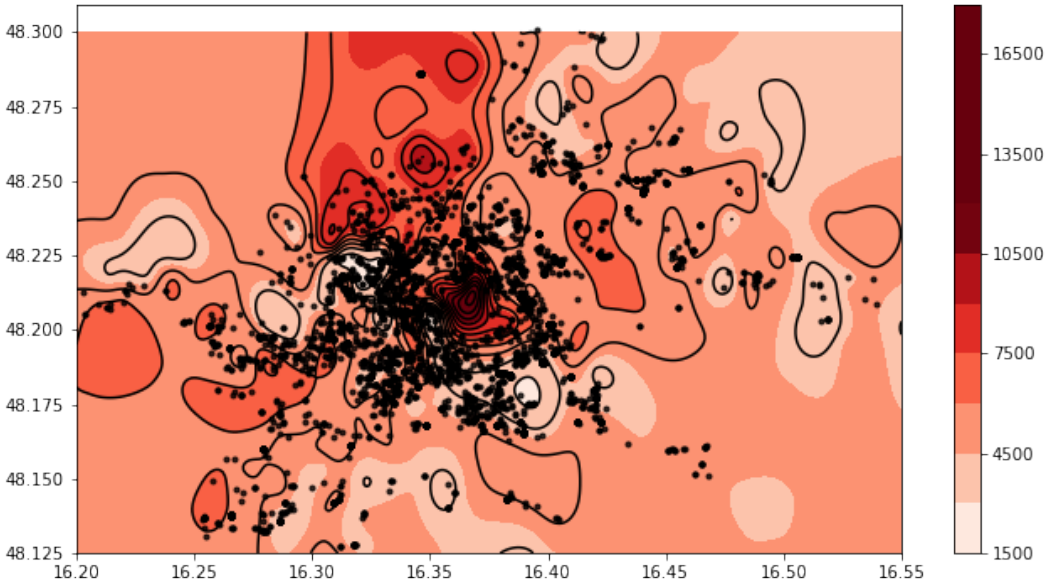
Figure 6: Thin-Plate Spline surface, without (top) and with helper points (bottom)

Note: The figures show estimated square-meter prices for new-built apartments in Vienna in 2020. Each line follows one latitude value. Each line is therefore a slice of a three-dimensional spline surface. The dark line indicates the price level along the latitude coordinates that pass through the city center of Vienna.

An alternative two-dimensional depiction of the two different spline outputs is presented in the contour plots in Figure 7a and Figure 7b. In particular, the estimates at the corners of the map are more in tune with actual price levels when helper points are included. Our helper point method thus prevents price spikes that are not based on actual price observations but rather the results of spline "overshooting".



(a) Spline without helper points



(b) Spline with helper points

Figure 7: Contour plot presentation of spline surfaces

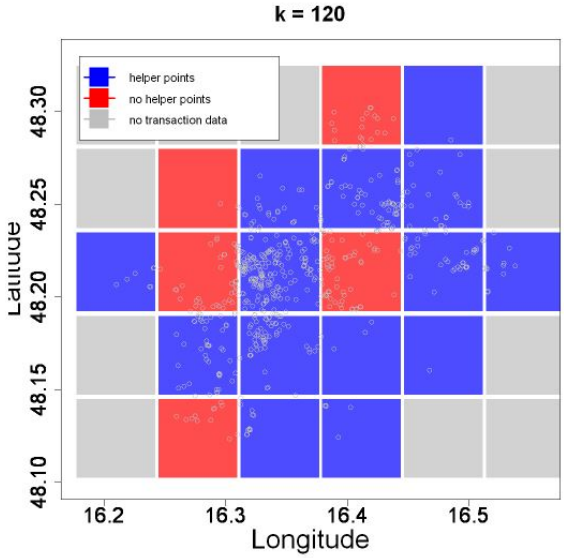
5.2 Testing the accuracy of splines with and without helper points

Figure 6 and Figure 7 clearly demonstrate that helper points can eliminate the wild overshooting behaviour of splines in those areas where no original data exist. The price estimates for these areas become much more realistic when helper points are included. For example, it is easy to see that square-meter

prices in the deepest Vienna woods (located in the top left corner of Figure 7a and Figure 7b should not exceed those of the most expensive city location. Potential investors are much better off when taking Figure 7b as a guide for suitable investment locations than Figure 7a. As no transaction data exist in those areas, these areas do not show up in the cost function of the original spline function and extreme spline values are left "unchecked". Helper points are effective because they introduce a cost to deviating too far from the local average values.

However, we also need to show how splines with helper points behave in areas where transaction data do exist. We do this via a series of out-of sample model estimations, each using a separate "cut-out" sample as testset. We need to use spatial (grouped) train-test splits as our data are spatially auto-correleated data (Valavi et al., 2018; Schratz et al., 2019).[25] Our testing method consists of placing a grid structure over the map of Vienna, which defines equal sized quadratic areas. We take each grid cell in turn as test set and run through the step-by-step procedure described in subsubsection 4.1.1 to estimate a spline with helper points price surface of Vienna.[26] Each time, we also estimate a thin-plate spline surface without helper points for comparison.

The individual results of these separate rounds of model estimation are shown in Figure 8 for a 5km-by-5km grid structure: grid cells for which the model with helper points achieved better accuracy (measured as Mean Absolute Error, MAE) are colored blue, while grid cells for which the traditional spline method achieved better results are colored red. Grid cells without any original data points automatically have zero estimation error independent of which spline method is used. These cells are colored grey. Note, however, that it is exactly these grey areas which show the most evidence of overshooting in the original spline method as illustrated above in Figure 6a or Figure 7a.



(a) Out-of-sample results for 17 helper points (min distance 2.5km)

Figure 8: Blue boxes illustrate the areas where splines with helper points achieve better accuracy (based on MAE), while red boxes illustrate those areas where splines without helper points achieve better results. Grey areas indicate areas without transaction data.

An important insight is that adding just a few helper points to the dataset helps to improve model accuracy; it is not necessary to include a large number of them. In Figure 8a, we placed only 18 helper

[25] When data are independently and identically distributed (IID), we can measure the performance of a model with a random test sample (i.e., a randomly chosen sample of the data observations which is kept separate from the model estimation process). However, with housing data the assumption of IID does not hold as we know that prices are spatially autocorrelated. Thus, using a typical random train-test split would lead to data leakage and thus to overly optimistic model performance estimations (see e.g. Valavi et al. (2018) or Schratz et al. (2019).

[26] The implemented algorithm is shown as a pseudo code in Appendix B.

points (the result of placing each with a minimum distance of 2.5km to any other point). In Figure 8b we placed almost ten times as many helper points (we lowerd the minimum distance between points to 1km). Table 2 summarizes the results by providing the mean estimates over the 30 separate rounds of model estimation. Independently of whether 18 or 173 helper points are placed, out mean out-of-sample accuracy is improved.

Next, we examine how the size of the test set influences model accuracy. For this we vary the grid-cell size between 1km and 10km. **??** illustrates the results with respect to MAE. As before, we use each grid cell in turn as test set. Knot values are kept fixed at 120. We find that splines with helper points improve mean accuracy for all grid sizes, but the effect is not linear.

	MAE		
	with HP	without HP	Difference
17 HPs (min dist. 2.5 km)	1154.187	1429.85	-19.3%
173 HPs (min dist. 1 km)	1063.509	1429.85	-25.6%

Table 1: Mean out-of-sample accuracy (MAE) of spline-with-helperpoint and spline-without-helperpoint estimations

One of the hyper-parameter values that needs to be specified in the *mcgv* package (Wood, 2011) is the number of knots used to estimate the thin-plate spline surface. Wood (2017) recommends setting the number of knots large enough to represent the underlying 'truth'. Choosing too few knots can lead to underfitting. On the other hand, choosing high knot values will increase computation time. Figure 9 illustrates the relationship between knot numbers and model accuracy in more detail. It shows that in-sample accuracy accuracy improves as more knots are included, and that in-sample accuracy is always better for the traditional spline estimation technique. However, figure 9b (not yet included) illustrates that splines with helper points provide better accuracy for out of sample MAE independently of the number of knots used. Also, for out-of-sample accuracy, there is little gain in choosing more than 120 knots, which is the number we have chosen for all spline estimations throughout our paper.

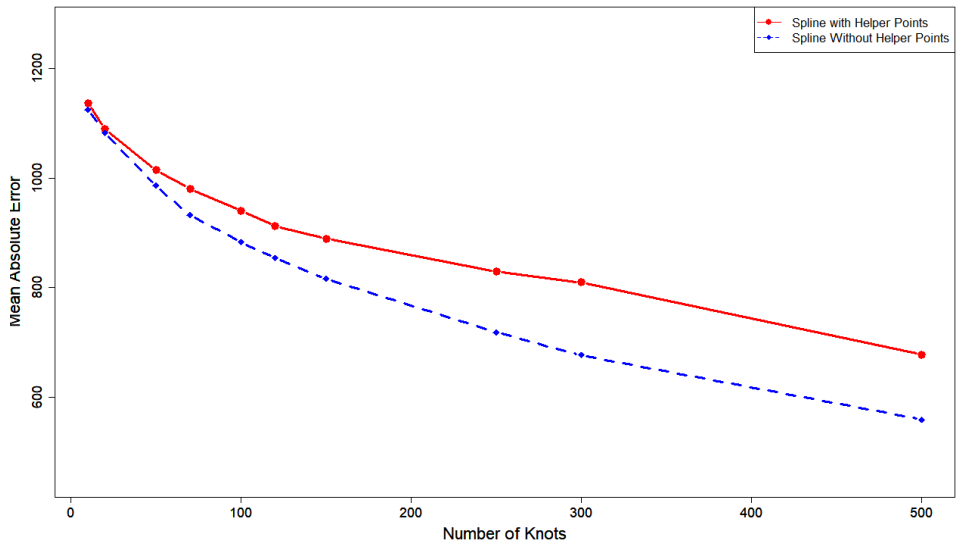


Figure 9: Influence of the number of knots on accuracy

Note: Figure 9 illustrates in-sample MAE achieved by the spline with and without helper points (17 helper points were included). These results are based on in-sample accuracy.

5.3 Additional testing procedure: using 2019 data as testset for 2020 splines

Introducing helper points not only prevents overshooting behavior in areas without data support, it can also improve model prediction in regions where data are less scarce. Data scarcity is not a zero-one phenomenon. Rather, there is a scarcity spectrum. In regions in which data are plentiful, the spline-with-helper-points method will not add any helper points. Hence it will not have much impact on the predictive performance of the spline surface in such areas. However, as scarcity increases, so do the number of helper points. It is thus in these regions with higher scarcity that model predictive performance may be most improved. Helper points thus not only prevent excessive spline overshooting in areas without data input, they can also prevent overfitting in data-scarce areas.

We investigate this issue by fitting a spline surface on 2020 data and then checking how well it predicts the prices of properties sold in 2019 (with these prices inflated by the average price rise between 2019 and 2020).

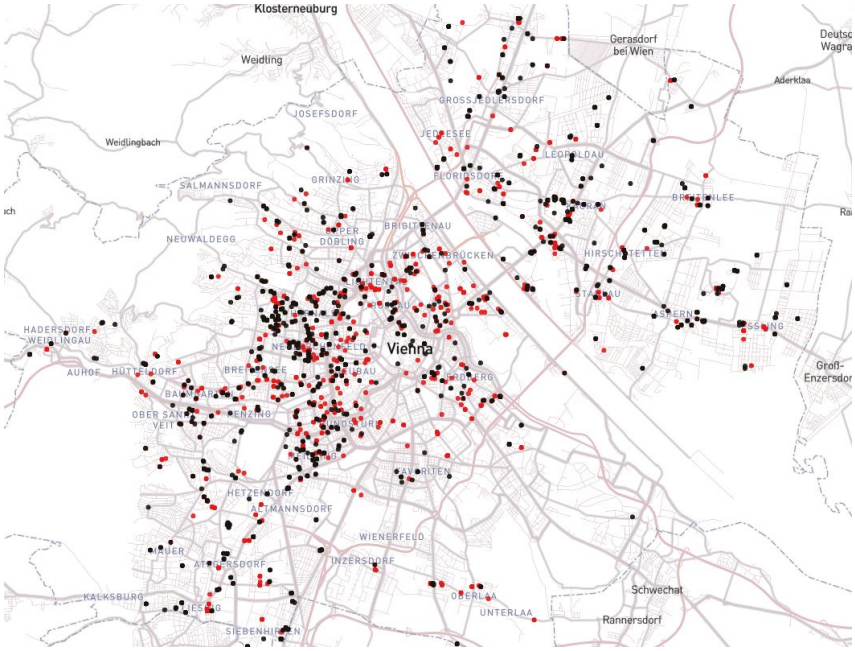


Figure 10: Transactions for new-built apartments in Vienna, 2019 and 2020
Note: Black dots indicate 2020 apartment transactions, while red dots indicate 2019 transactions.

[Insert Figure with grid structure here]

	MAE		
	with HP	without HP	Difference
17 HPs (min dist 2.5km)	968.7902	984.7067	-1.6%
173 HPs (min dist 1km)	962.93	984.7067	-2.2%

Table 2: Mean out-of-sample accuracy (MAE) of spline-with-helperpoint and spline-without-helperpoint estimations

NOTE: SECTION STILL WORK IN PROGRESS

6 Conclusion

We pursued three aims in this paper: First, we showed how to construct geospatial price surfaces using penalized regression splines with house price data. While there are many excellent technical books and papers on splines, we found almost no discussion on how to construct multilateral regression splines with geospatial data. We focused primarily on those parts of the spline literature relevant to real-estate applications.

Our second aim was to draw attention to a potential problem with applying penalized regression spline techniques in real-estate settings. Unlike other non-parametric smoothing techniques (such as locally weighted regression or kernel methods), spline models can react differently in different parts of the data distribution. However, splines are poor at extrapolating into areas with little or no data support. This poses a problem for spline applications with real estate data, as property sales are rarely evenly distributed throughout the landscape. Typically there will be multiple locations within the study area for which few or no transactions exist. These data gap areas tend to be problematic when constructing splines, as spline functions tend to overshoot in areas with little data support. This is a problem when using splines to depict cities' price gradients.

Our paper's third and foremost aim was to introduce a straightforward and powerful method to overcome this overshooting problem of splines by placing helper points in areas without data support. We used a dataset of transaction prices for new-built apartments in Vienna in 2020. New-built apartments are spread unevenly throughout the city area, and there are many gaps in the data support. We show that our method prevents overshooting and dramatically improves out-of-sample accuracy even when only few helper points are introduced.

References

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The economics of density: Evidence from the berlin wall. *Econometrica*, 83(6):2127–2189.

Allen, T. and Arkolakis, C. (2014). Trade and the Topography of the Spatial Economy *. *The Quarterly Journal of Economics*, 129(3):1085–1140.

Bao, H. X. and Wan, A. T. (2004). On the use of spline smoothing in estimating hedonic housing price models: Empirical evidence using hong kong data. *Real Estate Economics*, 32(3):487–507. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1080-8620.2004.00100.x.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brumback, B. A., Ruppert, D., and Wand, M. P. (1999). Comment on shively, kohn and wood.

Clapp, J. M. (2004). A semiparametric method for estimating local house price indices. *Real Estate Economics*, 32(1):127–160. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1080-8620.2004.00086.x.

Colwell, P. (1998). A primer on piecewise parabolic multiple regression analysis via estimations of chicago cbd land prices. *The Journal of Real Estate Finance and Economics*, 17:87–97.

Craig, S. G. and Ng, P. T. (2001). Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. *Journal of Urban Economics*, 49(1):100–120. https://www.sciencedirect.com/science/article/pii/S0094119000921867.

Crainiceanu, C. M., Ruppert, D., and Wand, M. P. (2005). Bayesian analysis for penalized spline regression using winbugs. *Journal of Statistical Software*, 14(14):1–24. https://www.jstatsoft.org/index.php/jss/article/view/v014i14.

d. Boor, C. (1978). *A Practical Guide to Splines*. Springer Verlag.

Davis, P. (1996). B-splines and geometric design. *SIAM News*, 29.

Diewert, E. and Shimizu, C. (2019). Residential Property Price Indexes: Spatial Coordinates versus Neighbourhood Dummy Variables. Microeconomics.ca working papers erwin_diewert-2019-11, Vancouver School of Economics. https://ideas.repec.org/p/ubc/pmicro/erwin_diewert-2019-11.html.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in sobolev spaces. In Schempp, W. and Zeller, K., editors, *Constructive Theory of Functions of Several Variables*, pages 85–100, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89 – 121. https://doi.org/10.1214/ss/1038425655.

Gibbons, S. and Machin, S. J. (2003). Valuing english primary schools. *Journal of Urban Economics*, 53:197–219.

Greiner, A. (2009). Estimating penalized spline regressions: theory and application to economics. *Applied Economics Letters*, 16(18):1831–1835. https://doi.org/10.1080/13504850701719629.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1(3):297 – 310. https://doi.org/10.1214/ss/1177013604.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Wiley Online Library. https://www.bibsonomy.org/bibtex/2cecb9f8ade557a3ef56727ee8605e57d/lautenschlager.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition. http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Hill, R. J. and Scholz, M. (2018). Can geospatial data improve house price indexes? a hedonic imputation approach with splines. *Review of Income and Wealth*, 64(4):737–756. https://onlinelibrary.wiley.com/doi/abs/10.1111/roiw.12303.

Kholodilin, K. A., Limonov, L. E., and Waltl, S. R. (2021). Housing rent dynamics and rent regulation in st. petersburg (1880–1917). *Explorations in Economic History*, 81:101398. https://www.sciencedirect.com/science/article/pii/S0014498321000164.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Liu, B., Mavrin, B., Niu, D., and Kong, L. (2016). House price modeling over heterogeneous regions with hierarchical spatial functional analysis. *2016 IEEE 16th International Conference on Data Mining (ICDM)*. http://dx.doi.org/10.1109/ICDM.2016.0134.

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics Data Analysis*, 28(2):193–209. https://www.sciencedirect.com/science/article/pii/S0167947398000334.

McMillen, D. (1996). One hundred fifty years of land values in chicago: A nonparametric approach. *Journal of Urban Economics*, 40(1):100–124. https://doi.org/10.1006/juec.1996.0025.

McMillen, D. P. (2001). Nonparametric employment subcenter identification. *Journal of Urban Economics*, 50(3):448–473. https://www.sciencedirect.com/science/article/pii/S0094119001922284.

Melser, D. and Hill, R. J. (2019). Residential Real Estate, Risk, Return and Diversification: Some Empirical Evidence. *The Journal of Real Estate Finance and Economics*, 59(1):111–146. https://ideas.repec.org/a/kap/jrefec/v59y2019i1d10.1007_s11146-018-9668-x.html.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Schoenberg, I. J. (1946a). Contributions to the problem of approximation of equidistant data by analytic functions: Part a.—on the problem of smoothing or graduation. a first class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(1):45–99. http://www.jstor.org/stable/43633538.

Schoenberg, I. J. (1946b). Contributions to the problem of approximation of equidistant data by analytic functions: Part b—on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(2):112–141. http://www.jstor.org/stable/43633544.

Schoenberg, I. J. (1964). Spline interpolation and best quadrature formulae. *Bulletin of the American Mathematical Society*, 70(1):143 – 148. https://doi.org/.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., and Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, 406:109–120.

Shalizi, C. R. (2013). *Advanced Data Analysis from an Elementary Point of View*. Cambridge University Press. http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

Thorsnes, P. and McMillen, D. P. (1998). Land value and parcel size: A semiparametric analysis. *The Journal of Real Estate Finance and Economics*, 17:233–244. https://doi.org/10.1023/A:1007772223239.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2018). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *bioRxiv*, 357798.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(3):364–372. http://www.jstor.org/stable/2984701.

Wahba, G. (1980). Ill posed problems: Numerical and statistical methods for mildly, moderately and severely ill posed problems with noisy data.

Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, Philadelphia. https://doi.org/10.1137/1.9781611970128.

Wahba, G. and Wang, Y. (2017). *Spline Functions: Overview*, pages 1–19. American Cancer Society. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat02388.pub2.

Wahba, G. and Wold, S. (1975). Periodic splines for spectral density estimation: the use of cross validation for determining the degree of smoothing. *Communications in Statistics*, 4:125–141.

Wakefield, J. (2013). *Spline and Kernel Methods*, pages 547–595. Springer New York, New York, NY.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00374.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N., Bravington, M. V., and Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(5):931–955.

A Appendix: Decision-trees

Figure A1 illustrates the output of a decision tree algorithm. It successively splits an area into relatively homogenous regional sub-groups subject to a predefined loss metric (e.g. RMSE). A Random Forest consists of many such decicion trees.

Example Decision Tree

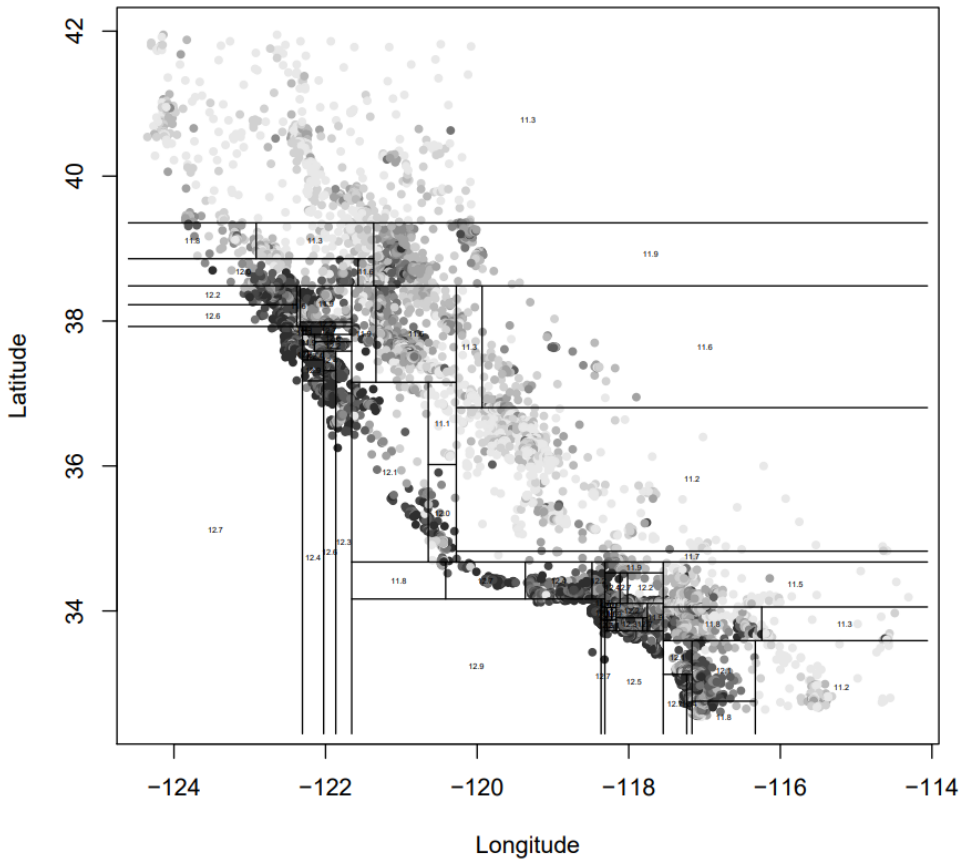


Figure A1: Illustration of a decision-tree based data segmentation
The algorithm brakes up the area into rectangular groups of similar price level

B Appendix: Pseudo algorithm for the implementation of the spline estimation with helper points

Algorithm 1 Geo-spatial cross validation with test size between 3-8%

$testSetSize \leftarrow 0$
$upperLimit \leftarrow 0.08 * nrow(data)$
$lowerLimit \leftarrow 0.03 * nrow(data)$
$maxDensity \leftarrow 2$, number of observations within 1km x 1km
$nHelperPoints \leftarrow 100$
for $i = 1, 2, \ldots, 100$ **do**
    while $testSetSize > upperLimit$ **or** $testSetSize < lowerLimit$ **do**
        create two random uniform points from Cartesian product of latitude and longitude
        combine these points to create rectangle
        $testSet \leftarrow$ all observations within rectangle
        $testSetSize \leftarrow nrow(testSet)$
    end while
    $trainSet \leftarrow$ all observations not in rectangle
    $dtModel \leftarrow$ train and tune decision tree model on $trainSet$
    $r(latitude, longitude) \leftarrow$ draw random point from Cartesian product of latitude and longitude
    $d(r) \leftarrow$ calculate density of 1km x 1km area around random point
    $helperPoints \leftarrow emptylist$
    $helperPointSize \leftarrow 0$
    while helperPointSize < nHelperPoints **do**
        if $d(r)$ < maxDensity **then**
            helperPoint ← append(dtModel.predict(r), r)
        end if
        helperPointSize ← nrow(helperPoints)
    end while
    splineNoHelperPoints ← train spline on latitude, longitude to predict square meter price on training set
    splineHelperPoints ← train spline on latitude, longitude to predict square meter price on training set and predicted square meter price on helperPoints
    calculate metric on testset for splineNoHelperPoints and splineNoHelperPoints
end for