Sandra Peterl BSc

Smooth Spatial and Time Effect Models to Forecast House Prices in Sydney

MASTER'S THESIS

to achieve the university degree of
Diplom-Ingenieurin

Master's degree programme: Financial and Actuarial Mathematics

submitted to

Graz University of Technology

Supervisor
Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig Friedl

Institute for Statistics

Graz, June 2017

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date

Signature

DANKSAGUNG

An dieser Stelle möchte ich mich bei denjenigen bedanken, die durch ihre fachliche und persönliche Unterstützung zum Gelingen dieser Masterarbeit beigetragen haben.

Mein größter Dank gebührt meinem Betreuer und Professor Dipl.-Ing. Dr.techn. Herwig Friedl, der mich stets bei der Entwicklung meiner Modelle unterstützt und meine Fragen jederzeit geduldig und hilfsbereit beantwortet hat. Ich bedanke mich auch für die vielen hilfreichen Anregungen und die konstruktive Kritik bei der Erstellung dieser Arbeit. Des Weiteren bedanke ich mich bei Dr. Robert Hill vom Institut der Volkswirtschaftslehre der Universität Graz für die erhaltenen Fördermittel zur Erstellung dieser Arbeit und die zur Verfügung gestellten Datensätze. Mit seiner Unterstützung war es mir möglich, auch wirtschaftliche Aspekte in meine Modelle und somit auch in meine Arbeit einfließen zu lassen. Ich bedanke mich auch bei Sofie Waltl und Angelika Renz für das Korrekturlesen dieser Arbeit.

Mein besonderer Dank gilt meiner Familie, insbesondere meinen Eltern, die mir mein Studium ermöglicht und mich in all meinen Entscheidungen unterstützt haben. Herzlich bedanken möchte ich mich auch bei meinen Freunden und Studienkollegen, die mich immer wieder ermutigt, mir mit vielen nützlichen Tipps die Arbeit erleichtert und mir eine schöne Studienzeit ermöglicht haben.

ABSTRACT

Regression is the most popular statistical methodology for analysing empirical problems in economics, life sciences and social sciences. Thus, there exists a large diversity of models from the simple linear regression model to the more complex nonparametric regression model. While the parametric model assumes the mean response variable to be a linear combination of the explanatory variables, the nonparametric model uses flexible techniques which enable the automatic and data-driven estimation of the nonlinear effects. In this thesis, the parametric and the nonparametric regression models are introduced, including detailed theory about the estimation procedures. The subclass of generalized additive models is then used to model renting and selling prices for houses located in Sydney, since this class of models allows the inclusion of spatial effects. Therefor two datasets are used which, additionally to the renting or selling price, contain information about the houses' characteristics. The derived house price model is further used to forecast the renting and selling prices of houses in Sydney for the nearest future. In contrast to most of the literature on forecasting house prices, which only model the average trend of the price for the future, the model derived in this thesis allows to forecast the price of any specific house with particular properties and location at a time of interest. The derived model as well as the forecast turn out to provide a good fit to the data.

ZUSAMMENFASSUNG

Regression ist die populärste statistische Methodik zur Analyse von empirischen Problemen in der Wirtschaft, den Lebenswissenschaften und den Sozialwissenschaften. Aus diesem Grund existiert eine breite Vielfalt an Modellen vom einfachen linearen bis hin zum komplexeren nicht-parametrischen Regressionsmodell. Während das parametrische Regressionsmodell die zu erwartende Response-Variable als Linearkombination der erklärenden Variablen darstellt, verwendet das nicht-parametrische Regressionsmodell flexible Techniken, welche die automatische und datengesteuerte Schätzung der nicht-linearen Effekte erlauben. In dieser Arbeit werden die parametrischen und die nicht-parametrischen Regressionsmodelle und die detaillierte Theorie zu den zugehörigen Schätzungsverfahren vorgestellt. Die Unterklasse der generalisierten additiven Modelle wird schließlich zur Modellierung von Miet- und Verkaufspreisen von Häusern in Sydney verwendet, da diese Modellklasse das Inkludieren geografischer Effekte erlaubt. Hierfür werden zwei Datensätze eingesetzt, welche zusätzlich zu den Miet- bzw. Verkaufspreisen auch Eigenschaften der Häuser beinhalten. Das hergeleitete Hauspreismodell wird des Weiteren zur Vorhersage von Miet- und Verkaufspreisen für Häuser in Sydney in naher Zukunft verwendet. Im Gegensatz zum Großteil der Literatur zu Vorhersagemodellen für Hauspreise, welche nur den durchschnittlichen Trend zukünftiger Preise modellieren, prognostiziert das in dieser Arbeit hergeleitete Modell den Preis eines speziellen Hauses mit gegebenen Eigenschaften und Standort zu einem interessierenden Zeitpunkt in der Zukunft. Es stellt sich heraus, dass das gefundene Modell und auch die Vorhersage die Daten gut abbilden.

Contents

List of Variables

n	number of observations
p	number of parameters in the linear model
$\boldsymbol{y}$	response vector
$\boldsymbol{\mu}$	mean of the response vector
$\boldsymbol{x}_i$	covariates in the linear model
$\boldsymbol{X}$	design matrix in the linear model
$\boldsymbol{\beta}$	vector of regression coefficients in the linear model
$\boldsymbol{\epsilon}$	vector of error terms
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\boldsymbol{H}$	hat matrix in the linear model
$\boldsymbol{\eta}$	linear predictor in the linear model
$\boldsymbol{\theta}$	canonical parameter of the exponential family
ϕ	dispersion parameter of the exponential family
$a(\cdot)$, $b(\cdot)$, $c(\cdot)$	specific functions in the definition of the exponential family
$V(\cdot)$	variance function of the exponential family
$g(\cdot)$	link function in the generalized linear model
$\boldsymbol{z}_i$	covariates in the nonparametric model
$f(\cdot)$	deterministic function of the covariate
$B_j(z)$	basis functions in the nonparametric model
d	number of basis functions
κ_i	knots used for the splines
m	number of knots used
$\boldsymbol{\gamma}$	vector of regression coefficients in the nonparametric model
l	degree of the spline
$\boldsymbol{Z}$	design matrix in the nonparametric model
$[a, b]$	domain of the regression spline
$B_j^l(z)$	B-spline basis functions of degree l
$\boldsymbol{\lambda}$	vector of smoothing parameters
w_i	function values at the knots ($f(\kappa_i) = w_i$)
$\boldsymbol{K}$	penalty matrix in the penalized least-squares procedure
$\boldsymbol{S}(\boldsymbol{\lambda})$	smoother matrix in the nonparametric model
$\boldsymbol{B}$	square root matrix of $\boldsymbol{K}$ such that $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{K}$
df	degrees of freedom
$\boldsymbol{b}_\lambda$	bias of the function f
y_i^*	new observation at z_i
ϵ_i^*	error term corresponding to y_i^*
$f^{(-i)}(z)$	estimated mean obtained when removing the ith observation
d_1, d_2	number of basis functions of the first and the second covariate, respectively, in bivariate smoothing
$B_j^{(1)}(z_1)$, $B_r^{(2)}(z_2)$	univariate basis functions of z_1 and z_2, respectively
$B_{jr}(z_1, z_2)$	basis functions in the bivariate nonparametric model
$\boldsymbol{K}_j$	penalty matrices in the penalized least-squares procedure in

	multivariate models
$\boldsymbol{D}_1$, $\boldsymbol{D}_2$	difference matrices in the bivariate nonparametric model
$J(f)$	penalty functional in the case of bivariate thin plate splines
$\eta(\cdot)$, $\phi_j(\cdot)$	specific functions in the definition of thin plate splines
$\boldsymbol{Z}_d$	best rank d approximation to $\boldsymbol{Z}$ in the sense of the spectral norm $\Vert\boldsymbol{Z} - \boldsymbol{Z}_d\Vert$
$\boldsymbol{\Gamma}$	orthonormal matrix of eigenvectors in the Choleski decomposition
$\boldsymbol{\Gamma}_d$	submatrix of $\boldsymbol{\Gamma}$ associated with the d largest eigenvalues
$\boldsymbol{\Omega}$	matrix consisting of the corresponding eigenvalues (of $\boldsymbol{\Gamma}$) in descending order
$\boldsymbol{\Omega}_d$	submatrix of $\boldsymbol{\Omega}$ associated with the d largest eigenvalues
$J_d^q(f)$	penalty functional in the case of q-dimensional thin plate splines in the d-dimensional space
Ω	region on the $x - y$ plane in the definition of SOAP film smooths
$J_\Omega(f)$	penalty functional in the case of SOAP film smooths
B	collection of closed loops on the $x - y$ plane in the definition of SOAP film smooths
$\rho(\cdot,\cdot)$, $\rho_i(\cdot,\cdot)$	specific functions in the definition of SOAP film smooths
$h(\cdot,\cdot)$, $g_i(\cdot,\cdot)$	specific functions in the representation of a SOAP film smoother
$f_b(r)$	boundary function in the case of SOAP film smooths when the value of the function on the boundary is not known
$J_b(f_b)$	penalty functional in the case of SOAP film smooths when the value of the function on the boundary is not known
J	number of basis functions corresponding to the term associated with the unknown function values at the boundary in the case of SOAP film smooths
$\nu_j(\cdot)$	basis functions corresponding to the term associated with the unknown function values at the boundary in the case of SOAP film smooths
$\boldsymbol{A}$	design matrix corresponding to the term associated with the unknown function values at the boundary in the case of SOAP film smooths
$\boldsymbol{S}_b$	smoother matrix corresponding to the term associated with the unknown function values at the boundary in the case of SOAP film smooths
λ_b	smoothing parameter corresponding to the term associated with the unknown function values at the boundary in the case of SOAP film smooths
$\boldsymbol{\alpha}$	vector of regression coefficients corresponding to the term associated with the unknown function values at the boundary in the case of SOAP film smooths
q	number of covariates in the higher-dimensional model
$\boldsymbol{\eta}^{lin}$	linear predictor in the structured additive model
$\boldsymbol{\eta}^{add}$	additive predictor in the structured additive model

k	number of covariates of the linear part in a structured additive model
$\boldsymbol{v}_i$	covariates in the structured additive model
$\boldsymbol{V}_j$	design matrices in the structured additive model
$\boldsymbol{Z}_{geo}$	design matrix of the geoadditive term in the additive model
$\boldsymbol{Z}_{1\|2}$	design matrix of the interaction term in the additive model
$\boldsymbol{Z}_{int}$	design matrix of the interaction term in the varying coefficient model
$\boldsymbol{\delta}$	vector of regression coefficients in the generalized additive model
r_P	Pearson residual
r_A	Anscombe residual
r_D	Deviance residual
$A(\cdot)$	function in the definition of Anscombe residuals

CHAPTER 1

Introduction

The housing market plays a decisive role in everyday life. The standard of living of every person is directly influenced by house rents and house prices, since everyone needs a place to live. Thus, house price developments have an essential impact on financial stability and real economic activity. The financial crisis in the year 2007 was for example provoked by the sudden and immense downturn in U.S. house prices in the year 2006. This crisis demonstrated the need of forecasting the trend of house prices for the nearest future for financial stability. Furthermore, the prediction of house price developments is also important in terms of mortgage default and property taxes, as well as investment and policy decisions.

For all these reasons, the goal of this thesis is to derive a statistical model which allows to forecast the renting and selling prices of houses located in Sydney. We use two datasets containing information about the rented or sold houses regarding the houses' characteristics as well as the geographic position and the day of advertisement or respectively the day of the sale. This thesis is divided into two parts, namely the theoretical and the practical part.

The theoretical part starts with the simplest class of regression models, namely the Parametric Regression Models in Chapter 2. These models assume the mean response variable to be a linear combination of the explanatory variables. The big advantage of these models is their simplicity. However, the assumption of a purely linear model is not suitable for many practical applications. Therefore, the more complex class of Nonparametric Regression Models is introduced in Chapter 3. The simplest subclass of these models are the Univariate Smoothing Models, which explain the expected response variable through a deterministic function f depending on one explanatory variable. This is done by defining the space of functions of which f is an element of by choosing a suitable basis. The following sections in Chapter 3 present the most important basis functions and their use. In Section 3.4 the univariate models are extended yielding the Bivariate Smoothing Models, where the response variable depends on two different covariates. This concept is further generalized to the class of Multivariate Smoothing Models. Additive Models, defined in Section 3.6, allow spatial effects by including geographic locations as covariates, which is important in terms of modeling house prices. Furthermore, it is possible to include different ways of interactions between the explanatory variables. All these generalizations lead to the class of Generalized Additive Models in Section 3.7. These models describe a function of the response variable through an additive model and the response variable is assumed to follow a distribution from a member of the linear exponential family, which is a specific class of distributions. The last section of Chapter 3 deals with the problem of model choice and diagnostics.

All the theory of Chapter 2 and 3 of this thesis is used to derive a model for the renting and selling prices of houses located in Sydney. The used datasets of rented and sold houses between the years 2008 and 2015 are described and analysed in detail in Chapter 4 to get a more detailed insight into the behaviour of the house price. The obtained information is then used in Chapter 5 to derive a hedonic price model for rented and sold houses in Sydney. All the calculations are done with the `mgcv` package of the program `R`, which is introduced in the first section of this chapter. We first try a model with normally distributed prices and go on with the assumption of renting and selling prices which follow a gamma distribution. We then model the house rents and the house prices seperately. In the last subsection we try a joint model for both datasets. The last Chapter 6 uses the derived joint model to forecast the renting and selling prices of Sydney for the calendar year 2016. Finally, the prediction of the price-rent ratio in the year 2016 is done, which is the ratio of property prices relative to rents. This ratio increased from about 25 in 2014 to 32 years of rent in 2016. This means that in 2016 the price of the house equals about 32 years of rent, whereas in the year 2014 one only had to pay about 25 yearly rents to cover the house price.

CHAPTER 2

Parametric Regression

2.1 Classical Linear Model

The *Classical Linear Model (LM)* is defined by McCullagh and Nelder (1989) as

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X\beta},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T}$ is a vector of observations, whose components are independently distributed with means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\mathsf{T}$. The $n \times p$ design matrix $\boldsymbol{X}$ consists of the explanatory variables or covariates x_{ij}, where each row $\boldsymbol{x}_i = (x_{i0}, \ldots, x_{i,p-1})$ of $\boldsymbol{X}$ refers to a different observation i and each column to a different covariate j. The $p \times 1$ vector of parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\mathsf{T}$ consists of the unknown values and has to be estimated from the data. Hence, for the mean of $\boldsymbol{y}$ it follows that

$$\mathrm{E}(y_i) = \mu_i = \sum_{j=0}^{p-1} x_{ij}\beta_j \qquad \text{for } i = 1, \ldots, n.$$

The first column of $\boldsymbol{X}$ is equal to one, which means that $x_{i0} = 1$ for $i = 1, \ldots, n$. Thus, the model takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i \qquad \text{for } i = 1, \ldots, n,$$

where the errors $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed with zero mean and variance σ^2. In most of the cases, the error terms are assumed to follow a normal distribution. In this special case, the response vector $\boldsymbol{y}$ is normally distributed with mean $\boldsymbol{X\beta}$ and variance $\sigma^2 \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix.

The estimator $\hat{\boldsymbol{\beta}}$ in the classical linear model can be found by minimizing the least-squares criterion

$$SSE(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X\beta}). \tag{2.1.1}$$

The value of $\boldsymbol{\beta}$ that minimizes (2.1.1) is found to be

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.$$

The vector of the fitted values can therefore be written as

$$\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ is a symmetric and idempotent matrix, called the *hat matrix.*

The classical linear model has a number of drawbacks, as described in detail in Fahrmeir, Kneib, Lang, and Marx (2013). In most of the cases, the error terms ϵ_i are assumed to be normally distributed, so the linear model is well suited for regression analyses when the response variable is continuous and at least approximately normally distributed. In addition, the mean of the response is assumed to be a linear combination of covariates. In some cases, the assumption of an approximately normally distributed response variable or the linear relationship can be ensured by using an appropriate transformation. Nonetheless, there are many applications where the response is not a continuous variable, but rather binary, categorical or a count variable. Furthermore, there are cases where the distribution of the continuous response variable is skewed, then it is often advantageous to use, for example, the gamma distribution. In the linear model, the error variances are assumed to be homoscedastic, which means that $\mathrm{E}(y_i) = \sigma^2$ for all $i = 1, \ldots, n$, thus all observations fluctuate with a constant variability around the regression surface. Additional to the homoscedastic variances, the errors are assumed to be uncorrelated, which means that $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$. There are some cases, where these assumptions are not realistic, especially for time series or longitudinal data.

Nevertheless, the classical linear model is very simple in its structure and well interpretable for its user. The estimate can be simply computed without much effort. After all, the linear regression model is very useful, if the dependence of the mean of y_i on $\boldsymbol{x}_i$ is linear or almost linear for $i = 1, \ldots, n$, since it can be simply computed and provides a concise description of the data. However, if the dependence of y_i on $\boldsymbol{x}_i$ is far from linear, modeling the mean linearly in the parameters would not always provide a good fit. Therefore, many extensions of the classical linear model exist.

2.2 Generalized Linear Model

The *Generalized Linear Model (GLM)* is an extension of the classical linear model, which can deal with some of the disadvantages of the classical linear model. According to McCullagh and Nelder (1989), the GLM consists of the following components:

- The *random component*: the response variable y follows the distribution of a member of the one parameter exponential family with $\mathrm{E}(y) = \mu$. The density or probability mass function of this exponential family is defined as

  $$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

  for some specific functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$. The parameter θ is called the *canonical parameter*. If ϕ is known, this is an exponential family model with canonical parameter θ. The parameter ϕ is called the *dispersion parameter*. The normal, the binomial, the poisson and the gamma distribution are all members of this exponential family. It can be shown that

  $$\begin{aligned} \mathrm{E}(y) &= b'(\theta) = \mu, \\ \mathrm{var}(y) &= a(\phi)b''(\theta) = a(\phi)V(\mu) \end{aligned}$$

  holds for the response variable y with a distribution from the exponential family. The variance of y is therefore the product of the function $V(\mu)$, the *variance function*, which depends on the canonical parameter only, and $a(\phi)$, which is independent of θ and depends only on ϕ.

- The *systematic component*: the covariates $\boldsymbol{x} = (x_0, \ldots, x_{p-1})^\mathsf{T}$ build the linear predictor η given by

  $$\eta = \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}.$$

- The *link funktion*: the link function $g(\cdot)$ describes the relation between the random and the systematic component as

  $$g(\mu) = \eta$$

  where $g(\cdot)$ is assumed to be a monotonic twice differentiable continuous function. Therefore, in the GLM, a function of the mean of the response variable is modeled linearly in some parameters. Some popular link functions for binary responses are

  - the *logit* link with $\eta = \log(\frac{\mu}{1-\mu})$,
  - the *probit* link with $\eta = \Phi^{-1}(\mu)$, where $\Phi(\cdot)$ is the normal cumulative distribution function and
  - the *complementary log-log* with $\eta = \log(-\log(1-\mu))$.

In the GLM, the parameter $\boldsymbol{\beta}$ is estimated by maximizing the likelihood criterion. This is done by calculating the first derivative of the log-likelihood function with respect to $\boldsymbol{\beta}$ and setting it to zero. Since the resulting system of equations is not linear in $\boldsymbol{\beta}$, it can be solved with the Newton-Raphson method. This procedure will not be discussed in this thesis, but can be read in McCullagh and Nelder (1989). Another method for estimating $\boldsymbol{\beta}$ is the *Iteratively Re-Weighted Least-Squares (IRLS)* procedure. This procedure is described in Section 3.7.1 in case of penalized likelihood estimation and works analogously for the GLM, when leaving out the penalty term.

In summary, a specific GLM is completely determined by the member of the exponential family, the choice of the link or response function and the definition and selection of the covariates. In the classical linear model, the mean and the linear predictor are identical. This special link function is called the *identity link* and both η and μ can take any value on the real line. When considering a poisson distribution, for example, it must be assured that $\mu > 0$, thus the log link $\eta = \log \mu$ may be more attractive than the identity link because η might be negative while μ must not be. The link function is also reasonable for the binomial distribution where $0 < \mu < 1$, thus the link function should map the interval $(0, 1)$ on the whole real line, which can be achieved for example with the probit link.

Despite the simple and well interpretable characteristics of a GLM, there are many practical applications where a purely linear model is not sufficient. This insufficiency could result from uncertainty about the specific form of an effect of the covariate on the response. The next chapter will therefore focus on flexible regression techniques for nonlinear effects.

CHAPTER 3

Nonparametric Regression

The main tool in nonparametric regression is the so-called *smoother.* Hastie and Tibshirani (1990) defined a smoother as a tool for summarising the trend of a response variable y as a function of a predictor variable z. The estimated trend is then less variable than y itself, which is the reason for the name *smoother.* A smoother is nonparametric because it does not assume a rigid form for the dependence of y on z. The estimate produced by a smoother is called a *smooth.* There are two main uses for smoothers. Firstly, a smoother can show the trend between a response variable y and a predictor variable z and thus has a descriptive use. The other main application is the estimation of the dependency of the mean of y on the predictor, as can be seen in the following sections.

3.1 Univariate Smoothing

The simplest form of a smoother is the flexible modeling of the effect of one continuous covariate on a continuous dependent variable and is called *Univariate Smoothing* or *Scatterplot Smoothing*, since the data can be best visualized in a scatter plot and the intention is to find a smooth function representing the effect of the covariate. The data is assumed to be given in the form (y_i, z_i), $i = 1, \ldots, n$, where the y_i are observations of the response variable and the z_i represent the corresponding values of the continuous covariate. The standard univariate nonparametric regression model, as described in Fahrmeir et al. (2013), is given by

$$y_i = f(z_i) + \epsilon_i \qquad \text{for } i = 1, \ldots, n. \tag{3.1.1}$$

The response variable can therefore be explained through a deterministic function of the covariate plus an additive error term. To simplify the estimation problem, qualitative constraints concerning the smoothness and therefore the continuity and the differentiability of the function f, are imposed. The error terms are assumed to be independent and identically distributed with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$. Hence, for the response variable it follows that

$$\mathrm{E}(y_i) = f(z_i) \quad \text{and} \quad \mathrm{var}(y_i) = \sigma^2 \qquad \text{for } i = 1, \ldots, n.$$

Wood (2006) explains in his book that f has to be represented in such a way that (3.1.1) becomes a linear model, to make it possible to estimate f. This can be done by defining the space of functions of which f is an element by choosing a basis. If $B_j(z)$ is the jth basis function, then f has the representation

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z) \tag{3.1.2}$$

for some values of the unknown parameters γ_j. This yields a linear model when substituting Equation (3.1.2) into (3.1.1). The linear representation of f is the reason, why Eilers and Marx (1996) wrote that the name *nonparametric* is not well chosen for spline smoothers, since they are described by parameters, although their number can be large. Therefore, they think it might be better to talk about *overparametric* techniques. In the next sections different basis functions for f are introduced.

3.1.1 Regression Splines

As a first approach for nonparametric regression, *Polynomial Splines* or *Regression Splines* are considered. The idea is to partition the domain of the covariates into intervals and estimate a polynomial for each interval seperately. Furthermore, smoothness restrictions are imposed to guarantee that the function values coincide at the interval boundaries. Fahrmeir et al. (2013) defined a regression spline as follows:

A function $f : [a, b] \to \mathbb{R}$ is called a *polynomial* or *regression spline* of degree $l \geq 0$ with knots $a = \kappa_1 < \cdots < \kappa_m = b$, if it fulfills the following conditions:

- $f(z)$ is $(l-1)$-times continuously differentiable. The special case of $l = 1$ corresponds to $f(z)$ being continuous but not differentiable. If $l = 0$, no smoothness requirements are imposed.
- $f(z)$ is a polynomial of degree l on the intervals $[\kappa_j, \kappa_{j+1})$ defined by the knots.

The global smoothness of the spline is regulated by the degree l of the spline, whereas the diversity of available functions is driven by the number of knots m. There are different popular variants for representing the set of polynomial splines for a given degree and knots configuration: the *Truncated Power Series*, the *B-Splines* and the *Natural Cubic Splines*, which are discussed in detail in the following sections.

In summary, the regression splines are an attractive tool for nonparametric regression because of their computational cleanliness. However, the biggest drawback is the difficulty in choosing the number and the position of the knots. Section 3.1.5 will deal with this problem.

3.1.2 Truncated Power Series

The functions

$$B_1(z) = 1, \quad B_2(z) = z, \quad \dots, \quad B_{l+1}(z) = z^l$$
$$B_{l+2}(z) = (z - \kappa_2)_+^l, \quad \dots, \quad B_d(z) = (z - \kappa_{m-1})_+^l$$

are considered, where

$$(z - \kappa_j)_+^l = \begin{cases} (z - \kappa_j)^l & z \geq \kappa_j \\ 0 & \text{otherwise.} \end{cases}$$

It can be shown that there exists a uniquely determined linear combination of these $d = m + l - 1$ functions for each polynomial spline of degree l. The basis spanned by the functions $B_1, \dots, B_d$ is called *Truncated Power Series Basis (TP Basis)*. This is in fact a basis, since every polynomial spline can be uniquely represented using these functions and the polynomial splines form a d-dimensional vector space. In this case, the nonparametric regression model can be written as

$$\begin{aligned} y_i &= \gamma_1 + \gamma_2 z_i + \dots + \gamma_{l+1} z_i^l + \gamma_{l+2}(z_i - \kappa_2)_+^l + \dots + \gamma_{l+m-1}(z_i - \kappa_{m-1})_+^l + \epsilon_i \\ &= \sum_{j=1}^{d} \gamma_j B_j(z_i) + \epsilon_i. \end{aligned}$$

The first part of the model is a global polynomial of degree l, whereas the second part allows the use of local polynomials in every interval and fulfills the global smoothness restriction.

The main advantage of regression splines is the fact that the nonparametric regression problem can be written and estimated as a linear model with a large number of parameters γ_j. If the design matrix is defined as

$$\boldsymbol{Z} = \begin{pmatrix} B_1(z_1) & \dots & B_d(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \dots & B_d(z_n) \end{pmatrix} = \begin{pmatrix} 1 & z_1 & \dots & z_1^l & (z_1 - \kappa_2)_+^l & \dots & (z_1 - \kappa_{m-1})_+^l \\ \vdots & & & & & & \vdots \\ 1 & z_n & \dots & z_n^l & (z_n - \kappa_2)_+^l & \dots & (z_n - \kappa_{m-1})_+^l \end{pmatrix},$$

the linear model

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{3.1.3}$$

is obtained, where $\boldsymbol{y} = (y_1, \dots, y_n)^\mathsf{T}$ is the vector of the response variables, $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_d)^\mathsf{T}$ the coefficient vector and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^\mathsf{T}$ the vector of the error terms. This linear model can be estimated with the usual least-squares criterion and the estimator is thus

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}, \tag{3.1.4}$$

and the estimated mean function is

$$\hat{f}(z) = \boldsymbol{Z}_1^\mathsf{T}(z)\hat{\boldsymbol{\gamma}}, \tag{3.1.5}$$

where $\boldsymbol{Z}_1(z) = (B_1(z), \ldots, B_d(z))^\mathsf{T}$.

Despite the simplicity of the TP basis, alternative basis are often used in practice. This results from the fact that the calculation of the TP basis functions can cause numerical instabilities for covariates with large values. Furthermore, if knots are very close to each other, the TP basis functions are nearly linear dependent. Therefore, the B-Splines are a numerically favorable polynomial spline basis.

3.1.3 B-Splines

Basic Splines or shortly *B-Splines* are piecewise polynomials which are joined together smoothly at the given knots, so that the smoothness requirements are fulfilled (Eilers and Marx, 1996). More specifically, a B-spline of degree l consists of $l+1$ polynomial pieces of degree l and they join at l inner knots. The derivatives up to order $l-1$ are continuous at the joining points and the B-Spline overlaps with $2l$ polynomial pieces of its neighbours. Furthermore, it is positive on a domain spanned by $l+2$ knots and is zero everywhere else. At a given point z, there are $l+1$ nonzero B-Splines.

The function $f(z)$ can then be represented as a linear combination of $d = m+l-1$ basis functions as

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z).$$

The basis functions are defined recursively as

$$B_j^0(z) = I(\kappa_j \leq z \leq \kappa_{j+1}) = \begin{cases} 1 & \kappa_j \leq z \leq \kappa_{j+1} \\ 0 & \textit{otherwise.} \end{cases}$$

for order $l = 0$ and

$$B_j^l(z) = \frac{z - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(z)$$

for a degree $l \geq 1$. For this definition, m interior knots $\kappa_1, \ldots, \kappa_m$ and $2l$ outer knots are used, so the knots sequence is $\kappa_{1-l}, \kappa_{1-l+1}, \ldots, \kappa_{m+l-1}, \kappa_{m+l}$. Some advantages of the B-spline basis are its local definition and the fact that the basis functions are bounded from above so that there are no numerical problems. Furthermore, the estimation of a polynomial spline in B-spline representation can be performed, as well as the TP basis as a linear model with the design matrix

$$\boldsymbol{Z} = \begin{pmatrix} B_1^l(z_1) & \ldots & B_d^l(z_1) \\ \vdots & & \vdots \\ B_1^l(z_n) & \ldots & B_d^l(z_n) \end{pmatrix}.$$

This design matrix has some special characteristics due to the properties of the B-spline basis functions. Every row of the design matrix sums up to one and therefore the intercept is contained in the span of the basis. Furthermore, the design matrix $\boldsymbol{Z}$ mainly consists of zeros. The resulting linear model has the same form as the TP basis model (3.1.3) and the estimators are given as in (3.1.4) and (3.1.5).

As already mentioned, the main drawback of the regression spline is the free choice of the position and number of knots used. To overcome this problem of the dependence on the chosen knots, the estimation problem can be regularized through the introduction of a roughness penalty. The resulting splines are called *P-Splines* and described in detail in Eilers and Marx (1996). The idea of the roughness penalty is to introduce an additional penalty term in the least-squares criterion to prevent overfitting and then minimize the resulting *Penalized Least-Squares (PLS) Criterion.* This procedure is discussed in Section 3.2.

3.1.4 Cubic Smoothing Splines

A *Smoothing Spline* is not constructed like the splines already discussed in the previous sections. A smoothing spline is obtained when minimizing the *Penalized Sum of Squares*

$$\sum_{i=1}^{n}(y_i - f(z_i))^2 + \lambda \int_a^b f''(z)^2 dz \tag{3.1.6}$$

among all functions $f(z)$ which are twice differentiable (Hastie and Tibshirani, 1990). The first term minimizes the squared error of the model, whereas the second term penalizes curvature in the function. The parameter λ is called the *smoothing parameter.* Small values of λ produce wiggly curves whereas large values of λ produce smoother ones. For $\lambda \to 0$ an interpolating spline is obtained and for $\lambda \to \infty$ the solution of (3.1.6) is a straight line. Therefore, the selection of the smoothing parameter λ is a crucial step in the application of smoothing splines, which will be discussed in Section 3.3.

It can be shown that there exists an explicit, unique minimizer of Criterion (3.1.6). This minimizer is called a *Natural Cubic Spline.* A function $f(z)$ is called a natural cubic spline based on the knots $a \leq \kappa_1 < \cdots < \kappa_m \leq b$, if the following requirements are fulfilled:

- $f(z)$ is a cubic polynomial spline for the given knots, which means that $f(z)$ is a regression spline of degree 3.
- $f(z)$ satisfies the *natural boundary conditions* $f''(a) = f''(b) = 0$, which means that $f(z)$ is linear in the intervals $[a, \kappa_2]$ and $[\kappa_{m-1}, b]$.

Firstly, it can be shown that for given points (κ_i, w_i) ($i = 1, \ldots, m$ and $m \geq 2$), there exists a unique natural cubic spline f which interpolates the points (κ_i, w_i), which means that $f(\kappa_i) = w_i$ for $i = 1, \ldots, m$. The proof of this statement can be found in Green and

Silverman (1994). A possible choice for obtaining the smoothest possible function that interpolates given points would be the function which minimizes $\int f''(z)^2 dz$ among all twice differentiable functions which interpolate the data. This is the reason why the PLS criterion has the form (3.1.6).

As a next step, it will be shown that among all functions that are continuous on $[a, b]$, have absolutely continuous first derivatives and interpolate the values $w_1, \ldots, w_n$ at the points $\kappa_1, \ldots, \kappa_m$, the natural cubic spline is the one that minimizes

$$\int_a^b f''(z)^2 dz.$$

To prove this statement, the idea of Green and Silverman (1994) is followed. Denote $f(z)$ as the natural cubic spline for the points $\kappa_1, \ldots, \kappa_m$ and let $\tilde{f}(z)$ be an interpolant of (κ_i, w_i) other than $f(z)$. Furthermore, let $h(z) = \tilde{f}(z) - f(z)$. Integrating by parts and using the natural boundary conditions $f''(a) = f''(b) = 0$ yields

$$\begin{aligned}\int_a^b f''(z)h''(z)dz &= f''(b)h'(b) - f''(a)h'(a) - \int_a^b f'''(z)h'(z)dz \qquad (3.1.7)\\ &= -\int_a^b f'''(z)h'(z)dz\\ &= -\sum_{i=1}^{m-1} f'''(\kappa_i^+) \int_{\kappa_i}^{\kappa_{i+1}} h'(z)dz\\ &= -\sum_{i=1}^{m-1} f'''(\kappa_i^+)(h(\kappa_{i+1}) - h(\kappa_i)) = 0,\end{aligned}$$

where the equality in line 2 results from the fact that $f(z)$ is made up of piecewise cubic polynomials, thus $f'''(z)$ is constant over any interval (κ_i, κ_{i+1}). The last equality holds because $h(\kappa_i) = 0$ for $i = 1, \ldots, m$, since both f and $\tilde{f}$ interpolate the values w_i. It follows that

$$\begin{aligned}\int_a^b \tilde{f}''(z)^2 dz &= \int_a^b (f''(z) + h''(z))^2 dz\\ &= \int_a^b f''(z)^2 dz + 2\int_a^b f''(z)h''(z)dz + \int_a^b h''(z)^2 dz\\ &= \int_a^b f''(z)^2 dz + \int_a^b h''(z)^2 dz \geq \int_a^b f''(z)^2 dz,\end{aligned}$$

with equality only if h is linear on $[a, b]$. Since $h(\kappa_i) = 0$ for $i = 1, \ldots, m$, this can only happen, if h is zero, which means that $f = \tilde{f}$.

Moreover, it can be shown that among all functions that are continuous on $[a, b]$ and have absolutely continuous first derivatives, the natural cubic spline $f(z)$ is the function that minimizes Criterion (3.1.6). Wood (2006) proved this assumption by supposing that there is some other function $\hat{f}$ which minimizes (3.1.6). Then $(\kappa_i, \hat{f}(\kappa_i))$ can be interpolated using the cubic spline $f(z)$. Since they both minimize the penalized least-squares criterion, their sum of squares have to be the same. Given what was just shown above, the natural cubic spline $f(z)$ has a smaller integrated squared second derivative which would mean that $f(z)$ yields a smaller sum of squares term. That is a contradiction, unless $\hat{f} = f$. The existence and uniqueness of the minimizer of (3.1.6) will not be shown in this thesis, but can be read in Green and Silverman (1994).

To sum up, smoothing splines seem to be the ideal smoother. Nevertheless, one drawback of smoothing splines is that they have as many free parameters as there are observations, which is no problem with univariate smoothing, but becomes a problem as soon as there are more covariates. Consequently, there are implementations where the basis is reduced. One possibility is to reduce the basis utilizing a spectral decomposition of the design matrix. This aspect is discussed in in Section 3.4.2 in the context of thin plate splines.

3.1.5 Influence of the knots and the basis dimension

The choice of the number and the position of the knots is an important step when modeling with splines. If too many knots are used, this results in overfitting the data and too few knots lead to underfitting. There are algorithms for optimizing the number and the position of the knots, as for example proposed by Friedman and Silverman (1989). In general, there is no all-purpose rule for this problem. The more knots we use, the more flexible the estimated function is, but also the more wiggly the function gets. Therefore, in every modeling application, the optimal number of knots has to be seperately considered depending on the concrete problem.

Additional to the number of knots, the position of the knots along the covariate axis has to be chosen. In practice, there are three commonly used techniques:

- *Equidistant knots*: This is a very common and easy approach. Therefor the domain $[a, b]$ is split into $m-1$ intervals to obtain the knots

$$\kappa_j = a + (j-1) \cdot h \quad \text{for } j = 1, \ldots, m \quad \text{with } h = \frac{b-a}{m-1}.$$

- *Quantile-based knots*: The $(j-1)/(m-1)$-quantiles, for $j = 1, \ldots, m$, of the observed values $z_1, \ldots, z_n$ are used as knots. The advantage of this approach is, that many knots are placed in areas with a large number of observations.
- *Visual knot choice*: The decision of the placement of the knots can also be made by studying the scatter plot of the data.

Another main challenge is the selection of the basis dimension. Kim and Gu (2004) have done some empirical studies and suggested the use of $d = cn^{2/9}$ for cubic splines, where n is the size of the data. They suggested using $c \approx 10$ based on their simulation. Many other authors proposed to choose the basis dimension depending on the number of covariates as well as on the sample size. It is indeed difficult to say, if the chosen basis dimension is a good pick. However, as discussed in Wood (2006), it is important to note that the exact size of basis dimension is not that critical. He proposed to decide roughly, in any particular application, how large the basis dimension is fairly certain to provide adequate flexibility. The basis dimension only gives an upper bound on the flexibility of the term whereas the smoothing parameter controls the actual effective degrees of freedom, which are explained in the next section.

3.2 Penalized Least-Squares Criterion

As already mentioned, the problem of the number and the placement of the knots as well as the controlling of the smoothness of the curve, can be solved by introducing a penalty for the coefficients of the corresponding basis functions. This means that instead of the usual residual sum of squares, the penalized least-squares criterion is used, where a penalty term is added to the residual sum of squares. There are many different ways for creating such a penalty, depending on the basis which is used for the model.

When considering a TP basis of degree l, Fahrmeir et al. (2013) proposed that the sum of squared coefficients

$$\sum_{j=l+2}^{d} \gamma_j^2$$

is a good choice for a penalty, since the coefficients with indexes $j = l + 2, \ldots, d$ are associated with the truncated powers. In this case, the penalized residual sum of squares

$$PLS(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2$$

is minimized. The smoothing parameter $\lambda \geq 0$ controls the influence of the penalty term. For $\lambda \to 0$, the penalty term disappears and the penalized residual sum of squares equals the standard residual least-squares. For $\lambda \to \infty$, the $PLS(\lambda)$ is dominated by the penalty term and thus $\hat{\gamma}_j = 0$ for $j = l+2, \ldots, d$, which means that the estimate for $f(z)$ is a polynomial of degree l. The selection of the smoothing parameter λ is discussed in the next section.

If $f(z)$ is represented with B-splines or smoothing splines, an appropriate penalty is less obvious. Since the derivatives of a function respresent a measure for the variability, penalties based on the derivatives are attractive. The most common penalty is of the form

$$\int f''(z)^2 dz. \tag{3.2.1}$$

For B-splines, the construction of this penalty is a simple approach. This yields from the fact that the derivative of the B-spline can easily be computed. For every single basis function, the first derivative is

$$\frac{\partial}{\partial z} B_j^l(z) = l \cdot \left(\frac{1}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z) - \frac{1}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(z) \right).$$

This statement can be proved with complete induction to the degree l, which is just computational effort and will therefore not be shown in this thesis. The first derivative of the polynomial spline can then be written as

$$\frac{\partial}{\partial z} \sum_{j=1}^{d} \gamma_j B_j^l(z) = l \cdot \sum_{j=1}^{d} \frac{\gamma_j - \gamma_{j-1}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z).$$

The derivative can thus be written in terms of the differences of the basis coefficients and the basis functions of one lower degree. Higher-order derivatives can be expressed analogously using higher-order differences. Eilers and Marx (1996) therefore proposed to use the finite differences of the coefficients of adjacent B-splines as penalty. The PLS criterion in this case has the form

$$PLS(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=k+1}^{d} (\Delta^k \gamma_j)^2,$$

where Δ^k denotes the kth-order differences, recursively defined by

$$\begin{aligned}
\Delta^1 \gamma_j &= \gamma_j - \gamma_{j-1}, \\
\Delta^2 \gamma_j &= \Delta^1 \gamma_j - \Delta^1 \gamma_{j-1} = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2}, \\
&\vdots \\
\Delta^k \gamma_j &= \Delta^{k-1} \gamma_j - \Delta^{k-1} \gamma_{j-1}.
\end{aligned}$$

If the penalty term is chosen, the resulting PLS criterion can be minimized. In the rest of the thesis, the chosen penalty will be (3.2.1). However, the following calculations can simply be adapted to other penalties. Since $f(z)$ is assumed to be respresented with basis functions as

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z),$$

the penalty term can be written as

$$\int f''(z)^2 dz = \sum_{i=1}^{d} \sum_{j=1}^{d} \gamma_i \gamma_j \int B_i''(z) B_j''(z) dz = \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{K} \boldsymbol{\gamma}$$

with the vector of coefficients $\boldsymbol{\gamma}$ and the penalty matrix $\boldsymbol{K}$ defined elementwise by

$$\boldsymbol{K}_{ij} = \int B_i''(z)B_j''(z)dz.$$

The resulting PLS criterion has the form

$$PLS(\lambda) = (\boldsymbol{y} - \boldsymbol{Z\gamma})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{Z\gamma}) + \lambda\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{K\gamma},$$

since the regression model has the linear form $\boldsymbol{y} = \boldsymbol{Z\gamma} + \boldsymbol{\epsilon}$. The minimization of this optimization criterion yields the estimate

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y}.$$

Thus, the difference to the estimate in the classical linear model is the additional $\lambda\boldsymbol{K}$ term. The estimated function is

$$\hat{f} = \boldsymbol{Z}\hat{\boldsymbol{\gamma}} = \boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y} = \boldsymbol{S}(\lambda)\boldsymbol{y},$$

where $\boldsymbol{S}(\lambda) = \boldsymbol{Z}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}^\mathsf{T}$ is called the $n \times n$ *smoother matrix.* The meaning of this matrix will be explained later.

For computation it is not advisable to use these expressions. Orthogonal methods are to be preferred regarding greater numerical stability. For this methods, note that

$$\left(\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{pmatrix} - \begin{pmatrix}\boldsymbol{Z}\\ \sqrt{\lambda}\boldsymbol{B}\end{pmatrix}\boldsymbol{\gamma}\right)^\mathsf{T}\left(\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{0}\end{pmatrix} - \begin{pmatrix}\boldsymbol{Z}\\ \sqrt{\lambda}\boldsymbol{B}\end{pmatrix}\boldsymbol{\gamma}\right) = (\boldsymbol{y} - \boldsymbol{Z\gamma})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{Z\gamma}) + \lambda\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{K\gamma}, \quad (3.2.2)$$

where $\boldsymbol{B}$ is any square root of the matrix $\boldsymbol{K}$ such that $\boldsymbol{B}^\mathsf{T}\boldsymbol{B} = \boldsymbol{K}$. As described in Wood (2004) and Wood (2011), $\boldsymbol{B}$ can be obtained by a spectral decomposition or pivoted Choleski decomposition. It can be seen in (3.2.2) that the least-squares term on the right hand side is just a least-squares criterion for a model in which the model matrix $\boldsymbol{Z}$ has been augmented with a square root of the matrix $\boldsymbol{B}$ and the response vector $\boldsymbol{y}$ with d zeros. Once the matrix $\boldsymbol{B}$ has been calculated, the augmented least-squares problem on the left hand side can be solved using orthogonal methods to solve the original penalized least-squares model on the right hand side.

The smoother matrix $\boldsymbol{S}(\lambda)$ gives a measure to evaluate the approximate dimension of the smoothness for an estimated function. The *equivalent degrees of freedom* or *effective number of parameters* of a smoother are defined as

$$df(\boldsymbol{S}(\lambda)) = \mathrm{tr}(\boldsymbol{S}(\lambda)),$$

where $\mathrm{tr}(\boldsymbol{S}(\lambda)) = \sum_{i=1}^{n}\boldsymbol{S}_{ii}(\lambda)$ is the trace of the matrix $\boldsymbol{S}(\lambda)$. This measure results from ideas for the linear model, where the number of parameters can be computed as $\mathrm{tr}(\boldsymbol{H}) = p$ from the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}$. In fact, the number of parameters is reproduced in the model without penalization, while in the case of penalization the penalty effectively reduces the number of parameters. For polynomial splines, the effective number of parameters equals the number of basis functions, whereas for penalized splines the effective number of parameters decreases with an increasing smoothing parameter. Furthermore, there are two other definitions of the equivalent degrees of freedom (Hastie and Tibshirani, 1990). They are given by $\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^\mathsf{T})$ and $n - \mathrm{tr}(2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{S}^\mathsf{T})$. These definitions are useful for different purposes and can be motivated again with the linear regression model. For smoothing splines it can be shown that $\mathrm{tr}(\boldsymbol{S}\boldsymbol{S}^\mathsf{T}) \leq \mathrm{tr}(\boldsymbol{S}) \leq \mathrm{tr}(2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{S}^\mathsf{T})$.

3.3 Selection of the Smoothing Parameter

As already seen in the previous sections, the smoothing parameter plays an important role in the estimation of nonparametric regression models and thus has to be considered in detail. There are different automatic procedures available for choosing the smoothing parameter. One possible choice is to choose the smoothing parameter by minimizing the *Akaike Information Criterion (AIC)*, of which there are slightly different definitions. Fahrmeir et al. (2013) defined it as

$$AIC = n \log(\hat{\sigma}^2) + 2(df + 1), \tag{3.3.1}$$

where $\hat{\sigma}^2 = \sum (y_i - \hat{f}(z_i))^2/n$ is the maximum likelihood estimator for σ^2 under the considered model and df are the degrees of freedom of the model. In general, the AIC is defined as

$$AIC = -2 \log L(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2 | \boldsymbol{y}) + 2(df + 1),$$

where $L(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2 | \boldsymbol{y})$ is the likelihood function estimated for the maximum likelihood estimators $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$, based on a normally distributed response $\boldsymbol{y}$ with mean $\boldsymbol{\mu}$ and variance $\sigma^2 \boldsymbol{I}_n$. When calculating $-2 \log L(\hat{\boldsymbol{\mu}}, \hat{\sigma}^2 | \boldsymbol{y})$, we obtain $n \log \hat{\sigma}^2$ plus a constant. Thus, the essential part of the AIC is given as (3.3.1). When minimizing the AIC, we therefore maximize the log-likelihood, but also use the number of parameters as a form of penalization for the model complexity. In contrast to linear models, the AIC uses the effective number of parameters instead of the actual number of parameters.

A more common approach for the selection of the smoothing parameter is the *Cross Validation Criterion (CV Criterion)*. This criterion results from the fact that both the bias and the variance of a fitted smooth function depend on the smoothing parameter. Let the bias be defined as $\boldsymbol{b}_\lambda = \mathrm{E}(\boldsymbol{S}(\lambda)\boldsymbol{y}) - f = \boldsymbol{S}(\lambda)f - f$. It is not possible to decrease both bias and variance simultaneously. Therefore, a measure which includes both of them is minimized, namely the *Average Mean Squared Error (MSE)*, given as

$$\begin{aligned} MSE(\hat{f}(z)) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\hat{f}(z_i) - f(z_i))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{var}(\hat{f}(z_i)) + \frac{1}{n}\sum_{i=1}^{n} b_\lambda^2(z_i) \\ &= \frac{\mathrm{tr}(\boldsymbol{S}(\lambda)\boldsymbol{S}(\lambda)^\mathsf{T})}{n}\sigma^2 + \frac{\boldsymbol{b}_\lambda^\mathsf{T}\boldsymbol{b}_\lambda}{n}, \end{aligned}$$

with $\boldsymbol{b}_\lambda = (b_\lambda(z_1), \dots, b_\lambda(z_n))^\mathsf{T}$. Is the amount of smoothing increased, then the bias will increase while the variance will decrease, and conversely when the amount of smoothing is decreased. This is in fact true, since $\mathrm{tr}(\boldsymbol{S}(\lambda)\boldsymbol{S}(\lambda)^\mathsf{T})$ decreases, when the amount of smoothing increases, whereas the elements of $\boldsymbol{b}_\lambda$ increase and conversely. Since a naive approximation of the MSE is the average residual sum of squares

$$\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{f}(z_i))^2,$$

which is minimized for $\hat{f}(z_i) = y_i$, this procedure is not useful. As a consequence, the squared prediction error for new observations is rather used for determining an optimal smoothing parameter. The *Average Predictive Squared Error (PSE)* is defined by Hastie and Tibshirani (1990) as

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(y_i^* - \hat{f}(z_i))^2,$$

where y_i^* is a new observation at z_i, which means that $y_i^* = f(z_i) + \epsilon_i^*$ with ϵ_i^* independent of ϵ_i. It can be shown that $PSE(\lambda) = MSE(\lambda) + \sigma^2$. The CV works as follows: A smoothing parameter λ is fixed and the smooth is estimated at z_i by leaving out points (y_i, z_i) one at a time, based on the remaining $n-1$ observations. Let $\hat{f}^{(-i)}(z)$ denote the estimated mean obtained when removing the observation (y_i, z_i). The estimated mean will then be used to predict the function value $f(z_i)$ at the eliminated observation. By doing this for every point (y_i, z_i), the cross validation criterion is obtained as

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{f}^{(-i)}(z_i))^2.$$

The selection of the smoothing parameter then works as follows: $CV(\lambda)$ is computed for a range of values of λ and the minimizing $\hat{\lambda}$ is selected. The cross validation criterion is justified by the fact that

$$\mathrm{E}(CV(\lambda)) \approx PSE(\lambda), \tag{3.3.2}$$

and furthermore, that the minimizer of $CV(\lambda)$ is close to the minimizer of $PSE(\lambda)$. Property (3.3.2) yields from

$$\begin{aligned} \mathrm{E}\left(y_i - \hat{f}^{(-i)}(z_i)\right)^2 &= \mathrm{E}\left(y_i - f(z_i) + f(z_i) - \hat{f}^{(-i)}(z_i)\right)^2 \\ &= \sigma^2 + \mathrm{E}\left(f(z_i) - \hat{f}^{(-i)}(z_i)\right)^2, \end{aligned}$$

where the cross-product term is zero because $\hat{f}^{(-i)}(z_i)$ is independent of y_i. It follows analogously that

$$\mathrm{E}\left(y_i^* - \hat{f}(z_i)\right)^2 = \sigma^2 + \mathrm{E}\left(f(z_i) - \hat{f}(z_i)\right)^2 .$$

With the additional assumption $\hat{f}^{(-i)}(z_i) \approx \hat{f}(z_i)$, the claim in (3.3.2) is justified. The average mean-squared error thus differs from the average predictive squared error only by the constant σ^2. Wood (2006) argumented that complicated models are always preferred over simpler ones, when just looking at their ability to fit the data from which they were estimated. If the model is chosen in order to maximize the ability to predict data to which the model was not fitted, the problem vanishes.

As a next step, the calculation of $\hat{f}^{(-i)}(z_i)$, the fit at z_i with the ith point removed, named by Hastie and Tibshirani (1990) as the *jackknifed fit* at z_i, will be discussed. To do so, the corresponding smoother for $n-1$ points must be defined, while leaving out the point (z_i, y_i). Therefore, a definition of $\hat{f}^{(-i)}(z_i)$ is given, by utilizing the smoother matrix $\boldsymbol{S}(\lambda)$. This yields

$$\hat{f}^{(-i)}(z_i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{\boldsymbol{S}_{ij}(\lambda)}{1 - \boldsymbol{S}_{ii}(\lambda)} y_j. \tag{3.3.3}$$

To prove (3.3.3), the elements of each row of $\boldsymbol{S}(\lambda)$ are thought as weights, which makes sense since the smoother matrix is constant preserving, that is $\boldsymbol{S}(\lambda)\mathbf{1} = \mathbf{1}$, for an n-vector of ones $\mathbf{1}$. This implies that the sum of the weights in each row is one. To calculate $\hat{f}^{(-i)}(z_i)$, the weight on the ith observation is set to zero while the remaining weights are increased such that they sum up to one. This yields (3.3.3).

The Statement (3.3.3) implies an important relationship, namely

$$\hat{f}^{(-i)}(z_i) = \sum_{\substack{j=1 \\ j\neq i}}^{n} \boldsymbol{S}_{ij}(\lambda) y_j + \boldsymbol{S}_{ii}(\lambda)\hat{f}^{(-i)}(z_i). \tag{3.3.4}$$

In terms of deletion diagnostics, (3.3.4) says that if we add a new point that lies exactly on the regression surface, that point does not change the fitted regression. It must be pointed out that the above statement does not hold in full generality for all types of smoothers. However, it is still used to approximate the CV and it can be shown that it is valid for cubic smoothing splines. To see that, assume that $\hat{f}^{(-i)}$ minimizes the PLS criterion

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} (y_j - f(z_j))^2 + \lambda \int f''(z)^2 dz \tag{3.3.5}$$

for a sample size of $n-1$. Suppose the point $(z_i, \hat{f}^{(-i)}(z_i))$ is added to the sample. Then the value of (3.3.5) stays the same and therefore, $\hat{f}^{(-i)}$ must still minimize the PLS criterion (3.3.5) for sample size n because if there was another cubic spline producing a smaller value of (3.3.5), this spline would also produce a smaller value over the original $n-1$ points.

From (3.3.3) it follows that

$$y_i - \hat{f}^{(-i)}(z_i) = \frac{y_i - \hat{f}(z_i)}{1 - \boldsymbol{S}_{ii}(\lambda)},$$

and thus $\hat{f}^{(-i)}(z_i)$ can be computed using $\hat{f}(z_i)$ and $\boldsymbol{S}_{ii}(\lambda)$ and it is not necessary to actually remove the ith point from the data and recompute the smooth.

To sum up, the cross validation sum of squares can be written as

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(z_i)}{1 - \boldsymbol{S}_{ii}(\lambda)}\right)^2.$$

Since it was not known for a long time, how to compute the diagonal elements $\boldsymbol{S}_{ii}(\lambda)$ for a smoothing spline in $O(n)$ operations, another cross validation was introduced, namely the *Generalized Cross Validation (GCV)* (Craven and Wahba, 1978). The GCV replaces $\boldsymbol{S}_{ii}(\lambda)$ by its average value $\mathrm{tr}(\boldsymbol{S}(\lambda))/n$, which yields

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}(z_i)}{1 - \mathrm{tr}(\boldsymbol{S}(\lambda))/n}\right)^2.$$

The sum of the diagonal elements of the smoother matrix $\boldsymbol{S}(\lambda)$, the trace of $\boldsymbol{S}(\lambda)$, corresponds to the definition of the equivalent degrees of freedom. The original motivation for the GCV is no longer valid, since there currently are algorithms that can compute $\boldsymbol{S}_{ii}(\lambda)$ for the cubic smoothing spline in $O(n)$ operations. Hutchinson and de Hoog (1985) described an algorithm for computing the diagonal elements of $\boldsymbol{S}(\lambda)$. Nevertheless, Wood (2006) explained that the GCV in contrast to the CV is invariant under orthogonal transformations of the data.

It can be shown that both criteria, AIC and GCV, are asymptotically equivalent (Fahrmeir et al., 2013), but can be clearly different in some applications. The optimal smoothing parameter can thus be found by defining a grid of candidate smoothing parameters and choosing the value that minimizes the criterion. Care has to be taken since it cannot be guaranteed that the function has a unique minimum. Direct optimization is possible for univariate smoothing models, but is no longer efficient when considering more than one covariate. For the multivariate regression model, there are algorithms for an efficient minimization of the AIC and the GCV. The regression model with multiple predictors will be the focus in the following sections.

3.4 Bivariate Smoothing

Thus far, nonparametric regression models with one single continuous covariate were considered. In this section, the concept of bivariate smoothing is introduced. The idea of bivariate smoothing is to model the predictor variable y_i as a function of two continuous covariates z_{i1} and z_{i2}, which yields

$$y_i = f(z_{i1}, z_{i2}) + \epsilon_i, \qquad \text{for } i = 1, \dots, n. \tag{3.4.1}$$

In many applications bivariate smoothing is reasonable, for example when considering longitudinal and latitudinal data. In this special case, it is necessary to model a function depending on both, longitude and latitude. In general, bivariate or even multiple smoothing with more than two covariates is required, when there is interaction between the covariates. Bivariate smoothing models are obtained by generalizing the univariate smoothing models. Some common methods for bivariate smoothing are, for example, the *Tensor Product P-Splines*, the *Thin Plate Splines* and the *SOAP Film Smooths*, which are introduced in the following sections.

3.4.1 Tensor Product P-Splines

According to Fahrmeir et al. (2013), the *Tensor Product Bases* are obtained when considering pairwise products of two univariate bases constructed for univariate smooths. Firstly, the univariate bases for the covariates z_1 and z_2 are constructed, yielding the basis functions $B_j^{(1)}(z_1)$ $(j = 1, \dots, d_1)$ and $B_r^{(2)}(z_2)$ $(r = 1, \dots, d_2)$. By multiplying these basis functions, the basis functions of the tensor product bases are obtained and have the form

$$B_{jr}(z_1, z_2) = B_j^{(1)}(z_1) \cdot B_r^{(2)}(z_2) \qquad \text{for } j = 1, \dots, d_1, \quad r = 1, \dots, d_2.$$

The function f in (3.4.1) thus has the representation

$$f(z_1, z_2) = \sum_{j=1}^{d_1} \sum_{r=1}^{d_2} \gamma_{jr} B_{jr}(z_1, z_2),$$

with regression coefficients γ_{jr}, defined for $j = 1, \dots, d_1$ and $r = 1, \dots, d_2$. At first sight, it seems that the tensor product approaches are much more complex than the univariate ones. It is indeed possible to represent the tensor product splines in form of a large linear model and therefore, the tensor product splines are not that complex as one would expect. The design matrix $\boldsymbol{Z}$ can be written with rows

$$\boldsymbol{z}_i^\mathsf{T} = (B_{11}(z_{i1}, z_{i2}), \dots, B_{d_1 1}(z_{i1}, z_{i2}), \dots, B_{1 d_2}(z_{i1}, z_{i2}), \dots, B_{d_1 d_2}(z_{i1}, z_{i2}))$$

for $i = 1, \dots, n$ and the corresponding vector of regression coefficients as

$$\boldsymbol{\gamma} = (\gamma_{11}, \dots, \gamma_{d_1 1}, \dots, \gamma_{1 d_2}, \dots, \gamma_{d_1 d_2})^\mathsf{T}.$$

Thus, bivariate smoothing approaches can also be estimated within the scope of linear models, since it follows that

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}.$$

Nevertheless, it is important to compute the estimator $\hat{\boldsymbol{\gamma}}$ in a numerically efficient way, since the number of parameters is much larger as in the univariate case. This calculation can for example be done by utilizing univariate tensor product B-splines for $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ and benefiting from the sparse structure of the design matrix, as already mentioned in Section 3.1.3.

As already discussed in Section 3.2 in the case of univariate smoothing, a penalty term has to be added to the usual least-squares problem to ensure obtaining an adequate smooth curve. In the two-dimensional case, one way to construct penalties, is to use the Kronecker products of univariate penalty matrices. This means that the resulting penalty matrix has the form

$$\boldsymbol{K} = \boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1 + \boldsymbol{K}_2 \otimes \boldsymbol{I}_{d_1}, \tag{3.4.2}$$

where $\boldsymbol{I}_d$ is the d-dimensional identity matrix. The Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ of a $n \times p$ matrix $\boldsymbol{A}$ and a $r \times q$ matrix $\boldsymbol{B}$ is defined as the $nr \times pq$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1p}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{n1}\boldsymbol{B} & a_{n2}\boldsymbol{B} & \cdots & a_{np}\boldsymbol{B} \end{pmatrix}.$$

Then the properties

$$(\boldsymbol{A} \otimes \boldsymbol{B})^\top = \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top \qquad \text{and} \qquad (\boldsymbol{A}\boldsymbol{B}) \otimes (\boldsymbol{C}\boldsymbol{D}) = (\boldsymbol{A} \otimes \boldsymbol{C})(\boldsymbol{B} \otimes \boldsymbol{D}) \tag{3.4.3}$$

hold for matrices of appropriate order. The matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$ in (3.4.2) can be calculated as

$$\boldsymbol{K}_1 = \boldsymbol{D}_1^\top \boldsymbol{D}_1 \qquad \text{and} \qquad \boldsymbol{K}_2 = \boldsymbol{D}_2^\top \boldsymbol{D}_2,$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are the matrices of univariate first order differences in z_1 and z_2 direction, respectively. The penalty matrix (3.4.2) is obtained as follows: Applying the matrix $\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1$ to the vector $\boldsymbol{\gamma}$ yields row-wise first-order differences. Thus, the sum of all squared row-wise differences is

$$\boldsymbol{\gamma}^\top (\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1)^\top (\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1)\boldsymbol{\gamma} = \sum_{r=1}^{d_2} \sum_{j=2}^{d_1} (\gamma_{jr} - \gamma_{j-1,r})^2.$$

Analogously, the squared column-wise differences are

$$\boldsymbol{\gamma}^\top (\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1})^\top (\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1})\boldsymbol{\gamma} = \sum_{j=1}^{d_1} \sum_{r=2}^{d_2} (\gamma_{jr} - \gamma_{j,r-1})^2.$$

Summing up the squared row-wise and the squared column-wise differences yields the penalty matrix

$$\boldsymbol{K} = (\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1)^\top (\boldsymbol{I}_{d_2} \otimes \boldsymbol{D}_1) + (\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1})^\top (\boldsymbol{D}_2 \otimes \boldsymbol{I}_{d_1}).$$

Using Properties (3.4.3), the penalty matrix can be written in the form (3.4.2) and therefore the resulting penalty has the form

$$\lambda \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{K} \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}^{\mathsf{T}} (\boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1 + \boldsymbol{K}_2 \otimes \boldsymbol{I}_{d_1}) \boldsymbol{\gamma}.$$

The idea of building two-dimensional penalties based on row-wise or column-wise differences can also be extended to difference matrices of a higher order. In this case, the penalties have the form

$$\lambda \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{K} \boldsymbol{\gamma} = \lambda \boldsymbol{\gamma}^{\mathsf{T}} \left(\boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1^{(k_1)} + \boldsymbol{K}_2^{(k_2)} \otimes \boldsymbol{I}_{d_1} \right) \boldsymbol{\gamma},$$

where $\boldsymbol{K}_1^{(k_1)}$ and $\boldsymbol{K}_2^{(k_2)}$ are the univariate penalty matrices of orders k_1 and k_2. A penalty based on squared second-order differences can thus be obtained with $k_1 = k_2 = 2$.

One of the main drawbacks of tensor product P-splines is the fact that, as with univariate splines, the optimal number and position of knots has to be chosen. The problem that may occur is that there may be data regions with no observations, so it is impossible to estimate the corresponding coefficients of the basis functions. The same problem may occur in the univariate case, when the data has large gaps. However, this problem is more common in the two-dimensional case and can be eliminated by adding a penalty for regularizing the estimation problem. Another way to overcome the above mentioned problem, is the use of *Thin Plate Splines* or *SOAP Film Smooths.*

3.4.2 Thin Plate Splines

Thin Plate Splines are a way of constructing bivariate basis functions by using *Radial Bases.* According to Fahrmeir et al. (2013), a radial basis function is a function of the Euclidean distance between a knot $\boldsymbol{\kappa} = (\kappa_1, \kappa_2)$ and an observation point $\boldsymbol{z} = (z_1, z_2)$, which means that

$$B_{\kappa}(\boldsymbol{z}) = B(\|\boldsymbol{z} - \boldsymbol{\kappa}\|) = B(r),$$

where B is a suitably chosen scalar function and $r = \|\boldsymbol{z} - \boldsymbol{\kappa}\| = ((z_1 - \kappa_1)^2 + (z_2 - \kappa_2)^2)^{0.5}$ is the Euclidean distance. The name *radial bases* results from the fact that the contour plots of the radial basis functions consist of circular contour lines, whereas the contour plots of tensor product B-splines clearly deviate from circles. The chosen knots of a radial basis are typically a subset of the observation points, which means that $\{\kappa_1, \ldots, \kappa_m\} \subset \{z_1, \ldots, z_n\}$, so the distribution of the radial basis functions matches the distribution of the data. Thus, it is no problem for the radial bases if there are regions in the data with no observations.

In the one-dimensional case, the natural cubic spline can be found by finding the smoothest curve that interpolates a given set of data and fulfills smoothness restrictions in terms of integrated squared second derivative. In the bivariate case, a functional $J(f)$ has to be found, which measures the overall roughness of $f = f(z_1, z_2)$ in an analogous way as in the univariate case. Green and Silverman (1994) imposed some desiderata for a suitable roughness functional $J(f)$ and defined an appropriate one as

$$J(f) = \int\int_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f}{\partial z_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial z_1 \partial z_2} \right)^2 + \left(\frac{\partial^2 f}{\partial z_2^2} \right)^2 \right] dz_1 dz_2. \tag{3.4.4}$$

If the second derivatives of f are square-integrable over $\mathbb{R}^2$, then the penalty function $J(f)$ will be finite. By minimizing $J(f)$ subject to certain constraints, which will be discussed later, we obtain the thin plate splines. The penalty function $J(f)$ does indeed quantify the roughness of f, since $J(f)$ will be large if f exhibits high local curvature because then the second derivatives will be large. According to Green and Silverman (1994), the penalty function $J(f)$ has some good properties. It can for example be shown that if the coordinates in $\mathbb{R}^2$ are rotated, then $J(f)$ is unaffected. Furthermore, $J(f)$ is always non-negative, and it is zero if and only if f is a linear function. The details of these properties can be read in Green and Silverman (1994).

The importance of the thin plate splines is the fact that they have an optimality property. More precisely, the thin plate spline uniquely minimizes $J(f)$ subject to the interpolation conditions $f(\kappa_{i1}, \kappa_{i2}) = w_i$ for $i = 1, \dots, n$, for a given set of values $w_1, \dots, w_n$ and knots $\boldsymbol{\kappa}_i = (\kappa_{i1}, \kappa_{i2})$. The exact proof of this statement will not be given in this thesis, but major steps of the proof will be provided by discussing further properties of thin plate splines. A sketch of the proof is given in Green and Silverman (1994).

Firstly, a thin plate spline will be defined in general, without the penalty function $J(f)$. Suppose the dataset $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ in $\mathbb{R}^2$ is given. A function f is called a thin plate spline if and only if it is of the form

$$f(\boldsymbol{z}) = \sum_{i=1}^{n} \gamma_i \eta(\|\boldsymbol{z} - \boldsymbol{z}_i\|) + \sum_{j=0}^{2} \beta_j \phi_j(\boldsymbol{z}), \tag{3.4.5}$$

for suitable constants γ_i $(i = 1, \dots, n)$ and β_j $(j = 0, 1, 2)$. The function η is given as

$$\eta(r) = \begin{cases} \frac{1}{16\pi} r^2 \log(r^2) & \text{for } r > 0 \\ 0 & \text{otherwise,} \end{cases}$$

and the functions ϕ_j are defined for a point $\boldsymbol{z}_i = (z_{i1}, z_{i2})$ by

$$\begin{aligned} \phi_0(\boldsymbol{z}_i) &= 1, \\ \phi_1(\boldsymbol{z}_i) &= z_{i1}, \\ \phi_2(\boldsymbol{z}_i) &= z_{i2}, \end{aligned}$$

so that every linear function can be written as a linear combination of the ϕ_j. In the above definition (3.4.5), $\|\boldsymbol{z}\|$ denotes the Euclidean norm of a vector $\boldsymbol{z}$, which means that

$\|\boldsymbol{z}\| = \boldsymbol{z}^\mathsf{T}\boldsymbol{z}$. Furthermore, the function f in (3.4.5) is called a *natural* thin plate spline, if the vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n)^\mathsf{T}$ satisfies

$$\boldsymbol{X}^\mathsf{T}\boldsymbol{\gamma} = \boldsymbol{0}, \tag{3.4.6}$$

where $\boldsymbol{X}$ is defined by

$$\boldsymbol{X} = \begin{pmatrix} 1 & \boldsymbol{z}_1 \\ 1 & \boldsymbol{z}_2 \\ \vdots & \vdots \\ 1 & \boldsymbol{z}_n \end{pmatrix}. \tag{3.4.7}$$

From the restriction (3.4.6) it follows immediately that

$$\sum_{i=1}^{n} \gamma_i = \sum_{i=1}^{n} \gamma_i \boldsymbol{z}_i = 0.$$

As a next step, it can be shown that if f is a thin plate spline, then $J(f)$ is finite if and only if f is a natural thin plate spline. Furthermore, if f is a natural thin plate spline, then

$$J(f) = \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\gamma},$$

where $\boldsymbol{Z}$ is a $n \times n$ matrix with components

$$\begin{aligned} \boldsymbol{Z}_{ij} &= \eta(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|) \\ &= \frac{1}{16\pi}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2 \log \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2, \end{aligned}$$

and $\boldsymbol{Z}_{ii} = 0$.

Finally, for given values $w_1, \dots, w_n$, there exists a unique natural thin plate spline f on the set $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$, which are distinct non-collinear points in $\mathbb{R}^2$, such that

$$f(\boldsymbol{z}_i) = w_i \qquad \text{for } i = 1, \dots, n.$$

Thus, for a given set of values, there exists a unique natural thin plate spline that interpolates the given values. When putting all the above statements about thin plate splines together, it can be shown that there exists a natural thin plate spline interpolant which uniquely minimizes $J(f)$.

The representation (3.4.5) can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$ is defined as in (3.4.7), $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\mathsf{T}$ is the corresponding coefficients vector and the matrix $\boldsymbol{Z}$, as already defined before, contains the radial basis functions evaluated at the observed covariate values, which means that

$$\boldsymbol{Z}_{ij} = B_j(z_{i1}, z_{i2}) = \eta(\|\boldsymbol{z}_i - \boldsymbol{z}_j\|).$$

To obtain an identifiable version, an additional restriction has to be made, since there are $n+3$ parameters and therefore the model is overspecified. As already mentioned, a suitable choice is the restriction $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\gamma} = \mathbf{0}$, which ensures that the linear part of the model is orthogonal to the radial basis functions part. Furthermore, as outlined before, the penalty term (3.4.4) can be written as

$$\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\gamma},$$

so in this case, the penalty matrix equals the design matrix. Hence, the criterion which has to be minimized can be represented as

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\gamma})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\gamma}, \tag{3.4.8}$$

subject to the constraint

$$\boldsymbol{X}^{\mathsf{T}}\boldsymbol{\gamma} = \mathbf{0}.$$

The problem with this minimization is that when minimizing (3.4.8), a system of $(n+3) \times (n+3)$ equations has to be solved. It is therefore reasonable to obtain low rank approximations to the thin plate spline. Wood (2003) proposed using a spectral decomposition of the design matrix $\boldsymbol{Z}$ to obtain an optimal approximation as close to the solution as possible. The idea is to find the parameter space basis of a given rank that perturbs the given problem as little as possible and solve the resulting low rank problem. Firstly, the spectral decomposition

$$\boldsymbol{Z} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^{\mathsf{T}}$$

has to be computed, where $\boldsymbol{\Gamma}$ is an orthonormal matrix of eigenvectors and $\boldsymbol{\Omega}$ consists of the corresponding eigenvalues in descending order. The idea is to replace $\boldsymbol{Z}$ by $\boldsymbol{Z}_d$, where $\boldsymbol{Z}_d$ is the best rank d approximation to $\boldsymbol{Z}$ in the sense of the spectral norm $\|\boldsymbol{Z} - \boldsymbol{Z}_d\|$, where the spectral norm of a matrix $\boldsymbol{A}$ corresponds to the square root of the largest eigenvalue of the positive semidefinite matrix $\boldsymbol{A}$. Therefore, the matrices $\boldsymbol{\Gamma}_d$ and $\boldsymbol{\Omega}_d$ are defined as submatrices of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$, respectively, associated with the d largest eigenvalues. Wood (2003) showed, that for $\boldsymbol{Z}_d$ it follows that

$$\boldsymbol{Z}_d = \boldsymbol{\Gamma}_d\boldsymbol{\Omega}_d\boldsymbol{\Gamma}_d^{\mathsf{T}}.$$

When replacing $\boldsymbol{Z}$ by $\boldsymbol{Z}_d$, the original estimation problem is projected into the optimal d-dimensional subspace, where d can be chosen such that the approximation error is as small as possible. The exact proof and procedure is discussed in detail in Wood (2003).

It is also possible to define the penalty functional $J(f)$ in general form for more than two dimensions. When considering q points $\boldsymbol{z}_1, \dots, \boldsymbol{z}_q$ in the d-dimensional space, $J(f)$ has the form

$$J_d^q(f) = \sum_{\alpha_1 + \dots + \alpha_q = d} \frac{d!}{\alpha_1! \dots \alpha_q!} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\frac{\partial^d f}{\partial \boldsymbol{z}_1^{\alpha_1} \dots \partial \boldsymbol{z}_q^{\alpha_q}} \right)^2 \prod_{j=1}^{q} d\boldsymbol{z}_j.$$

Wahba (1990) showed that a general thin plate spline is the solution of the following problem: Find the function f that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(z_{i1},\ldots,z_{id}))^2 + \lambda J_d^q(f).$$

The details of this general case is discussed in detail in Wahba (1990) and we will not go into detail in this thesis.

3.4.3 SOAP Film Smooths

Conventional smoothing methods have the disadvantage that they have problems when smoothing across boundary features. This problem may occur when considering a geographical area where the boundaries of the domain matter, such as for instance a bay. The conventional smoothing methods tend to smooth across the water, resulting in inappropriate estimates on either side of the bay. Thus, another class of smoothers has to be introduced to deal with complicated subregions of $\mathbb{R}^2$, while still remaining computationally efficient. This class of smooths is introduced by Wood, Bravington, and Hedley (2008) and is called the *SOAP Film Smooth.*

Firstly, this class of smooths is motivated by a physical model. Consider a region Ω on the $x-y$ plane in which we are interested. A loop of wire is taken to follow the boundary of the region on the plane. The known function values at the boundary are given by the vertical displacement of the loop above the plane. The name *SOAP film smooth* results from the fact, that an appropriate smooth function over this domain is obtained by considering a soap film supported by the boundary wire. For defining this soap film, the assumption that the vertical displacement of the wire is small enough, has to be made. Then the height of the soap film inside the boundary is given by a function $f = f(x, y)$ satisfying the boundary conditions and

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0.$$

Furthermore, the soap film should distort smoothly from its minimum energy configuration by moving vertically towards the data. Therefore, an appropriate measure of the total degree of distortion is

$$J_\Omega(f) = \int_\Omega \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}\right)^2 dxdy. \tag{3.4.9}$$

It must be pointed out that in contrast to the other sections in this thesis, the variables x and y here denote the covariates in x- and y-direction, respectively. The value of the function f at the point (x, y) will be defined by z. There are three differences between the penalty functional (3.4.9) and the penalty functional (3.4.4) from the thin plate spline:

- The functional (3.4.9) is integrated over Ω instead of the whole $x-y$ plane as in the functional of the thin plate splines,

- in the penalty functional (3.4.9) there is no mixed second derivative term,
- and in the functional of the thin plate splines, the different terms are being squared separately, whereas in the case of SOAP film smooths, the sum of the second derivatives is squared.

The latter difference is the most important one. It allows the second derivatives with respect to x and y to be traded off against each other, which means, that the space of functions for which $J_\Omega(f)$ is zero is infinite dimensional. Thus, functions with zero penalty can meet any boundary condition by being curved enough.

Assume that there are n data points $z_1, \ldots, z_n$ on the domain $(x_1, y_1), \ldots, (x_n, y_n)$. The SOAP film smooths are obtained by minimizing

$$\sum_{i=1}^{n} (z_i - f(x_i, y_i))^2 + \lambda J_\Omega(f)$$

with respect to f and subject to some known boundary conditions. Wood et al. (2008) proved the SOAP film interpolant theorem in their article: Consider a smooth function $f^*(x, y)$ over a $x - y$ plane and let B be a collection of closed loops in that plane, such that no two loops intersect and one outer loop encloses all the others. Furthermore, let Ω be the region which is made up of all points (x, y) which are interior to an odd total number of loops. The function $f^*(x, y)$ is assumed to be exactly known on B and $z_i = f^*(x_i, y_i)$ for $i = 1, \ldots, n$ are observations of f^* for locations (x_i, y_i) in Ω. Let $f(x, y)$ be the function with the following properties:

- It interpolates the known values for f^* in B and the z_i at the points (x_i, y_i) for $i = 1, \ldots, n$.
- It satisfies the *Poisson equation*

  $$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0 \tag{3.4.10}$$

  on B.
- It minimizes the *Laplace equation*

  $$J_\Omega(f) = \int_\Omega \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)^2 dxdy. \tag{3.4.11}$$

Then f is the function such that

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \rho, \tag{3.4.12}$$

where

$$\frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} = 0, \tag{3.4.13}$$

except at the points (x_i, y_i) for $i = 1, \dots, n$ and $\rho = 0$ on B. The proof of this statement can be read in Wood et al. (2008). The Laplace Equation (3.4.11) and the Poisson Equation (3.4.10) are among the most well studied in the theory of partial differential equations due to their frequent occurence in physics. The same construction as described above can be used for bounded domains in $\mathbb{R}^k$ for any positive integer k. In this thesis, we restrict ourselves to the two-dimensional case.

It must be pointed out that for one-dimensional regions, the SOAP film penalty (3.4.9) equals the cubic spline penalty (3.2.1). According to (3.4.12) and (3.4.13), the SOAP film interpolant of (x_i, y_i) on the interval $[a, b]$ is defined in the one-dimensional case by

$$\frac{\partial^2 f}{\partial x^2} = \rho, \qquad \text{where} \quad \frac{\partial^2 \rho}{\partial x^2} = 0,$$

except at the points x_i where $f(x_i) = z_i$ for $i = 1, \dots, n$. Furthermore, the boundary conditions are that $f(a)$ and $f(b)$ are known and $f''(a) = f''(b) = 0$ holds. Thus, the second derivative of f, ρ, is piecewise linear with derivative discontinuities at the points x_i for $i = 1, \dots, n$. It follows that f is a natural cubic spline interpolant.

Wood et al. (2008) also proved the theorem of SOAP film smoothing. For this theorem, let the setup be exactly as in the SOAP film interpolant theorem, except that the values z_i are now measured with error. Let $\boldsymbol{f} = (f(x_1, y_1), \dots, f(x_n, y_n))^\mathsf{T}$, then the function $f(x, y)$ which minimizes

$$\|\boldsymbol{z} - \boldsymbol{f}\|^2 + \lambda J_\Omega(f) \tag{3.4.14}$$

subject to the known conditions on B, must satisfy (3.4.12) and (3.4.13). This characterization is the key to the computation of the SOAP film smoothers.

As a next step, the smoother and the basis-penalty representation are constructed. Therefore denote $\rho_i(x, y)$ as the function which is zero on B, satisfies Property (3.4.13) in Ω except at the single point (x_i, y_i) and

$$\int_\Omega \rho_i(x, y) dx dy = 1.$$

Then every function $\rho(x, y)$ which satisfies Property (3.4.13) in Ω expect at the points (x_i, y_i) for $i = 1, \dots, n$ can be written as a linear combination of the ρ_i with coefficients γ_i because of the linearity of (3.4.13). This yields

$$\rho(x, y) = \sum_{i=1}^{n} \gamma_i \rho_i(x, y),$$

and for the penalty it follows that

$$J_\Omega = \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{S} \boldsymbol{\gamma},$$

where the matrix $\boldsymbol{S}$ is given by

$$\boldsymbol{S}_{ij} = \int_{\Omega} \rho_i(x,y)\rho_j(x,y)dxdy.$$

To write f in terms of γ_i for $i = 1, \ldots, n$, define $h(x,y)$ as the solution to (3.4.12) with $\rho(x,y) = 0$ for all x, y and subject to the boundary condition that $f(x,y)$ is known on B. Let $g_i(x,y)$ be the solution of (3.4.12) with $\rho(x,y) = \rho_i(x,y)$ and the boundary condition that f is zero on B. Then the SOAP film smoother can be respresented as

$$f(x,y) = h(x,y) + \sum_{i=1}^{n} \gamma_i g_i(x,y)$$

because of the linearity of (3.4.12). Due to the fact, that the Laplace and the Poisson equation are well studied problems, the computation of h, g_i, ρ_i, and thus the matrix $\boldsymbol{S}$, is computationally straightforward.

To sum up, the minimization criterion (3.4.14) becomes the standard penalized regression problem of minimizing

$$\|\boldsymbol{z} - \boldsymbol{h} - \boldsymbol{Z}\boldsymbol{\gamma}\|^2 + \lambda \boldsymbol{\gamma}^{\mathsf{T}} \boldsymbol{S} \boldsymbol{\gamma}$$

with respect to $\boldsymbol{\gamma}$. The design matrix $\boldsymbol{Z}$ is defined as $\boldsymbol{Z}_{ij} = g_j(x_i, y_i)$ and the vector $\boldsymbol{h} = (h(x_1,y_1), \ldots, h(x_n,y_n))^{\mathsf{T}}$. The procedure of choosing the smoothing parameter λ is the same as already mentioned in Section 3.3.

There may be applications where the data lies in a problematic boundary region, but we do not have specific knowledge about the value of the function on that boundary. In this case, the values of the function on the boundary have to be modeled. To do so, assume that B consists only of a single closed loop. The loop should be parameterized in terms of r, where r is the distance along the loop from some fixed starting point on B. Thus, the coordinates of B are given as $(x_B(r), y_B(r))$. Let $f_b(r) = f(x_B(r), y_B(r))$ be the boundary function. Suppose that $f_b(r)$ has the basis representation

$$f_b(r) = \sum_{j=1}^{J} \alpha_j \nu_j(r) \tag{3.4.15}$$

for some parameters α_j and known basis functions $\nu_j(r)$. This representation is a cyclic penalized regression spline smoother in r, that is a function that has the same value and first few derivatives at its upper and lower boundaries. The penalty functional

$$J_b(f_b) = \boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{S}_b \boldsymbol{\alpha}$$

should be associated with the smoother, where $\boldsymbol{S}_b$ is a smoother matrix of known coefficients. It can then be shown that the function $h(x,y)$ has the representation

$$h(x,y) = \sum_{j=1}^{J} \alpha_j h_j(x,y),$$

where $h_j(x, y)$ is the solution of (3.4.12) with $\rho(x, y) = 0$ and the boundary condition resulting from setting $\alpha_j = 1$ and $\alpha_i = 0$ for $i \neq j$ in (3.4.15). The calculation of the $h_j(x, y)$ is easy, using the same method as for the $g_i(x, y)$. Details of the computation can be read in Wood et al. (2008).

Finally, given the basis functions and penalty for the boundary model, the criterion which has to be minimized with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, can be rewritten as

$$||\boldsymbol{z} - \boldsymbol{A}\boldsymbol{\alpha} - \boldsymbol{Z}\boldsymbol{\gamma}||^2 + \lambda\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{S}\boldsymbol{\gamma} + \lambda_b\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{S}_b\boldsymbol{\alpha},$$

where $\boldsymbol{A}_{ij} = h_j(x_i, y_i)$. The smoothing parameters λ and λ_b can be selected by an appropriate criterion, like the AIC or the GCV. The restriction to the case where B is a single closed loop can be generalized without any extra difficulties. For each loop of B a cyclic smooth can be used, each with an associated penalty. Thus, the SOAP film smooths are in practice very convenient because of their computational efficieny. Since the model can be evaluated with basis functions and a quadratic penalty for the smoothers, the computational and theoretical machinery, which is already available, can be used.

3.5 Higher-Dimensional Smoothing

The ideas of bivariate smoothing can also be applied to higher-dimensional regression problems. The models of higher-dimensional surfaces have the form

$$y_i = f(z_{i1}, \ldots, z_{iq}) + \epsilon_i \qquad \text{for } i = 1, \ldots, n, \tag{3.5.1}$$

for covariates $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q$. These models are used when there is interaction between the covariate values $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q$. Higher-dimensional tensor product splines result for example when considering all possible interactions of the univariate splines for $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q$. The penalty matrices for model (3.5.1) can be constructed as in the bivariate case with the Kronecker product. For the special case $q = 3$ the penalty matrix has the form

$$\boldsymbol{I}_{d_3} \otimes \boldsymbol{I}_{d_2} \otimes \boldsymbol{K}_1 + \boldsymbol{I}_{d_3} \otimes \boldsymbol{K}_2 \otimes \boldsymbol{I}_{d_1} + \boldsymbol{K}_3 \otimes \boldsymbol{I}_{d_2} \otimes \boldsymbol{I}_{d_1},$$

where $\boldsymbol{K}_1$, $\boldsymbol{K}_2$ and $\boldsymbol{K}_3$ are the univariate penalty matrices for the covariates $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$. When considering radial bases, the construction of the higher-dimensional surface is even simpler, since the basis functions only depend on the Euclidean distance between two points, which is defined in $\mathbb{R}^q$ as well as in $\mathbb{R}^2$.

Nevertheless, higher-dimensional functions which are estimated nonparametrically rely on a large number of parameters and therefore, the estimation is computationally expensive. Furthermore, a good coverage of the observation domain is only obtained when the number of parameters for each dimension is larger than what is used in the univariate smoothing problem. This problem is called the *curse of dimensionality* and means that neighbourhoods with a fixed number of points become less local as the dimensions increase. By doing so, the problem of a large number of parameters that need to be estimated occurs again. Consequently, it is advisable not to use surfaces with dimension higher than $q = 2$. In practice, it is common to use an additive structure for the function f, which means that

$$f(z_{i1}, \ldots, z_{iq}) = f_1(z_{i1}) + \cdots + f_q(z_{iq}) \qquad \text{for } i = 1, \ldots, n.$$

This yields the class of additive models, which will be discussed in detail in the next section.

3.6 Additive Model

In the previous sections, various methods were introduced for flexibly modeling the effect of a continuous covariate z on the response variable y without a restrictive functional form of the effect $f(z)$. The methods were extended to the bivariate and the higher-dimensional case with continuous covariates $z_1, \ldots, z_q$. However, there are cases where the effect of the continuous covariates $z_1, \ldots, z_q$ on the response y cannot be described by a simple functional form. Therefore, it is reasonable to model these effects in form of a function $f(z_1, \ldots, z_q)$. As already seen in the previous section, the estimation of such high-dimensional functions is problematic. Consequently, a more restrictive additive structure for the effect of the covariates is used in practice.

Consider observations $(y_i, x_{i1}, \ldots, x_{ik})$ for $i = 1, \ldots, n$ of a continuous response vector $\boldsymbol{y}$ and covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$, and observations $(z_{i1}, \ldots, z_{iq})$ for $i = 1, \ldots, n$ of additional continuous covariates $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q$. The effects of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ on $\boldsymbol{y}$ are modeled through a linear predictor, whereas the effects of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q$ on $\boldsymbol{y}$ are modeled nonparametrically. An *Additive Model* is defined by Fahrmeir et al. (2013) as

$$\begin{aligned} y_i &= f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad (3.6.1) \\ &= f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \eta_i^{lin} + \epsilon_i \\ &= \eta_i^{add} + \epsilon_i \end{aligned}$$

with

$$\eta_i^{lin} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \quad \text{and} \quad \eta_i^{add} = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \eta_i^{lin},$$

and $i = 1, \ldots, n$. The error terms ϵ_i are assumed to be independent and identically distributed with zero mean and variance σ^2. In the special case where the linear predictor η_i^{lin} is zero, model (3.6.1) is called *partial linear model* or *semiparametric model.* The advantage of model (3.6.1) is that it is additive in the predictor effects, which means that once the model is fitted, the predictor effects can be examined separately. However, the additive model has an identification problem, since if we add a constant $a \neq 0$ to the function $f_1(z_{i1})$ and subtract the same constant at the same time from the second function $f_2(z_{i2})$, the sum

$$f_1(z_{i1}) + f_2(z_{i2}) = f_1(z_{i1}) + a + f_2(z_{i2}) - a$$

and also the predictor η_i^{add}, remain the same. Hence, it is necessary to impose further restrictions by fixing the level of the functions. This is usually done by "centering the functions around zero", which means that the restrictions

$$\sum_{i=1}^{n} f_1(z_{i1}) = \cdots = \sum_{i=1}^{n} f_q(z_{iq}) = 0$$

are imposed. It must be pointed out that the functions $f_1, \ldots, f_q$ in (3.6.1) need not to be univariate and smooth. They can also be higher-dimensional, as well as categorical variable terms and there might be interactions as well. These generalizations are discussed in Section 3.7. This section focuses on the case, where the functions $f_1, \ldots, f_q$ are univariate and smooth.

For the estimation of the additive model, each function f_j for $j = 1, \ldots, q$ can be approximated by

$$f_j(z_j) = \sum_{l=1}^{d_j} \gamma_{jl} B_l(z_j)$$

with basis functions B_l. The basis functions can be B-splines, smoothing splines or other basis functions and it is also possible to choose different types of basis functions for the different functions f_j. Each vector $\boldsymbol{f}_j = (f_j(z_{1j}), \ldots, f_j(z_{nj}))^\mathsf{T}$ can then be written as

$$\boldsymbol{f}_j = \boldsymbol{Z}_j \boldsymbol{\gamma}_j,$$

where $\boldsymbol{\gamma_j} = (\gamma_{j1}, \ldots, \gamma_{jd_j})^\mathsf{T}$ is the vector of regression coefficients and the design matrix $\boldsymbol{Z}_j$ is defined as

$$\boldsymbol{Z}_j[i, l] = B_l(z_{ij}).$$

Thus, model (3.6.1) can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{3.6.2}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\mathsf{T}$ is the vector of the response values, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\mathsf{T}$ the vector of regression coefficients of the linear part of the predictor with corresponding design matrix $\boldsymbol{X}$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\mathsf{T}$ is the vector of the error terms. If the coefficient vectors $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q$ are estimated unrestrictedly using ordinary least-squares, the model (3.6.2) would just be a large linear model. However, the estimation is in general done by introducing penalties of the form $\lambda_j \boldsymbol{\gamma}_j^\mathsf{T} \boldsymbol{K}_j \boldsymbol{\gamma}_j$ for each function $\boldsymbol{f}_j$, to ensure smoothness properties of the estimates. This procedure is explained later in the case of structured additive regression models.

3.6.1 Geoadditive Model

The data for geoadditive models is in general given through observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ for $i = 1, \dots, n$, where $\boldsymbol{y}$ is the response vector and $\boldsymbol{x}$, $\boldsymbol{z}$ the covariates. Furthermore, there are additional values s_i for $i = 1, \dots, n$, which are associated with a geographic location index s. There are two different cases for the location index s. On the one hand, s can be discrete by taking the values $s \in \{1, \dots, d\}$. In this case, s_i denotes a specific region, in which the individual or the unit i was observed. These specific regions can be, for example, districts or postcode areas. On the other hand, s can be a continuous variable given by the coordinates of $\mathbb{R}^2$, which means that for every unit i, the corresponding location s_i is exactly known and given as a pair of longitude and latitude. The geoadditive model has the form

$$\begin{aligned} y_i &= \eta_i^{add} + f_{geo}(s_i) + \epsilon_i \qquad (3.6.3) \\ &= f_1(z_{i1}) + \dots + f_q(z_{iq}) + f_{geo}(s_i) + \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i, \end{aligned}$$

for $i = 1, \dots, n$. Thus, additional to the assumptions for the additive model, the predictor is expanded by a spatial effect $f_{geo}(s_i)$ of the location variable s_i for $i = 1, \dots, n$. Fahrmeir et al. (2013) described the spatial effect $f_{geo}(s)$ as a surrogate for unobserved spatial variables not included in the data. Depending on whether s is continuous or discrete, different methods for estimating the function $f_{geo}(s)$ can be applied. For discrete s, the property $s_i = s$ implies that the ith observation belongs to region s with $s \in \{1, \dots, d\}$. In this case, $f_{geo}(s_i)$ denotes the spatial effect of region $s = s_i$.

In geoadditve models, the vector $\boldsymbol{f}_{geo} = (f_{geo}(s_1), \dots, f_{geo}(s_n))^\top$ of the spatial effect can be represented as

$$\boldsymbol{f}_{geo} = \boldsymbol{Z}_{geo} \boldsymbol{\gamma}_{geo}$$

with the corresponding vector of regression coefficients $\boldsymbol{\gamma}_{geo} = (\gamma_{geo,1}, \dots, \gamma_{geo,d})^\top$. The $n \times d$ design matrix $\boldsymbol{Z}_{geo}$ is an incidence matrix, which means that

$$\boldsymbol{Z}_{geo}[i, s] = \begin{cases} 1 & \text{if } s_i = s \\ 0 & \text{otherwise.} \end{cases}$$

The geoadditive model can thus be represented in matrix notation as

$$\boldsymbol{y} = \boldsymbol{Z}_1 \boldsymbol{\gamma}_1 + \dots + \boldsymbol{Z}_q \boldsymbol{\gamma}_q + \boldsymbol{Z}_{geo} \boldsymbol{\gamma}_{geo} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

which is just the additive model (3.6.2) extended by the additional spatial effect $\boldsymbol{Z}_{geo}\boldsymbol{\gamma}_{geo}$. The variable selection and model choice in geoadditive regression models is discussed in more detail in Kneib, Hothorn, and Tutz (2009).

3.6.2 Model with Interactions

The additive and the geoadditive model are purely main effect models with nonparametric functions $f_1, \dots, f_q$ and an optional spatial effect f_{geo}. However, it is also possible that there is interaction between two ore more covariates, so that a purely main effect model will not be sufficient. According to Fahrmeir et al. (2013), there are different types of interaction terms. The most common type of interaction is the nonparametric interaction between two continuous covariates z_1 and z_2. This interaction can be modeled by adding a smooth function $f_{1|2}(z_1, z_2)$ to the predictor. This can be done in two different ways.

On the one hand, the effect of z_1 and z_2 can be considered by the function $f_{1|2}(z_1, z_2)$, but the one-dimensional main effects $f_1(z_1)$ and $f_2(z_2)$ are not included in the model. This yields

$$y_i = f_{1|2}(z_{i1}, z_{i2}) + f_3(z_{i3}) + \cdots + f_q(z_{iq}) + \eta_i^{lin} + \epsilon_i \qquad \text{for } i = 1, \ldots, n. \tag{3.6.4}$$

As in the additive model without interaction terms, the level of $f_{1|2}(z_1, z_2)$ is not identified and thus the function has to be centered around zero, which means that it has to be assumed that

$$\sum_{i=1}^{n} f_{1|2}(z_{i1}, z_{i2}) = 0.$$

When using model (3.6.4), it must be assured that the existing dataset contains enough data combinations of z_1 and z_2 to estimate a two-dimensional surface. Otherwise, it is advisable to use a pure main effects model.

On the other hand, a two-dimensional function $f_{1|2}(z_1, z_2)$ can be added to the predictor in addition to the main effect functions $f_1(z_1)$ and $f_2(z_2)$. This results in the model

$$y_i = f_1(z_{i1}) + f_2(z_{i2}) + f_{1|2}(z_{i1}, z_{i2}) + f_3(z_{i3}) + \cdots + f_q(z_{iq}) + \eta_i^{lin} + \epsilon_i, \tag{3.6.5}$$

for $i = 1, \ldots, n$. In this case, the problem of identifiability is more complicated. Identifiability is guaranteed if in addition to centering all functions $f_i(z_i)$ around zero, all interaction terms $f_{1|2}(z_{i1}, z_{i2})$ are centered around zero as well. By doing so, the interaction $f_{1|2}(z_1, z_2)$ can be seen as a deviation from the main effects.

In both cases (3.6.4) and (3.6.5), the interaction term $f_{1|2}(z_1, z_2)$ is generally modeled with tensor product P-splines or radial bases. When the interaction vector is defined as

$$\boldsymbol{f}_{1|2} = (f_{1|2}(z_{11}, z_{12}), \ldots, f_{1|2}(z_{n1}, z_{n2}))^{\mathsf{T}},$$

the large linear model can be written as

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{Z}_{1|2}\boldsymbol{\gamma}_{1|2} + \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{Z}_{1|2}$ is the design matrix of the interaction term with corresponding coefficient vector $\boldsymbol{\gamma}_{1|2}$, so that

$$\boldsymbol{f}_{1|2} = \boldsymbol{Z}_{1|2}\boldsymbol{\gamma}_{1|2}.$$

The coefficients $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q, \boldsymbol{\gamma}_{1|2}$ can then be estimated using the penalized least-squares method described in the next section in case of the structured additive models.

Another way of possible interaction, is the one between a continuous variable z_1 and a binary variable $x_1 \in \{0, 1\}$, modeled through the interaction term $f_{z_1|x_1}(z_1) \cdot x_1$. This yields the model

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + f_{z_1|x_1}(z_{i1})x_{i1} + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \qquad (3.6.6)$$

where except of the added interaction term, the model is the same as the additive model (3.6.1). Once again, the function $f_{z_1|x_1}$ has to be centered around zero, otherwise the model is not identifiable. Model (3.6.6) is then called a *Varying Coefficients Model.* If the binary variable $x_1 = 0$, then $f_1(z_1)$ is the nonlinear effect of z_1, whereas if $x_1 = 1$, then $f_1(z_1) + f_{z_1|x_1}(z_1) + \beta_1$ is the nonlinear effect of z_1. The sum $f_{z_1|x_1}(z_1) + \beta_1$ is called the *varying effect* for $x_1 = 1$ depending on z_1. The variable z_1 in this case is also called the *effect modifier* of x_1, and x_1 is called the *interaction variable.*

The same procedure can be done for a three-level categorical variable $x \in \{1, 2, 3\}$, which is coded by the dummy variables x_1 and x_2 as

$$(x_1, x_2) = \begin{cases} (1, 0) & \text{if } x = 1 \\ (0, 1) & \text{if } x = 2 \\ (0, 0) & \text{if } x = 3. \end{cases}$$

This results in the model

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + f_{z_1|x_1}(z_{i1})x_{i1} + f_{z_1|x_2}(z_{i1})x_{i2} + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i.$$

The interpretation is similar to the one before. The effect of z_1 is given as $f_1(z_1) + f_{z_1|x_1}(z_1) + \beta_1$ if $x = 1$, and as $f_1(z_1) + f_{z_1|x_2}(z_1) + \beta_2$ if $x = 2$, whereas the nonlinear effect of z_1 is given as $f_1(z_1)$ if $x = 3$. However, we will restrict ourselves to the case of one binary variable x_1. For representing the varying coefficients model as a large linear model, the functions $f_1, \dots, f_q, f_{z_1|x_1}$ can be modeled with the basis function approach as

$$f_{z_1|x_1}(z_1) = \sum_{l=1}^{d} \gamma_{int,l} B_l(z_1) \quad \text{and} \quad f_i(z_i) = \sum_{l=1}^{d} \gamma_l B_l(z_i) \qquad \text{for } i = 1, \dots, q$$

with the vector of coefficients $\boldsymbol{\gamma}_{int} = (\gamma_{int,1}, \dots, \gamma_{int,d})^\mathsf{T}$. Therefore, the interaction term $\boldsymbol{f}_{int} = (f_{z_1|x_1}(z_{11})x_{11}, \dots, f_{z_1|x_1}(z_{n1})x_{n1})^\mathsf{T}$ can be represented as

$$\boldsymbol{f}_{int} = \boldsymbol{Z}_{int} \boldsymbol{\gamma}_{int},$$

where the design matrix $\boldsymbol{Z}_{int}$ is defined by

$$\boldsymbol{Z}_{int}[i, l] = B_l(z_{i1})x_{i1}.$$

Thus, the varying coefficients model can be written as

$$\boldsymbol{y} = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{Z}_{int}\boldsymbol{\gamma}_{int} + \boldsymbol{X\beta} + \boldsymbol{\epsilon}.$$

The concept of varying coefficients is also useful in geoadditive models. In this case, an interaction term $f_{geo|z}(s) \cdot z$ between the location variable s and a binary or even multicategorical variable z is included in the geoadditive model (3.6.3). This model is then called a *model with spatially varying coefficients* or *geographically weighted regression model.*

3.6.3 Structured Additive Regression Model

The various terms in a model discussed in the previous sections can also be combined into one model. This means, a model can include all main effects and spatial effects, as well as interaction terms. According to Fahrmeir et al. (2013), the resulting model is called a *Structured Additive Regression (STAR) Model* and has the form

$$y_i = f_1(v_{i1}) + \cdots + f_q(v_{iq}) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad \text{for } i = 1, \ldots, n, \quad (3.6.7)$$

where $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$ are n-dimensional covariates of different types, constructed from the original variables. For example, $f_1(v_{i1}) = f_1(z_{i1})$ can be a nonlinear effect of $\boldsymbol{z}_1$, $f_2(v_{i2}) = f_{geo}(s_i)$, $f_3(v_{i3}) = f_{1|2}(z_{i1}, z_{i2})$, and so on.

Defining the vectors $\boldsymbol{f}_j = (f_j(v_{1j}), \ldots, f_j(v_{nj}))^\mathsf{T}$ of functions evaluated at the observations of the covariates $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$, it follows that

$$\boldsymbol{f}_j = \boldsymbol{V}_j\boldsymbol{\gamma}_j,$$

and thus, the model (3.6.7) can be written in matrix notation as

$$\boldsymbol{y} = \boldsymbol{V}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \quad (3.6.8)$$

for suitable design matrices $\boldsymbol{V}_j$ and corresponding coefficient vectors $\boldsymbol{\gamma}_j$. For simplicity, we write $LS = LS(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_q, \boldsymbol{\beta})$ for the penalized least-squares criterion, which is given by

$$LS = (\boldsymbol{y} - \boldsymbol{V}_1\boldsymbol{\gamma}_1 - \cdots - \boldsymbol{V}_q\boldsymbol{\gamma}_q - \boldsymbol{X\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{V}_1\boldsymbol{\gamma}_1 - \cdots - \boldsymbol{V}_q\boldsymbol{\gamma}_q - \boldsymbol{X\beta}) + \sum_{j=1}^{d} \lambda_j \boldsymbol{\gamma}_j^\mathsf{T}\boldsymbol{K}_j\boldsymbol{\gamma}_j. \quad (3.6.9)$$

The penalty matrices $\boldsymbol{K}_j$ depend on the specific type of the function f_j and the chosen penalty. Thus, a STAR model is characterized by the matrices $\boldsymbol{V}_j$ and the penalty matrices $\boldsymbol{K}_j$.

For the minimization of (3.6.9), we first assume that λ_j is known for all $j = 1, \ldots, n$. The estimation of the smoothing parameters λ_j will be discussed later. There are two different ways to minimize (3.6.9): iterative minimization with the backfitting alorigthm or direct minimization of the penalized least-squares criterion. The idea of the backfitting algorithm is to iteratively estimate $\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_q, \hat{\boldsymbol{\beta}}$ by smoothing of some partial residuals. From (3.6.8) it follows that

$$\boldsymbol{f}_j \approx \boldsymbol{y} - \boldsymbol{f}_1 - \cdots - \boldsymbol{f}_{j-1} - \boldsymbol{f}_{j+1} - \cdots - \boldsymbol{f}_q - \boldsymbol{X\beta},$$

when the errors $\boldsymbol{\epsilon}$ are neglected. If we already have estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{f}}_l$ for $l \neq j$, then

$$\boldsymbol{y} - \sum_{l \neq j} \hat{\boldsymbol{f}}_l - \boldsymbol{X\beta}$$

can be interpreted as a residual vector without $\hat{\boldsymbol{f}}_j$. To obtain an estimate for $\boldsymbol{f}_j$, the smoother $\boldsymbol{S}_j(\lambda_j)$ has to be applied to the residual vector. This yields

$$\hat{\boldsymbol{f}}_j = \boldsymbol{S}_j(\lambda_j)(\boldsymbol{y} - \sum_{l \neq j} \hat{\boldsymbol{f}}_l - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{3.6.10}$$

The updated $\hat{\boldsymbol{\beta}}$ is defined as

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{y} - \hat{\boldsymbol{f}}_1 - \cdots - \hat{\boldsymbol{f}}_q). \tag{3.6.11}$$

When using basis function approaches, all estimators can be respresented as

$$\hat{\boldsymbol{f}}_j = \boldsymbol{V}_j(\boldsymbol{V}_j^\mathsf{T}\boldsymbol{V}_j + \lambda_j\boldsymbol{K}_j)^{-1}\boldsymbol{V}_j^\mathsf{T}(\boldsymbol{y} - \sum_{l \neq j} \hat{\boldsymbol{f}}_l - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

This estimation was already derived in Section 3.2 in the case of univariate smoothing. In the case of STAR models, the form of $\boldsymbol{V}_j$ and the penalty matrix $\boldsymbol{K}_j$ depend on the type of smoother used. The backfitting algorithm works as follows: Firstly, the values $\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_q$, $\hat{\boldsymbol{\beta}}$ are initialized with some starting values. Then the estimates for $\hat{\boldsymbol{f}}_j$ and $\hat{\boldsymbol{\beta}}$ are updated using (3.6.10) and (3.6.11), until the estimated functions do not differ more than a small given increment in two subsequent iterations.

Another way to compute the penalized least-squares estimator is through direct minimization of Criterion (3.6.9) with a non-iterative procedure. To minimize the penalized least-squares criterion, the first derivatives of the penalized least-squares criterion have to be computed with respect to the unknown parameters. The first derivatives are

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\gamma}_1} LS &= -\boldsymbol{V}_1^\mathsf{T}\boldsymbol{y} + (\boldsymbol{V}_1^\mathsf{T}\boldsymbol{V}_1 + \lambda_1\boldsymbol{K}_1)\boldsymbol{\gamma}_1 + \boldsymbol{V}_1^\mathsf{T}\boldsymbol{V}_2\boldsymbol{\gamma}_2 + \cdots + \boldsymbol{V}_1^\mathsf{T}\boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{V}_1^\mathsf{T}\boldsymbol{X\beta},\\
\frac{\partial}{\partial \boldsymbol{\gamma}_2} LS &= -\boldsymbol{V}_2^\mathsf{T}\boldsymbol{y} + \boldsymbol{V}_2^\mathsf{T}\boldsymbol{V}_1\boldsymbol{\gamma}_1 + (\boldsymbol{V}_2^\mathsf{T}\boldsymbol{V}_2 + \lambda_2\boldsymbol{K}_2)\boldsymbol{\gamma}_2 + \boldsymbol{V}_2^\mathsf{T}\boldsymbol{V}_3\boldsymbol{\gamma}_3 + \cdots + \boldsymbol{V}_2^\mathsf{T}\boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{V}_2^\mathsf{T}\boldsymbol{X\beta},\\
&\vdots\\
\frac{\partial}{\partial \boldsymbol{\gamma}_q} LS &= -\boldsymbol{V}_q^\mathsf{T}\boldsymbol{y} + \boldsymbol{V}_q^\mathsf{T}\boldsymbol{V}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{V}_q^\mathsf{T}\boldsymbol{V}_{q-1}\boldsymbol{\gamma}_{q-1} + (\boldsymbol{V}_q^\mathsf{T}\boldsymbol{V}_q + \lambda_q\boldsymbol{K}_q)\boldsymbol{\gamma}_q + \boldsymbol{V}_q^\mathsf{T}\boldsymbol{X\beta},\\
\frac{\partial}{\partial \boldsymbol{\beta}} LS &= -\boldsymbol{X}^\mathsf{T}\boldsymbol{y} + \boldsymbol{Z}^\mathsf{T}\boldsymbol{V}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{Z}^\mathsf{T}\boldsymbol{V}_q\boldsymbol{\gamma}_q + \boldsymbol{X}^\mathsf{T}\boldsymbol{X\beta}.
\end{aligned}$$

Setting these derivatives to zero, yields the system of equations

$$\begin{pmatrix} \boldsymbol{V}_1^\mathsf{T}\boldsymbol{V}_1 + \lambda_1\boldsymbol{K}_1 & \boldsymbol{V}_1^\mathsf{T}\boldsymbol{V}_2 & \dots & \boldsymbol{V}_1^\mathsf{T}\boldsymbol{V}_q & \boldsymbol{V}_1^\mathsf{T}\boldsymbol{X} \\ \boldsymbol{V}_2^\mathsf{T}\boldsymbol{V}_1 & \boldsymbol{V}_2^\mathsf{T}\boldsymbol{V}_2 + \lambda_2\boldsymbol{K}_2 & \dots & \boldsymbol{V}_2^\mathsf{T}\boldsymbol{V}_q & \boldsymbol{V}_2^\mathsf{T}\boldsymbol{X} \\ \vdots & & \ddots & \vdots & \vdots \\ \boldsymbol{V}_q^\mathsf{T}\boldsymbol{V}_1 & \dots & \boldsymbol{V}_q^\mathsf{T}\boldsymbol{V}_{q-1} & \boldsymbol{V}_q^\mathsf{T}\boldsymbol{V}_q + \lambda_q\boldsymbol{K}_q & \boldsymbol{V}_q^\mathsf{T}\boldsymbol{X} \\ \boldsymbol{Z}^\mathsf{T}\boldsymbol{V}_1 & \dots & \boldsymbol{Z}^\mathsf{T}\boldsymbol{V}_{q-1} & \boldsymbol{Z}^\mathsf{T}\boldsymbol{V}_q & \boldsymbol{Z}^\mathsf{T}\boldsymbol{X} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_q \\ \beta \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_1^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{V}_2^\mathsf{T}\boldsymbol{y} \\ \vdots \\ \boldsymbol{V}_q^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{y} \end{pmatrix}.$$

For small q and a relatively modest number of knots used, for example when modeling $\boldsymbol{f}_1, \dots, \boldsymbol{f}_q$ with P-splines, the direct estimation of the above system of equations works well. If $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_q$ are high-dimensional, the solution of the system of equations has to be solved iteratively, for example with the *Gauß-Seidel algorithm*. This algorithm is not described here, but can be read in Buja, Hastie, and Tibshirani (1989).

The estimation of the model coefficients only works, if the smoothing parameters λ_j are known. There are two common approaches: the estimation with the *Un-Biased Risk Estimator (UBRE)* or the *Generalized Cross Validation (GCV)*. The decision on the estimation criterion is made based on the scale parameter ϕ, which equals the variance σ^2 in the normal distribution. The smoothing parameter is estimated by minimizing the UBRE when σ^2 is known, whereas the GCV is used when it is unknown.

Let us first assume that the scale parameter is known. One possible way to choose the smoothing parameter $\boldsymbol{\lambda}$ would be to choose it in a way that the estimator $\hat{\boldsymbol{\mu}}$ is as close as possible to the true mean $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{y})$. To do so, an appropriate measure might be the expected Mean Squared Error (MSE) of the model, which is defined by Wood (2006) as

$$\mathrm{E}(M) = \mathrm{E}\left(\|\boldsymbol{\mu} - \boldsymbol{V}_1\hat{\gamma}_1 - \dots - \boldsymbol{V}_q\hat{\gamma}_q - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/n\right).$$

Using the smoother matrix $\boldsymbol{S}(\boldsymbol{\lambda})$ and $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, it follows that

$$\begin{aligned} M &= \|\boldsymbol{\mu} - \boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}\|^2/n = \|\boldsymbol{y} - \boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y} - \boldsymbol{\epsilon}\|^2/n \\ &= \frac{1}{n}(\|\boldsymbol{y} - \boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}\|^2 + \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^\mathsf{T}(\boldsymbol{y} - \boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y})) \\ &= \frac{1}{n}(\|\boldsymbol{y} - \boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}\|^2 - \boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon} - 2\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\mu} + 2\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{\epsilon}). \end{aligned}$$

When calculating the expected value of the last term, we obtain

$$\begin{aligned} \mathrm{E}(\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{\epsilon}) &= \mathrm{E}(\mathrm{tr}(\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{\epsilon})) = \mathrm{E}(\mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathsf{T})) = \mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda})\,\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\mathsf{T})) \\ &= \mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{I})\sigma^2 = \mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda}))\sigma^2, \end{aligned}$$

thus, for $\mathrm{E}(M)$ it follows that

$$\begin{aligned}\mathrm{E}(M) &= \frac{1}{n}\left(\mathrm{E}(\|\boldsymbol{y}-\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}\|^2 - \mathrm{E}(\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{\epsilon}) - 2\,\mathrm{E}(\boldsymbol{\epsilon}^\mathsf{T})\boldsymbol{\mu} + 2\,\mathrm{E}(\boldsymbol{\epsilon}^\mathsf{T}\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{\epsilon})\right)\\ &= \frac{1}{n}\left(\mathrm{E}(\|\boldsymbol{y}-\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}\|^2 - \mathrm{E}\left(\sum_{i=1}^n \epsilon_i^2\right) + 2\,\mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda}))\sigma^2\right)\\ &= \mathrm{E}(\|\boldsymbol{y}-\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}\|^2)/n - \sigma^2 + 2\,\mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda}))\sigma^2/n.\end{aligned}$$

For choosing the smoothing parameters, an estimate of this expected MSE is minimized, namely the *Un-Biased Risk Estimator (UBRE)* or *Mallow's C_p* (Mallows, 1973)

$$\nu_u(\boldsymbol{\lambda}) = \|\boldsymbol{y}-\boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}\|^2/n - \sigma^2 + 2\,\mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda}))\sigma^2/n.$$

For known σ^2, the estimation of $\boldsymbol{\lambda}$ with the UBRE works very well. Problems occur when σ^2 has to be estimated.

For an unknown scale parameter, the smoothing parameter is estimated with the *Generalized Cross Validation (GCV)*, which works with the mean squared prediction error instead of the mean squared error. The GCV in the univariate case, discussed in Section 3.3, can be generalized to

$$GCV = \frac{n\|\boldsymbol{y}-\hat{\boldsymbol{\mu}}\|^2}{(n-\mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda}))^2}.$$

The derivation works analogously to the univariate case and will therefore not be discussed in detail here. This derivation and details about the minimization of the UBRE and the GCV can be read in Wood (2006).

Another way to estimate the smoothing parameter is to minimize the AIC criterion, which is given for STAR models with Gaussian errors by

$$AIC = n\log(\hat{\sigma}^2) + 2(\mathrm{tr}(\boldsymbol{S}(\boldsymbol{\lambda})) + 1),$$

where $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{\eta}_i)^2/n$ is the maximum likelihood estimator of the error variance and $\boldsymbol{S}(\boldsymbol{\lambda})$ is the smoother matrix of the full STAR model. This criterion works for Gaussian errors since in this case $\hat{\boldsymbol{y}} = \boldsymbol{S}(\boldsymbol{\lambda})\boldsymbol{y}$. For non-Gaussian errors, the corresponding matrix from the model has to be used.

3.7 Generalized Additive Model

The *Generalized Additive Model (GAM)*, introduced by Hastie and Tibshirani (1990), follows from an structured additive model, as the generalized linear model follows from the classical linear model. That is, the predictor effects are still assumed to follow an additive structure, but the distribution of the responses as well as the link between the predictors and the mean can be quite general. Thus, the GAM has the form

$$g(\mu_i) = f_1(v_{i1}) + \cdots + f_q(v_{iq}) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \qquad \text{for } i = 1, \ldots, n,$$

where $\mu_i = \mathrm{E}(y_i)$ and the responses y_i are from some exponential family distribution. The function g is assumed to be a known, monotonic and twice differentiable link function. The functions $f_1, \ldots, f_q$ as well as the covariates $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ are defined as in the structured additive model (3.6.7). The generalized additive model therefore differs from a generalized linear model in that an additive predictor replaces the linear predictor. Many useful models fall into the class of generalized additive models, for example the logistic additive model, where the link function, when applied to binary response data, takes the form $g(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}/(1-\boldsymbol{\mu}))$.

3.7.1 The Penalized Iteratively Re-Weighted Least-Squares Procedure

For notational simplicity, we define the design matrix $\tilde{\boldsymbol{V}} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_q, \boldsymbol{X})$, containing all design matrices of the GAM. Furthermore, all regression coefficients are combined in one vector, namely $\boldsymbol{\delta}^\mathsf{T} = (\boldsymbol{\gamma}_1^\mathsf{T}, \ldots, \boldsymbol{\gamma}_q^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T})$. The estimation of the parameter vector $\boldsymbol{\delta}$ as well as $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_q$ can be done with the penalized likelihood function, defined by

$$l_p(\boldsymbol{\theta}(\boldsymbol{\delta})|\boldsymbol{y}) = l(\boldsymbol{\theta}(\boldsymbol{\delta})|\boldsymbol{y}) - \sum_{j=1}^{q} \lambda_j \boldsymbol{\delta}^\mathsf{T} \boldsymbol{K}_j \boldsymbol{\delta}, \tag{3.7.1}$$

where $l(\boldsymbol{\theta}(\boldsymbol{\delta})|\boldsymbol{y})$ denotes the log-likelihood function and the smoothing parameters λ_j again control the tradeoff between goodness of fit of the model and the smoothness. The estimaton of the smoothing parameters will be discussed later. In this section we assume that the smoothing parameters are known. In contrast to additive models, where the estimation was done with penalized least-squares, the GAMs are fitted by penalized likelihood maximization. To maximize the penalized likelihood function (3.7.1), Wood (2006) introduced the *Penalized Iteratively Re-Weighted Least Squares (P-IRLS) scheme*. For notational simplicity, define $\boldsymbol{K} = \sum_{j=1}^{q} \lambda_j \boldsymbol{K}_j$. For the P-IRLS algorithm, the derivatives of l_p with respect to δ_j have to be calculated and set to zero, since we want to maximize l_p. The log-likelihood function is defined by

$$l(\boldsymbol{\theta}(\boldsymbol{\delta})|\boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right).$$

and the derivative is

$$\begin{aligned}\frac{\partial l(\boldsymbol{\theta}(\boldsymbol{\delta})|\boldsymbol{y})}{\partial \delta_j} &= \sum_{i=1}^{n}\left(\frac{y_i - b'(\theta_i)}{\phi}\right)\frac{\partial \theta_i}{\partial \delta_j}\\ &= \sum_{i=1}^{n}\left(\frac{y_i - b'(\theta_i)}{\phi}\right)\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \delta_j},\end{aligned}$$

because of the chain rule. Furthermore, it follows that

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{\partial \theta_i}{\partial b'(\theta_i)} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}.$$

Thus, for the derivative of the penalized log-likelihood follows that

$$\frac{\partial l_p(\boldsymbol{\theta}(\boldsymbol{\delta})|\boldsymbol{y})}{\partial \delta_j} = \frac{\partial l(\boldsymbol{\theta}(\boldsymbol{\delta})|\boldsymbol{y})}{\partial \delta_j} - (\boldsymbol{K\delta})_j = \frac{1}{\phi}\sum_{i=1}^{n}\frac{y_i - \mu_i}{V(\mu_i)}\frac{\partial \mu_i}{\partial \delta_j} - (\boldsymbol{K\delta})_j = 0, \tag{3.7.2}$$

where $(\cdot)_j$ denotes the jth row of a vector. It must be pointed out that the weights $a(\phi)$ were set to ϕ, without loss of generality. The above equations are exactly those that would have to be solved to find $\hat{\boldsymbol{\delta}}$ in the penalized non-linear least-squares problem

$$S_p = \sum_{i=1}^{n}\frac{(y_i - \mu_i)^2}{\operatorname{var}(y_i)} + \boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{K\delta}, \tag{3.7.3}$$

where the variance of y_i is assumed to be known. This penalized least-squares problem is solved by calculating the derivatives with respect to δ_j, but the resulting system of equations is already given in (3.7.2). The P-IRLS is an iterative method. Assume we start with an estimated parameter $\hat{\boldsymbol{\delta}}^{[k]}$ at the kth iterate and vectors $\boldsymbol{\eta}^{[k]}$ and $\boldsymbol{\mu}^{[k]}$, calculated as $\eta_i^{[k]} = \tilde{\boldsymbol{V}}_i\hat{\boldsymbol{\delta}}^{[k]}$ and $\mu_i^{[k]} = g^{-1}(\eta_i^{[k]})$. Defining the diagonal matrix $\boldsymbol{V}_{[k]}$ with diagonal elements $\boldsymbol{V}_{[k]ii} = V(\mu_i^{[k]})$ yields

$$S_p = \|\sqrt{\boldsymbol{V}_{[k]}^{-1}}(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\delta}))\|^2 + \boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{K\delta}.$$

When replacing $\boldsymbol{\mu}$ by its first order Taylor expansion around $\hat{\boldsymbol{\delta}}^{[k]}$, S_p becomes

$$S_p \approx \|\sqrt{\boldsymbol{V}_{[k]}^{-1}}(\boldsymbol{y} - \boldsymbol{\mu}^{[k]} - \boldsymbol{J}(\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}^{[k]}))\|^2 + \boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{K\delta},$$

where $\boldsymbol{J}$ denotes the Jacobian matrix, given as $\boldsymbol{J}_{ij} = \frac{\partial \mu_i}{\partial \delta_j}\big|_{\hat{\boldsymbol{\delta}}^{[k]}}$. It follows that

$$g'(\mu_i)\frac{\partial \mu_i}{\partial \delta_j} = \tilde{\boldsymbol{V}}_{ij},$$

since $g(\mu_i) = \tilde{\boldsymbol{V}}_i\boldsymbol{\delta}$ and therefore,

$$\boldsymbol{J}_{ij} = \frac{\partial \mu_i}{\partial \delta_j}\bigg|_{\hat{\boldsymbol{\delta}}^{[k]}} = \frac{\boldsymbol{V}_{ij}}{g'(\mu_i^{[k]})}.$$

Defining $\boldsymbol{G}$ as a diagonal matrix with diagonal elements $\boldsymbol{G}_{ii} = g'(\mu_i^{[k]})$ yields that $\boldsymbol{J} = \boldsymbol{G}^{-1}\tilde{\boldsymbol{V}}$. Thus, the approximation for S_p becomes

$$\begin{aligned} S_p &\approx \|\sqrt{\boldsymbol{V}_{[k]}^{-1}}\boldsymbol{G}^{-1}\left[\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu}^{[k]}) + \boldsymbol{\eta}^{[k]} - \tilde{\boldsymbol{V}}\boldsymbol{\delta}\right]\|^2 + \boldsymbol{\delta}^\mathsf{T}\boldsymbol{K}\boldsymbol{\delta} \\ &= \|\sqrt{\boldsymbol{W}^{[k]}}(\boldsymbol{z}^{[k]} - \tilde{\boldsymbol{V}}\tilde{\boldsymbol{\beta}})\|^2 + \boldsymbol{\delta}^\mathsf{T}\boldsymbol{K}\boldsymbol{\delta}, \end{aligned} \tag{3.7.4}$$

where $\boldsymbol{z}^{[k]}$ is a vector of pseudodata given as

$$\boldsymbol{z}^{[k]} = g(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + \tilde{\boldsymbol{V}}_i\hat{\boldsymbol{\delta}}^{[k]},$$

and $\boldsymbol{W}^{[k]}$ is a diagonal matrix with diagonal elements

$$w_i^{[k]} = \frac{1}{V(\mu_i^{[k]})g'(\mu_i^{[k]})^2}.$$

To sum up, the P-IRLS method works as follows: For given estimates $\hat{\boldsymbol{\delta}}^{[k]}$ the pseudodata $\boldsymbol{z}^{[k]}$ and the weights $w_i^{[k]}$ have to be estimated. Using these values, S_p as given in (3.7.4) is minimized with respect to $\boldsymbol{\delta}$ to find $\hat{\boldsymbol{\delta}}^{[k+1]}$. The procedure is then again started with k replaced by $k+1$. These steps are iterated to convergence. The converged $\hat{\boldsymbol{\delta}}$ minimizes (3.7.3) and is thus the maximum likelihood estimate. It must be pointed out that only values for $\boldsymbol{\mu}^{[0]}$ and $\boldsymbol{\eta}^{[0]}$ have to be initialized at the beginning, but not for $\hat{\boldsymbol{\delta}}^{[0]}$. The method is usually started by setting $\mu_i^{[0]} = y_i$ and $\eta_i^{[0]} = g(\mu_i^{[0]})$ and the algorithm converges in most practical approaches. The P-IRLS is very similar to the IRLS algorithm for GLMs, where the algorithm works analogously just without the penalty term. The IRLS algorithm was introduced by Nelder and Wedderburn (1972) and the derivation of the P-IRLS can be read in McCullagh and Nelder (1989).

3.7.2 Estimation of the Smoothing Parameter

For estimating the smoothing parameter $\boldsymbol{\lambda}$ in GAMs, the minimization of the UBRE and the GCV from STAR models can be generalized. The GCV score in the generalized case is

$$GCV = \frac{n\sum_{i=1}^n V(\hat{\mu}_i)^{-1}(y_i - \hat{\mu}_i)^2}{(n - \operatorname{tr}(\boldsymbol{S}(\lambda))^2}$$

and the UBRE score

$$\nu_u(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^n V(\hat{\mu}_i)^{-1}(y_i - \hat{\mu}_i)^2 - \sigma^2 + \frac{2}{n}\operatorname{tr}(\boldsymbol{S}(\lambda))\sigma^2,$$

where $V(\mu_i)$ is the variance function, defined as in the GLM. Details about the derivation of the generalized UBRE and GCV can be read in Wood (2006). Algorithms and details regarding the minimization of the different criteria for estimating the smoothing parameters can be read in Gu and Wahba (1991), Wood (2000) and Wood (2004).

3.8 Model Choice and Diagnosis

3.8.1 Model Choice

This section focuses on how to carry out a regression analysis and to find a suitable model for a given dataset. As explained in Fahrmeir et al. (2013), the first step in every statistical application is the analysis of the dataset. It is important to get insight into the distribution and the behaviour of the variables. In this descriptive analysis of the data, the first step is the description of the distribution of the variables in the dataset. Furthermore, the data has to be checked for extreme values or limited information in some regions. Sometimes it is necessary to exclude extreme observations from the data. The problem of missing information is very common in the case of binary covariates when one category is not populated at all or only very sparsely. In this case it is often useful to combine individual categories or in the extreme case, a covariate may be useless for the analysis.

A next important step is the correlation analysis. This means that in case of continuous responses, graphical two-dimensional correlation analyses between the response values and each covariate gives an important insight in the type and the strength of the relationship. Important graphical tools for this step are scatter plots and box plots. It must be pointed out that we are only investigating two-dimensional correlations in this step, so that other correlations have to be considered in a further investigation. When considering discrete responses, the graphical tools are of limited use.

When trying to find a suitable model for a given dataset, the choice on the link function between the mean of the response and the additive predictor is very important. To obtain more flexibility, the link function is often estimated from the data. The logit link, for example, is symmetric, but the data must not be. In this case, it would be better to use an asymmetric link function, for example the complementary log-log. For linear models, there are approaches for estimating the link function. In nonparametric models, the choice on the link function has to be made by considering the properties of the data and also by comparing different models with different link functions.

An important issue is the decision about how the covariates should be included in the model. It is possible that there are nonlinear covariate effects as well as spatial effects and also interaction terms. A typical way to find a suitable model is to start with a pure main effects model and consider interactions in subsequent steps. In general, it is not advisable to include all variables in the model without reflecting their correlation to the response variable. Care has to be taken because there is always the risk of overfitting the data. However, it is important to try different models, for example with and without spatial effect or interaction terms and to compare these models with model choice criteria described in the next section. Sometimes there are covariate effects that are nearly linear, so that parts of the model can be modeled parametrically. An analysis of the heterogeneity effects provides information about covariates or interactions that are not already included in the model. This effect can be, for example, generated from geographic areas. When including new covariates in this case, for example a binary dummy variable, the spatial or cluster-specific effect should be decreased. Therefore, it is advisable to compare models with and without these spatial or cluster-specific effects. It must be pointed out that there will not be one perfect model, but rather several models which fit the data almost equally well.

3.8.2 Model Diagnostics

The tools of model diagnostic check the adequacy of a working model and also test assumptions, which were made for the modeling process, such as homoscedasticity of the error terms. For continuous responses, residual plots are a useful tool for model diagnostics. Appropriate residuals for nonparametric models are discussed later. This residual plots can provide information about misspecified covariate effects. This could mean that a nonlinear effect is wrongly modeled linearly, that there are missing covariates, or that there are missing interactions. As explained in the sections before, interactions should not be unnecessarily included in the model since they can increase the complexity of the model. Residual plots can also provide information about heteroscedastic or correlated error terms. In this case, Fahrmeir et al. (2013) discuss alternatives in Sections 3.1.2, 4.1.3. and 4.1.4. Last but not least, outliers can be detected when using residual plots. In semiparametric models, robust procedures do not exist, since this class of models is much less developed than linear models. One useful way to overcome this problem is to refit the model without the outlying observations and compare the difference in the model fits.

During the modeling process, we use different estimation concepts. Model estimation is for example done with penalized likelihood or penalized least-squares and the smoothing parameter estimation is done with the GCV or the AIC. To ensure model stability, it is necessary to estimate the model under different estimation concepts and compare the estimation results. The model is stable, when there are no large differences between the different estimation results.

The analysis of variance is a very useful tool for interpreting the effects of factors and their interactions in linear models. For generalized linear models, the residual sum of squares of non-normal distributions is no longer an appropriate measure. Therefore, another measure is introduced for generalized linear models, namely the *likelihood-ratio statistic* or also called *deviance* (Hastie and Tibshirani, 1990). The scaled deviance for a fitted model is defined by

$$\frac{1}{\phi}D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2(l(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) - l(\boldsymbol{y}|\boldsymbol{y})),$$

where $\hat{\boldsymbol{\mu}}$ is the parameter vector that maximizes the log-likelihood function $l(\boldsymbol{\mu}|\boldsymbol{y})$ over all $\boldsymbol{\mu}$. Thus, the deviance plays the role of the residual sum of squares for generalized models and can be used for estimating the goodness-of-fit and also for comparing nested models. The asymptotic distribution theory for generalized linear models is well known, whereas the distribution theory for nonparametric and additive models is undeveloped. Nevertheless, it is common to perform informal deviance tests using the χ^2 distribution to compare models. When using the `mgcv` package in `R` from Simon N. Wood, another tool for comparing different models is the *Deviance explained*, calculated as

$$\text{Deviance explained} = 1 - \frac{\text{Residual Deviance}}{\text{Null Deviance}},$$

where the null deviance shows how well the response variable is predicted by a model that includes only the intercept, whereas the residual deviance is the deviance of the fitted model. The deviance explained is given in percent and large values imply a good fit of the model.

As already mentioned before, model diagnostics mostly rely on the different types of residuals. The ordinary, the standardized, and the partial residuals can be defined in a similar way to those in the linear model, whereas there is no analogue definition for studentized residuals. Residuals can be used to investigate the adequacy of the considered model, relating to choice of the variance function, the link function and the predictor. Residuals can also reveal the presence of outliers in the data. There are also appropriate residuals for generalized linear models.

According to McCullagh and Nelder (1989), the *Pearson Residuals* are defined by

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}},$$

thus they are raw residuals scaled by the standard deviation of y, if the scale parameter $\phi = 1$. The main drawback of the Pearson residuals is that the distribution of r_P of non-normal distributions is often skewed, so it does not have the same properties as a normal-theory residual. Therefore, for defining the *Anscombe Residuals*, a function $A(y)$ is taken instead of y, where the function $A(\cdot)$ is chosen to make the distribution of $A(y)$ as close to the normal distribution as possible. The *Anscombe Residuals* are defined by

$$r_A = \frac{A(y) - A(\hat{\mu})}{A'(\hat{\mu})\sqrt{V(\hat{\mu})}},$$

where the function $A(\cdot)$ is given in generalized linear models as

$$A(\cdot) = \int \frac{d\mu}{V^{1/3}(\mu)}.$$

Finally, the *Deviance Residuals* are defined as

$$r_D^i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

where the unscaled deviance is assumed to be represented as

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i.$$

The values of the Anscombe and the deviance residuals are often very similar, although the functional forms for non-normal distributions are very different. Goodness-of-fit criteria such as AIC or the GCV can be used in case of GAMs for the variable selection and also the model choice.

CHAPTER 4

Exploratory Analysis of the House Price Data

4.1 Description of the House Price Data

In this chapter, the datasets of rented and sold houses in Sydney are described and analysed, before deriving a hedonic model describing the renting and selling prices of these houses in the next chapter. Two different datasets are used in this thesis. The first dataset contains rental prices of houses located in Sydney, whereas the other file consists of selling prices of houses in Sydney. In the rest of the thesis, the data from the first file will be named as "the rents" and the one from the second file as "the sales". There is information about 329.141 houses in the rents and 561.977 in the sales. Both datasets consist of a number of columns with information about each rented or sold house.

To distinguish between the different houses, each house in the rents and in the sales has its own ID. All the IDs are summarised in the variable ID2 and contained in the first column of each dataset. Many IDs can be found more than once, thus these houses were rented or sold repeatedly. For each transaction, there is a corresponding price, placed in the next column of the dataset. The variable PRICE is given in Australian Dollar, which will be abbreviated in this thesis with the sign $. It must be pointed out that the rents are given as weekly prices. The sale prices run from 851$ to 44.000.000$, whereas the cheapest house in the rents costed 20$ per week and the most expensive house 10.000$ per week. In addition to the house price or weekly rent, there is further information about every house given in the data, related to type, size and other characteristics of the house and most importantly the time of transaction for the sales and the date of advertisement for the rents, respectively.

Firstly, the area of Sydney was divided into 16 regions, labeled with the letters A to P. These regions are obtained by clustering postcode areas according to RESIDEX, an Australian provider of property information. The subdivision of Sydney into the different regions is shown in Figure 4.1. The allocation of the houses in the different regions is given by the variable REGION, which takes a value between the letters A and P.

region	label	postcode
Inner Sydney	A	2007 - 2011, 2015 - 2020
Eastern Suburbs	B	2021 - 2036
Inner West	C	2037 - 2050
Lower North Shore	D	2060 - 2069
Upper North Shore	E	2070 - 2077, 2079 - 2087
Mosman-Cremome	F	2088 - 2090
Manly-Warringah	G	2092 - 2097, 2099 - 2108
North Western	H	2110 - 2122, 2125 - 2126
Western Suburbs	I	2127 - 2128, 2130 - 2138, 2140 - 2145
Parramatta Hills	J	2146 - 2148, 2150 - 2159
Fairfield-Liverpool	K	2160 - 2168, 2170 - 2179
Canterbury-Bankstown	L	2190 - 2200
St. Georges	M	2203 - 2214, 2216 - 2223
Cronulla-Sutherland	N	2224 - 2234
Campbelltown	O	2556 - 2560, 2563 - 2570
Penrith-Windsor	P	2745, 2747 - 2750, 2752 - 2754, 2756 - 2763, 2765 - 2770, 2773 - 2775, 2777

Table 4.1: The regions of Sydney with the corresponding identification as letter and the corresponding postcode

Furthermore, every house in the datasets was assigned with the corresponding postcode and the geographic location, given as pair of longitude and latitude. There are 232 different postcodes in the area of Sydney given by the variable POSTCODE. The different regions and their corresponding postcodes are shown in Table 4.1. The variable LONGITUDE runs from $150°$ to $151,34°$ and the variable LATITUDE from $-34,2°$ to $-33,41°$. The variables REGION, POSTCODE, LONGITUDE and LATITUDE therefore give us exact information about the location of each house in the datasets.

Another column in the datasets, named as PROPTYPE, provides information about the type of the rented or sold house. There are five different types of houses in the datasets: Cottages, Houses, Semis, Terraces, and Villas. The houses in the rents and the sales also differ regarding their number of bedrooms, number of bathrooms, number of parking places and the size of the land area. The number of bathrooms and bedrooms both varies from one to six and is given in the columns BATH and BED, respectively. The variable PARKING shows us that the majority of the houses has between one and twelve parking places, but there is a smattering of houses with more than that. The land area of the sold and the rented houses, given by the variable AREA, ranges from 100 to 5000 square meters. Hence, all these variables give us information about the house characteristics additional to the geographic location of each house.

In addition to the abovementioned information about the house itself, the column DATE gives important information by specifying the date of advertisement or the date of the sale. These dates run from the 19th of June 2002 to the 31st of December 2014 in the rents and those for the sales from the 1st of January 2001 to the 31st of December 2014. The columns YEAR, MONTH, DAY consist of the specific information about the year, month or day, whereas the column TIME defines a continuous time scale calculated from the exact transaction date from house i as

$$\texttt{TIME}_i = \texttt{YEAR}_i + \frac{\texttt{MONTH}_i - 1 + \frac{\texttt{DAY}_i - 1}{30}}{12}.$$

It must be pointed out that with this time scale, we assume every month to have 30 days and every year to consist of 360 days. This continuous time scale is useful when modeling the trend of the renting or selling prices, as we will see in the next chapter.

Finally, it must be pointed out that there were many incomplete observations in the datasets, which were reconstructed by applying a specific algorithm. The reconstruction algorithm was performed by Sofie Waltl and is explained in detail in Waltl (2016b). The algorithm uses the fact, that most of the houses are sold or rented repeatedly, so if house characteristics were observed once, they can be reused if they are missing at another point in time.

To sum up, the two datasets for the rented and sold houses give us information about the time of advertisement for the rents or the day of the sale, respectively, as well as information about the properties of the rented or sold house, like the number of bedrooms and bathrooms, the size of the land area and the number of parking places. The geographic location of each house in the datasets is in addition given as region, corresponding postcode and longitude and latitude.

4.2 Analysis of the House Price Data

In this section, the two datasets of rented and sold houses in Sydney are analysed to get a deeper insight into the distribution of the variables and also the relationship between them. Since the goal of this thesis is to find a model to forecast house prices and rents for the nearest future, data from before the year 2008 may have limited impact. On the one hand that is due to the financial crisis in the year 2007, which had at that time an influence on the housing market and also the peak of a housing boom in 2004. On the other hand, in terms of forecasting it is better to mainly use current information. Going back into the past for more than six years is therefore not necessary. For these reasons, all of the following analysis and modeling in this thesis only uses the house rent and house price data from 2008 to 2014. This leaves us with information about 217.739 rented houses and 277.361 sold houses within the considered years.

number of parking places	**frequency**	
	rents	**sales**
1	103.471	87.234
2	72.445	112.258
3	8.520	18.226
4	3.936	9.109
5	618	1.527
6	395	1.144
7	50	206
8	48	214
9	15	55
10	7	35
11	3	13
12	4	17
13	-	4
14	-	6
15	-	2
16	-	1
18	-	1
20	-	3
21	1	-
23	-	1
24	-	1
25	1	-
40	1	-
111	1	-
222	1	-
255	-	1

Table 4.2: Frequency of the different number of parking places corresponding to the houses in the rents and in the sales

As already discussed in Section 3.8, the first important step in every statistical modeling application is the basic analysis of the data before deriving a suitable model. It is necessary to get to know the distribution of the variables in the data. Sometimes it is useful to exclude extreme observations to prevent sophisticating the model. For example, there are a few houses in the data with more than 14 parking places. More precisely, there are houses with for example 111, 222 or 255 parking places in the data. This distribution of the parking places is depicted in Table 4.2. Since there are too many extreme values, there probably are quite a few erroneous entries in this column. Therefore, the variable PARKING is excluded from further analysis.

As a next step, the different regions are analysed. In order to do that we are first interested in the distribution of the rented and sold houses in the datasets over the regions.

Table 4.2 shows us that compared to the other regions of Sydney, there are only few houses in the rents and in the sales that are located in the regions A and F. This may be due to the fact, that these regions are the inner city of Sydney, where apartments are more common than houses. We can see that there are more houses in the dataset of the sales and that these houses are more frequently located in the outer regions H to P. However, the distribution of the rented houses over the regions A to P shows nearly the same behaviour as the distribution of the sold houses. Furthermore, the means of the weekly rents and the house prices in the different regions were calculated for each year separately to get an overview of the geographical behaviour. It can be seen in Table 4.3, that there is a nearly linear increase in the mean renting and selling prices from 2008 to 2014 in every region. This correlation between the renting or selling price, respectively, and the time is also depicted in Figure 4.3. In order to get insight into the geographical behaviour of the house price, the mean rent and house price of every region was calculated and plotted in Figure 4.4. The darker the colour of the region in the map, the more expensive the region is. The map shows that the regions in the west of Sydney are the cheapest and the house price increases as the region gets closer to the east coast. The regions B and F are the most expensive ones, whereas the regions O, P, and K are the cheapest. These characteristics of the house price hold for both, the rents and the sales.

Next we are interested in the distribution of the different types of houses. Figures 4.5 and 4.6 show the frequency of every type of house in the regions A to P. It can be seen that terraces are mostly located in regions A to C at the east coast of Sydney, whereas villas most frequently stand in the outer regions H to P in the west. The other types of houses, namely the cottages, the houses, and the semis, are almost evenly distributed over the different regions. This behaviour of the distribution of the type of the house holds for both, the rents and the sales. In summary, the distribution of the type of the house as well as the corresponding mean prices are given in Table 4.4. Looking at Table 4.4 it seems, that the terrace is the most expensive type of house and the villa the cheapest one. However, when interpreting this prices care has to be taken, since terraces are mainly located in the most expensive areas, whereas the villas stand in the cheap west of Sydney. Thus, there might be a correlation between the region and the type of the house. Therefore, the mean prices of the different types of the houses have to be interpreted carefully. It can be seen in the table that the house is by far the most common type.

Furthermore, the influence of the number of bedrooms, the number of bathrooms and the size of the land area should be evaluated. Therefore, we calculate the mean prices depending on the number of bedrooms, the number of bathrooms and the size of the land area. The results are pictured in Table 4.5. It can be seen that the house price of the rents and the sales rises, if the number of bathrooms or bedrooms increases, which would be expected intuitively. The frequency of the different number of bedrooms and bathrooms, respectively, is also depicted in Figure 4.7, which shows that the most common number of bedrooms is three and four whereas the houses in the rents and also in the sales most frequently have one or two bathrooms. There is also a correlation between the number of bedrooms and the number of bathrooms. Figure 4.8 shows us, that if the number of bedrooms increases, also the number of bathrooms augments. This behaviour can also be seen in Table 4.6. Table 4.7 depicts the mean prices of the rents and of the sales depending on the number of bedrooms and bathrooms. It can be seen that the price rises for increasing number of bedrooms and also for increasing number of bathrooms. There are some outliers in the table, for example is the mean price for a rented house with five bathrooms and one bedroom only 210$. In fact, some of the values in the table have to be taken with care, since there are only a few houses with the corresponding number of bedrooms and bathrooms. In the case of the above mentioned example, there are only two houses with five bathrooms and one bedroom in the rents and those houses are located in the cheap regions I and J, respectively. Table 4.6 shows us, that there are only a few houses with more than four bathrooms and only one or two bedrooms in both, the rents and the sales. These outliers might be data errors.

The interpretation of the effect of the size of the land area on the house price is more difficult. It can be seen in Table 4.5 that the house price augments for increasing size of the land area, but only up to an area of 2.000 to 3.000 square meters. The house price then falls for an area greater than 3.000 square meters. A possible reason for this behaviour may lie in the fact, that houses with a large land area are more frequently located in the outer regions of Sydney in the west, where the house price is much lower than in the east. The variable AREA gives us the land area in square meter and not the area of the corresponding house itself, thus the price per square meter might be cheaper in the outer regions. The distribution of the houses with an area greater than 4.000 square meters over the regions A to P can be seen in Figure 4.9. Another explanation might be, that the variable AREA gives the size of the land area but not the actual size of the structure. Therefore, a large land area must not correspond with a large living area which might lead to a complicated nature of the effect. Analysing the land area conditional on the number of bedrooms, which is a proxy for the living area, might disentangle the effect.

In summary, the analysis of the dataset has shown important aspects, which are necessary for deriving a model for the house price. Firstly, there is a dependence of the house price on the date of the sale or the date of advertisement, respectively. Secondly, the spatial location of the house plays a decisive role in pricing the house. The location of the house measures many things at the same time. On the one hand, the location determines the distance to public amenities such as schools, parks, beaches, shopping centers and public transportation. The location may for example determine the school district and hence, which public school kids are assigned to. A school district with a good public school may therefore have higher house prices. On the other hand, the spatial location also measures things like air pollution, crime rates and traffic noise. Areas with higher crime rates or houses near main streets may have lower house prices. Furthermore, there is a correlation between the type of the house and the regional location and thus also between the house price and the type of the house. At this point it is difficult to say, whether the information of the type of the house gives us additional information about the house price, when there is already information about the geographic location. Last but not least, there is an influence of the number of bedrooms, the number of bathrooms as well as the size of the land area on the house price. With this knowledge, a model, which describes the price of a specific house subject to the just mentioned variables will be derived in the next chapter.

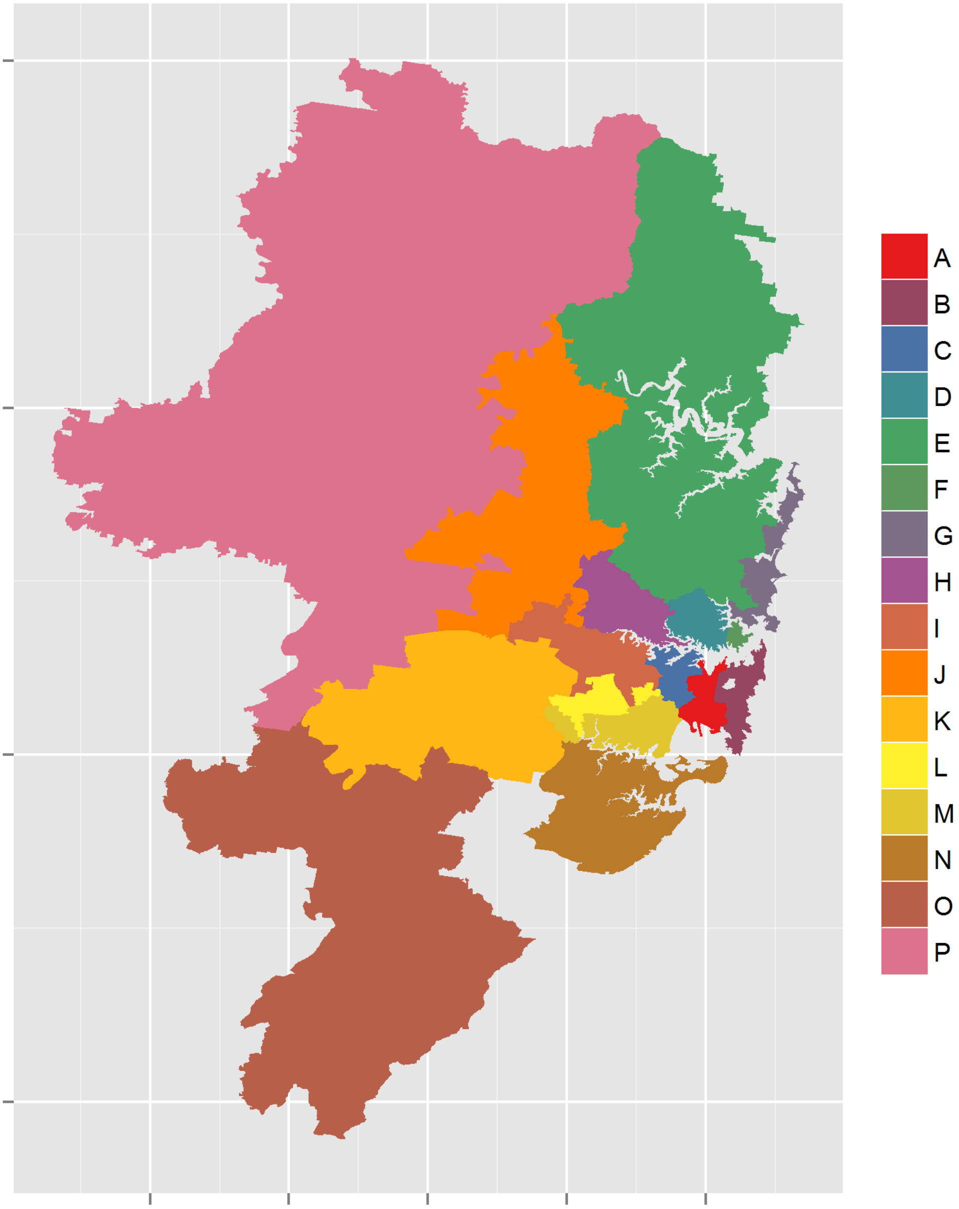


Figure 4.1: The subdivision of Sydney into 16 different regions named with the letters A to P

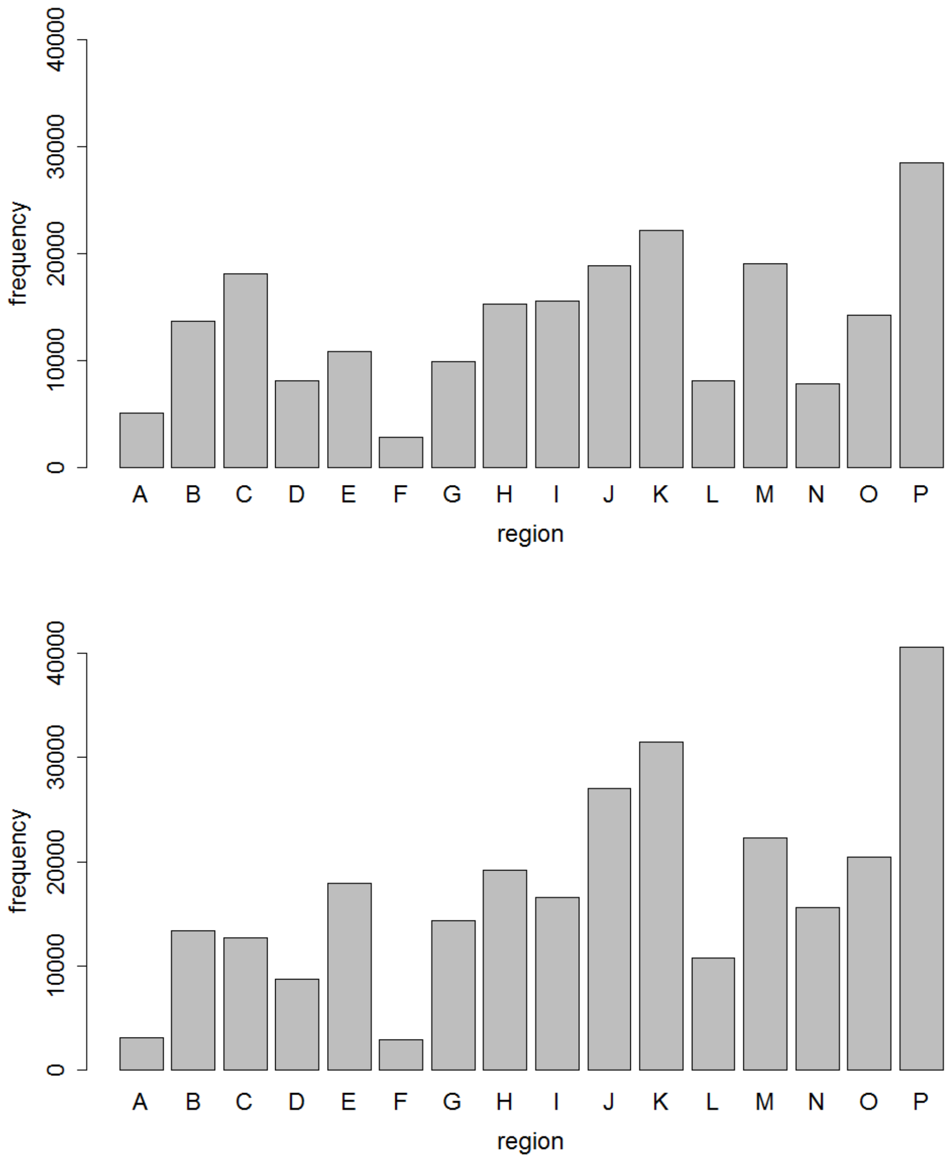


Figure 4.2: Frequency of rented (top) and sold (bottom) houses in the regions A to P

rents							
region	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
A	642	655	701	733	747	786	840
B	1.049	1.096	1.103	1.177	1.170	1.224	1.263
C	596	616	666	714	724	764	793
D	847	860	925	946	985	1.007	1.068
E	775	762	823	858	863	881	916
F	1.335	1.338	1.524	1.603	1.551	1.535	1.621
G	831	845	927	959	968	1.001	1.045
H	534	551	575	613	619	641	668
I	437	451	485	511	520	540	559
J	437	446	478	511	520	519	539
K	360	382	408	425	433	446	463
L	406	423	448	467	479	493	526
M	452	468	510	535	548	565	593
N	538	564	600	624	633	647	673
O	330	346	385	398	414	421	437
P	327	349	377	398	410	417	428
sales							
region	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**
A	839	916	1.011	1.033	1.001	1.169	1.343
B	1.958	1.908	2.090	1.981	1.903	2.232	2.477
C	909	946	1.103	1.081	1.071	1.224	1.374
D	1.454	1.397	1.565	1.546	1.546	1.676	1.904
E	1.004	1.030	1.118	1.097	1.116	1.237	1.451
F	2.510	2.243	2.496	2.608	2.540	2.588	3.018
G	1.235	1.242	1.348	1.277	1.262	1.379	1.508
H	800	831	967	951	956	1.080	1.297
I	615	666	779	773	797	900	1.084
J	523	539	601	599	623	699	831
K	382	411	457	467	486	538	634
L	472	493	586	596	616	688	818
M	648	687	786	791	801	885	1.041
N	741	788	886	866	887	945	1.068
O	339	358	390	395	412	446	518
P	359	378	406	413	429	465	543

Table 4.3: Mean prices of the weekly rents and the ones for the sales in thousand $, calculated for each region depending on the year

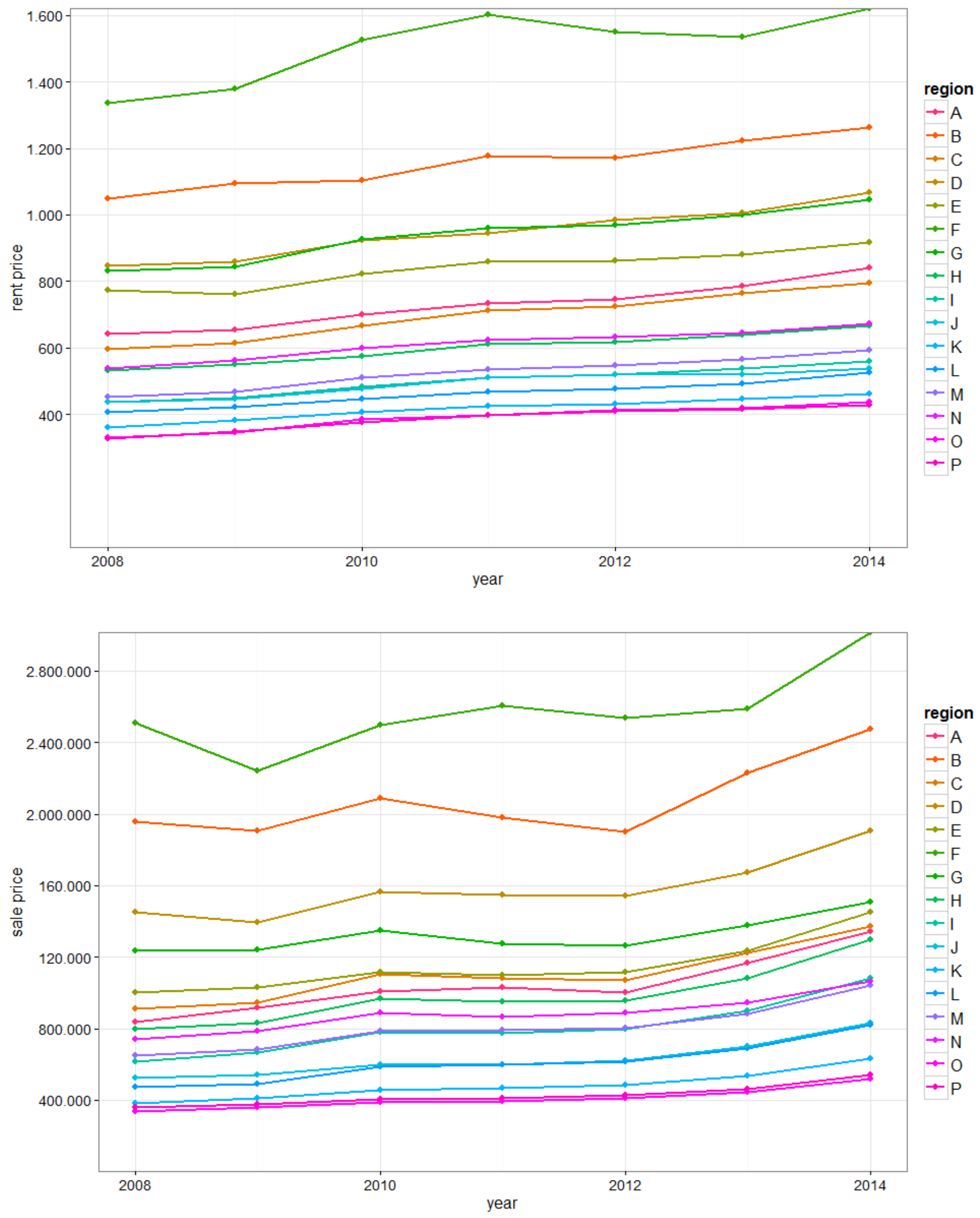


Figure 4.3: Trend of the mean renting (top) and the mean selling prices (bottom) in the regions A to P over the years 2008 until 2014

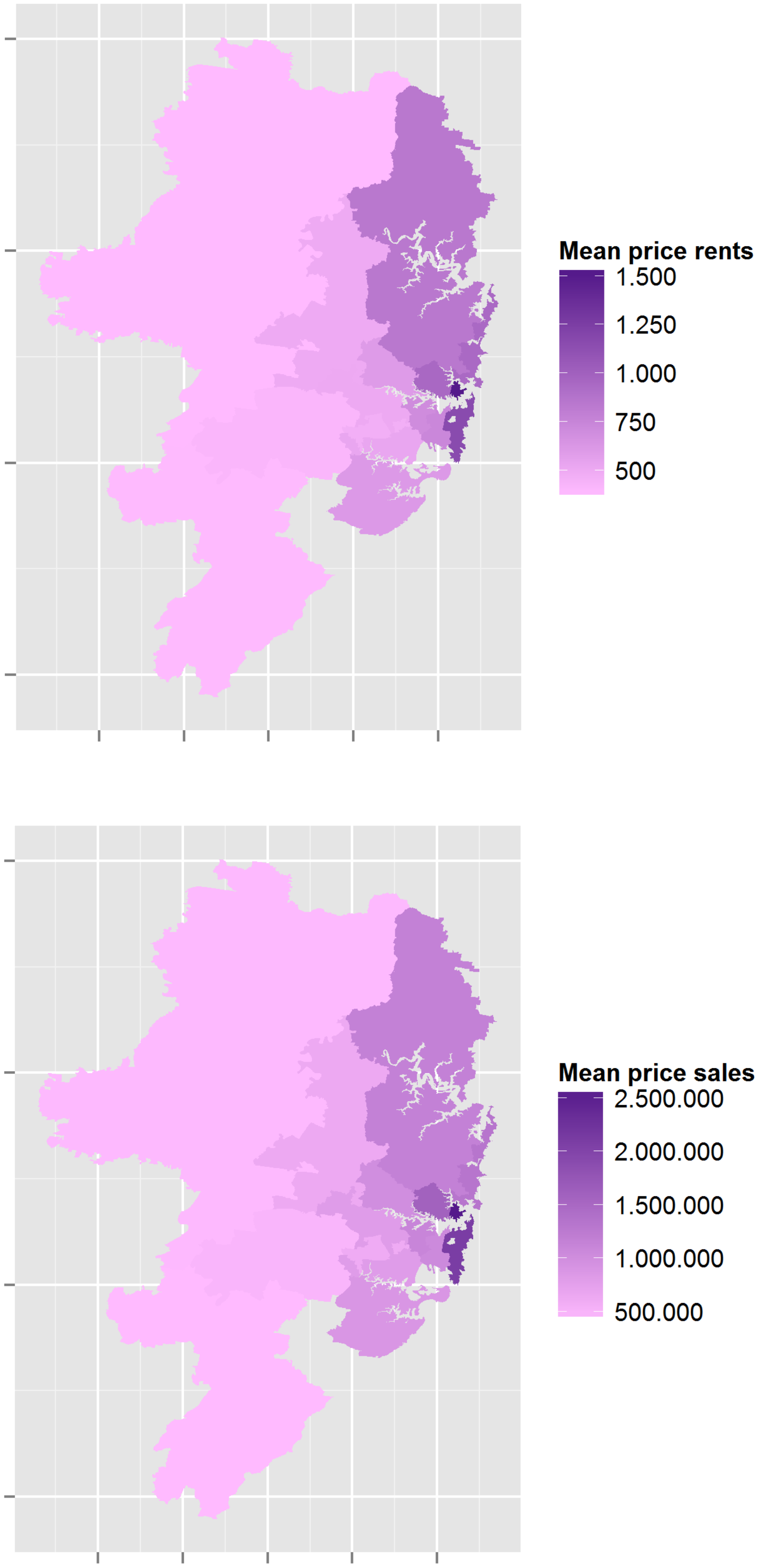
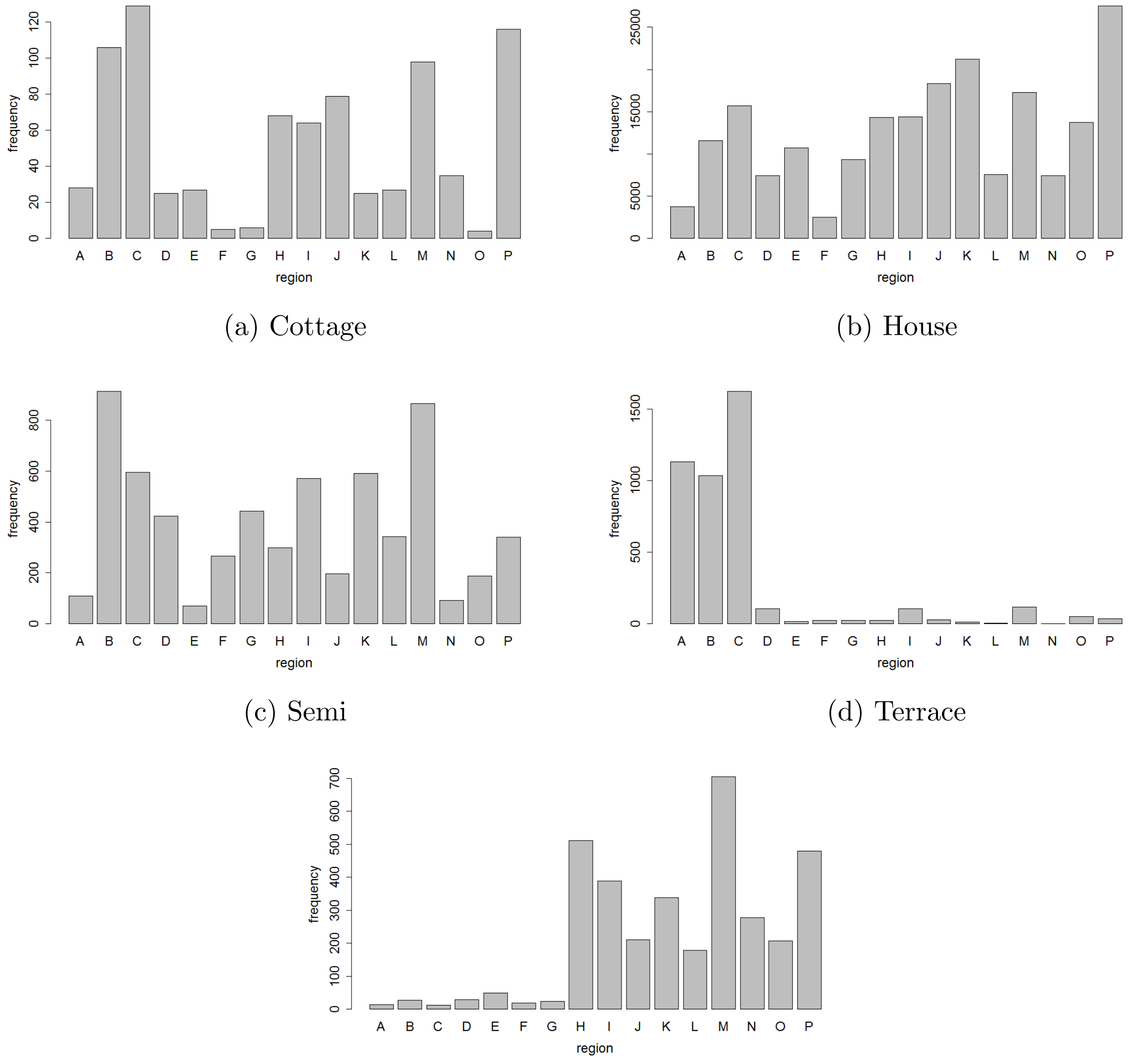


Figure 4.4: Mean renting (top) and selling (bottom) prices in the different regions of Sydney

	Type of House	Cottage	House	Semi	Terrace	Villa
rents	**frequency**	842	202.778	6.300	4.348	3.471
	mean price	611	616	559	803	440
sales	**frequency**	1.150	270.378	3.276	1.279	1.278
	mean price	846.105	853.211	764.423	1.161.793	485.000

Table 4.4: Frequency and mean prices of the different types of houses in the rents and in the sales



(a) Cottage

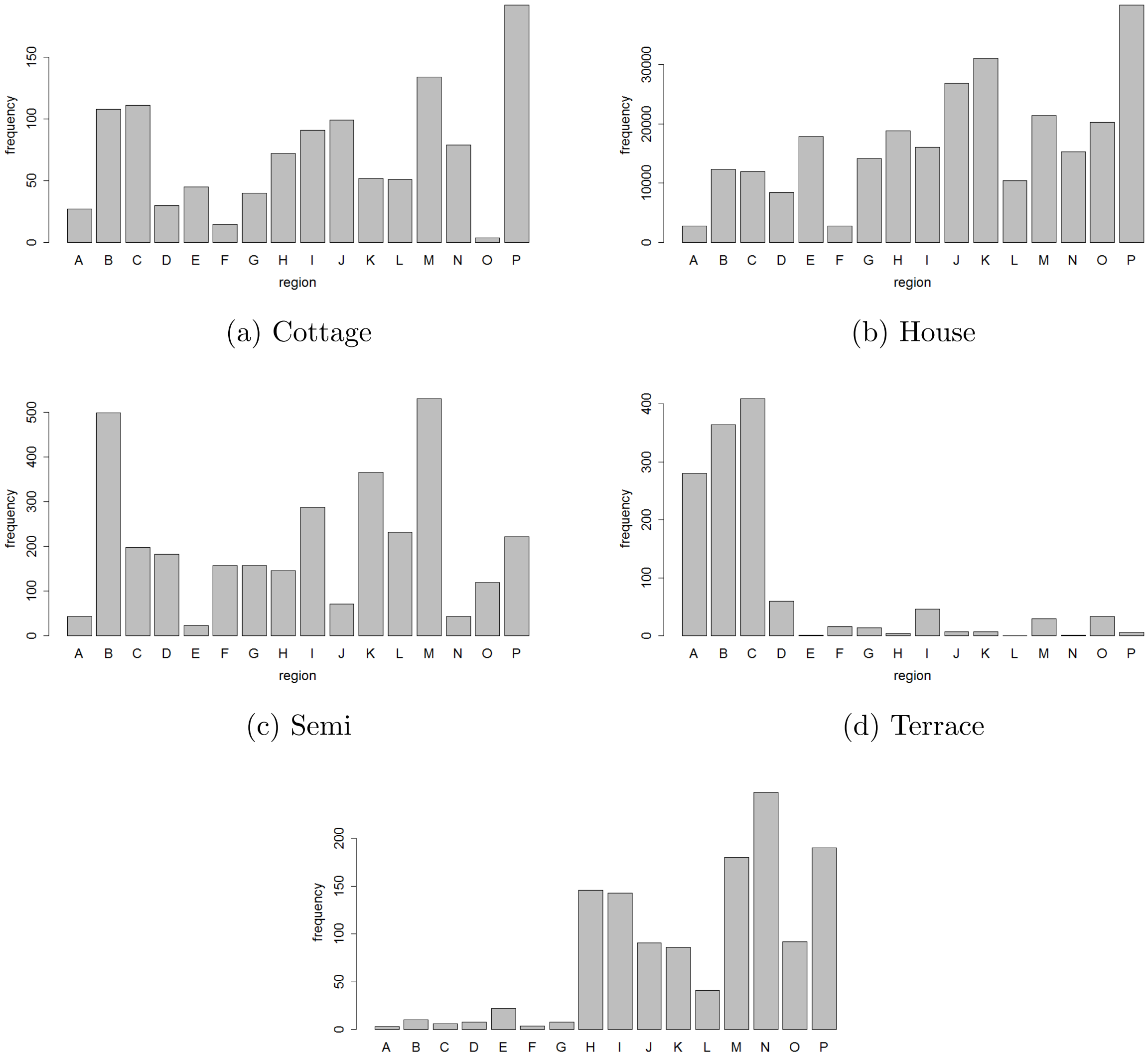
(b) House

(c) Semi

(d) Terrace

(e) Villa

Figure 4.5: Frequency of the different types of rented houses in the regions A to P



(a) Cottage

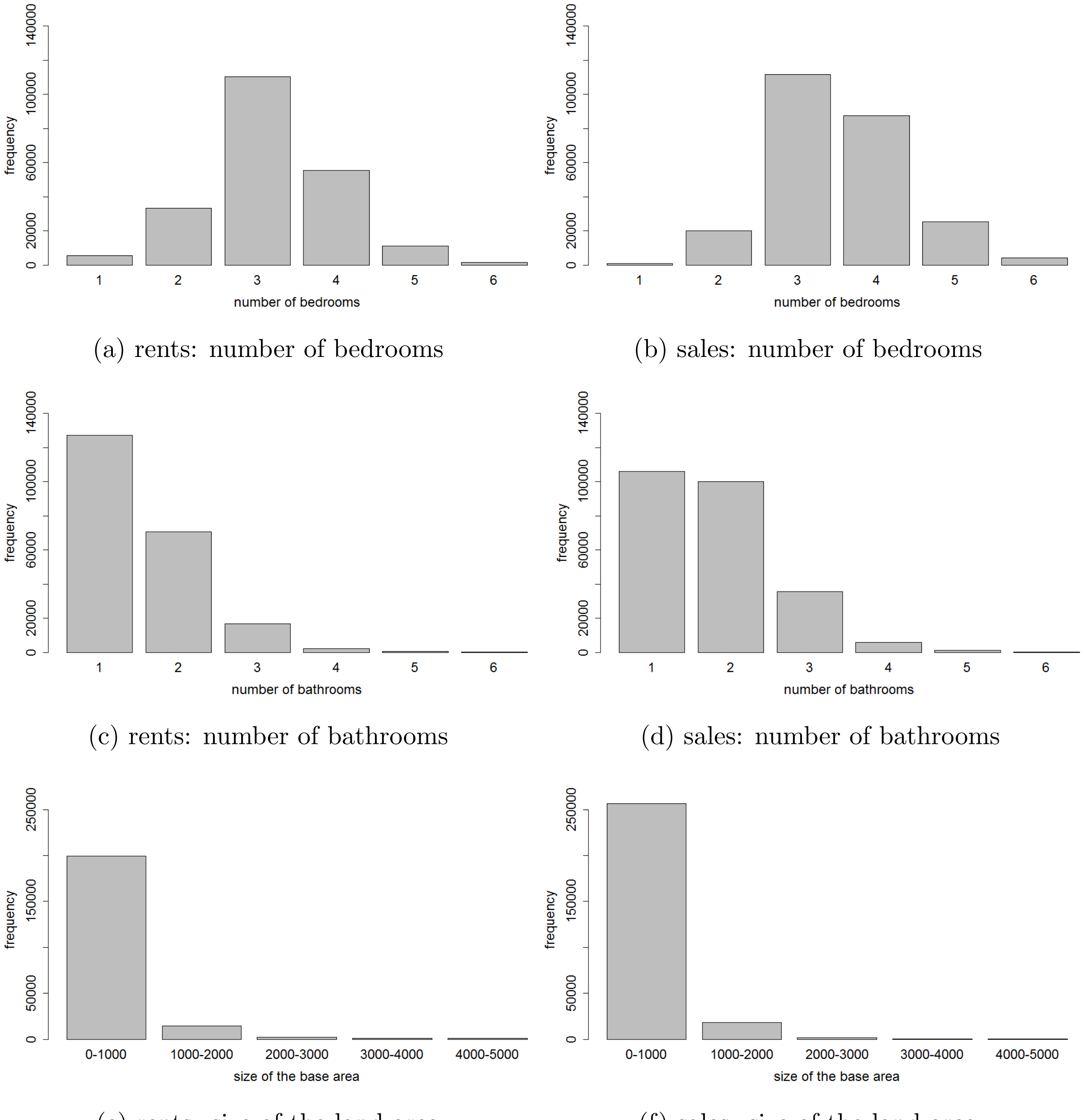
(b) House

(c) Semi

(d) Terrace

(e) Villa

Figure 4.6: Frequency of the different types of sold houses in the regions A to P



(a) rents: number of bedrooms

(b) sales: number of bedrooms

(c) rents: number of bathrooms

(d) sales: number of bathrooms

(e) rents: size of the land area

(f) sales: size of the land area

Figure 4.7: Frequency of the different number of bedrooms, bathrooms and the size of the land area for the rented (left) and for the sold (right) houses

	bedrooms	**1**	**2**	**3**	**4**	**5**	**6**
rents	**frequency**	5.379	33.368	110.452	55.325	11.042	1.559
	mean price	327	500	541	746	1.092	1.248
sales	**frequency**	703	20.007	111.708	87.507	25.240	4.172
	mean price	690.809	719.273	685.719	933.148	1.363.125	1.664.886

	bathrooms	**1**	**2**	**3**	**4**	**5**	**6**
rents	**frequency**	127.164	70.562	16.825	2.271	416	101
	mean price	480	695	1.098	1.734	2.301	2.686
sales	**frequency**	106.150	100.025	35.473	5.796	1.202	319
	mean price	618.261	853.000	1.284.733	2.160.210	3.285.156	4.317.466

	area	**0 - 1000**	**1000 - 2000**	**2000 - 3000**	**3000 - 4000**	**4000 - 5000**
rents	**frequency**	199.211	14.578	2.273	1.007	777
	mean price	611	681	636	604	581
sales	**frequency**	256.733	18.082	1.588	531	576
	mean price	813.614	1.307.374	1.654.988	1.390.910	1.032.276

Table 4.5: Frequency and mean renting and selling prices of houses with different number of bedrooms and bathrooms as well as the different sizes of the land area

rents							
	bedrooms	**1**	**2**	**3**	**4**	**5**	**6**
bathrooms	1	5.237	29.480	79.041	11.633	638	60
	2	100	2.551	28.861	33.744	4.437	483
	3	18	64	2.386	9.047	4.627	591
	4	3	1	94	779	1.097	270
	5	2	1	4	85	182	122
	6	3	0	1	9	53	29
sales							
	bedrooms	**1**	**2**	**3**	**4**	**5**	**6**
bathrooms	1	658	17.916	74.357	12.236	680	53
	2	18	1.830	33.738	54.191	8.947	998
	3	12	40	2.632	18.745	11.941	1.757
	4	5	16	103	1.690	2.853	881
	5	1	2	19	175	576	330
	6	0	1	8	28	108	114

Table 4.6: Frequency of houses with different number of bedrooms and bathrooms in the rents and in the sales

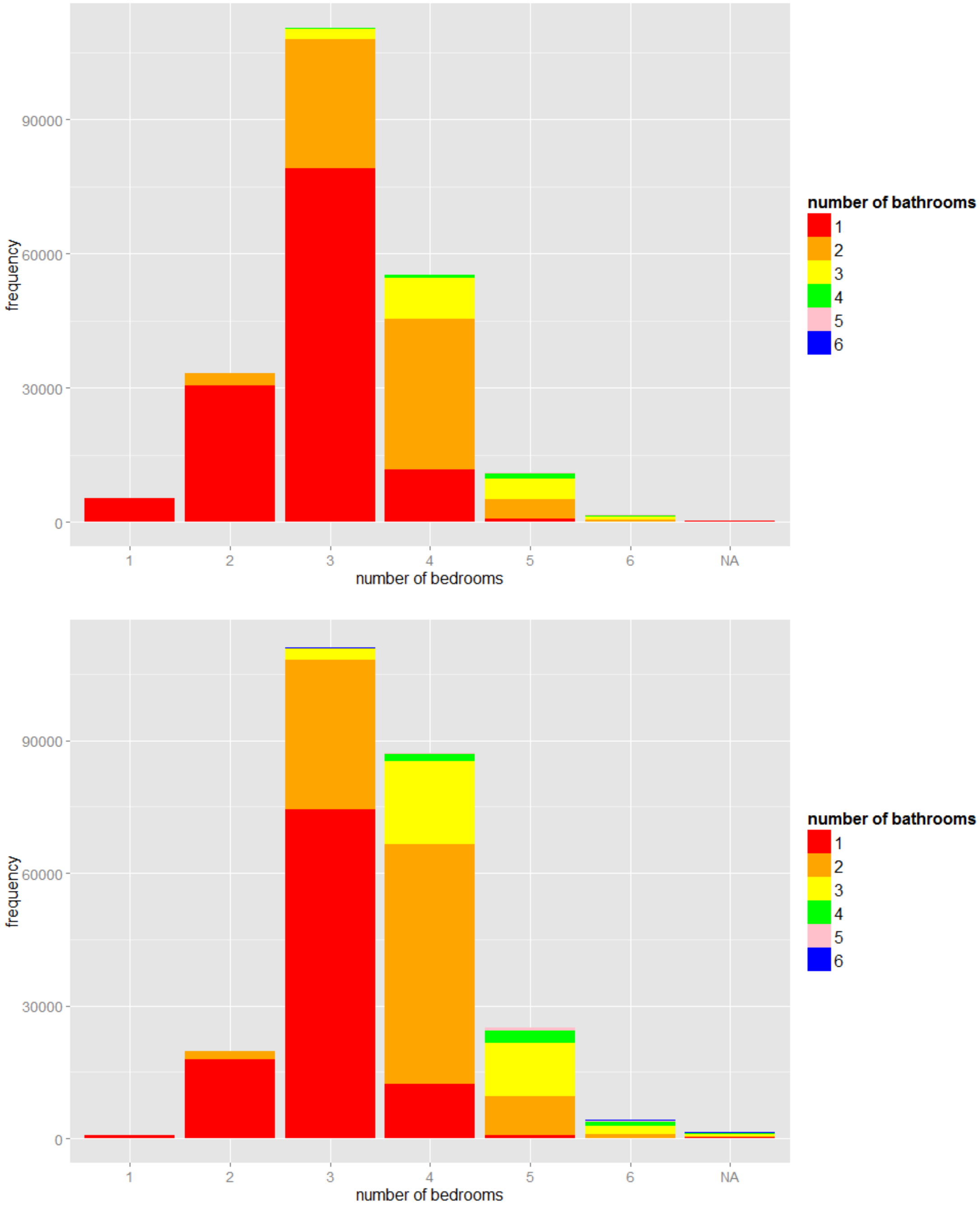


Figure 4.8: Frequency of the different number of bedrooms depending on the number of bathrooms for the rents (top) and for the sales (bottom)

rents							
	bedrooms	**1**	**2**	**3**	**4**	**5**	**6**
bathrooms	1	325	488	480	526	567	584
	2	418	614	673	705	801	765
	3	375	997	949	1.087	1.194	1.150
	4	400	950	1.421	1.662	1.832	1.707
	5	210	480	923	2.346	2.252	2.504
	6	318	-	1.300	2.326	2.952	2.642
sales							
	bedrooms	**1**	**2**	**3**	**4**	**5**	**6**
bathrooms	1	676	702	594	632	671	1.100
	2	1.113	868	841	841	954	926
	3	909	1.071	1.192	1.253	1.355	1.304
	4	800	1.233	2.318	2.179	2.221	2.080
	5	1.905	4.700	1.651	3.647	3.372	3.355
	6	-	2.900	1.160	4.626	5.158	4.337

Table 4.7: Mean prices of the weekly rents and the ones for the sales in thousand $ depending on the number of bedrooms and bathrooms of the house

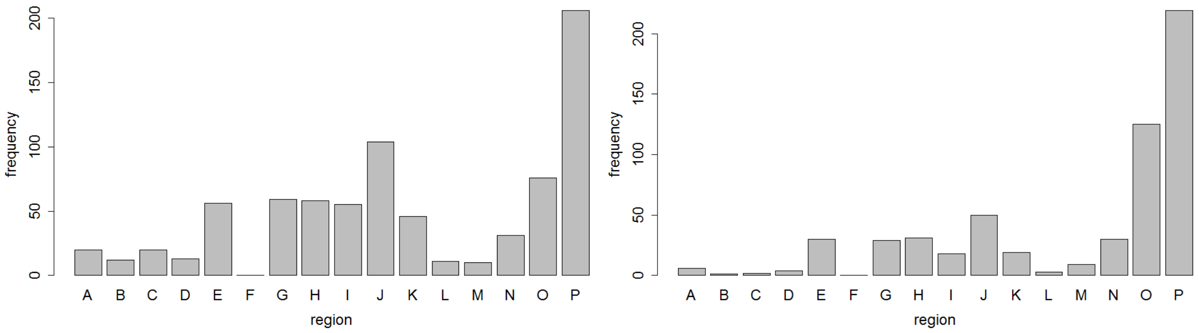


Figure 4.9: Frequency of the houses in the rents (left) and in the sales (right) with area greater than 4.000 square meters, over the regions of Sydney

CHAPTER 5

A Hedonic House Price Model for Sydney

There are many reasons, why it is important to model the trend of the house price and to forecast it for the nearest future. In their paper an de Meulen, Micheli, and Schmidt (2011) explained that variations in the house price have an important impact on the financial stability and the real economic activity. They argued that the decline of house prices in the year 2007 led to a huge deterioration of financial institutions balance sheets and consequently provoked the financial crisis. More specific, the trigger of the financial crisis were, amongst others, speculative price bubbles, which produced deep recessions accompanied by huge employment reductions (Kholodilin and Siliverstovs, 2014). Such a price bubble exists, if the reason that the price is high today is only because investors believe that the selling price is high tomorrow (Stiglitz, 1990). Leamer (2007) actually states in his paper that the housing market predicted eight out of ten post World War II recessions. Moreover, according to Gupta and Kabundi (2010) the trend of the house price can be an indicator for the future movement of inflation. On the other hand, if economy booms, there is expansion in the employment and also in the construction in the housing market due to the excess demand, which pushes the house prices upwards (Plakandaras, Gupta, Gogas, and Papadimitriou, 2015). For all these reasons, there is much interest in modeling and forecasting house prices for economic stability.

There are price indexes for explaining the housing market. Standard hedonic indexes are obtained with linear regression models and show average movements of average homes in average locations. Thus they can detect mean changes of house prices over time. An introduction to house price indexes with geospatial data can be found in Hill and Scholz (2014). It is also possible to derive other house price indexes. A hedonic house price index in continuous time for Sydney over the period from 2001 to 2011 is derived in Waltl (2016a). Furthermore, as proposed by Waltl (2016b), quantile regression models can be used instead of linear regression models to construct hedonic indexes for different locations. These models are used in order to do some analysis on the house sales in Sydney between 2001 and 2014.

For examining the housing market the price-rent ratio, which is the ratio of property prices relative to rents, is an indicator of the increase of the price of residential property. The main reason why the ratio is of interest is the user cost of housing. The user cost formula links the price-rent ratio with the expected capital gains and other rates (for example interest rate and property tax rate). The meaning of the price-rent ratio will be discussed in detail in Section 6.2. In terms of Sydney there are several papers of research work considering the Sydney housing market, for example the work of Hill and Syed (2016). Another example is the work of Hatzvi and Otto (2008), who used quarterly data of the 36 Local Government Areas of Sydney from 1991 to 2006. They found out that variations in price-rent ratios assume future real rent growth. Furthermore, their results suggested that the boom in Sydney property prices within the considered years can be explained with the lower long-term real interest rates. Thus, the asset pricing theory in Sydney can explain the residential property prices.

The goal of this chapter is to derive a generalized additive model, which shows the trend of the renting and selling prices of houses in Sydney over the period of the years 2008 to 2014. One big advantage of using splines when modeling geospatial data is the fact that in contrast to kernel estimators, splines have no problems at the boundaries of the considered domain, since they are estimated locally. Furthermore, hedonic price indexes show the average price movement of average homes, where averaging house characteristics like the size of the land area or the number of bedrooms is possible, but calculating an average location makes no sense. Therefore, it is reasonable to include the information about the location as a geospatial effect, as mentioned in Hill and Scholz (2014). The theory of including a geospatial effect in a model was already introduced in Section 3.6.1. The obtained model can then be used in the following chapter to forecast the rents and the house prices for the nearest future. All the following calculation is done with the program `R` using the `mgcv` package, which will be introduced shortly in the next section.

5.1 The `mgcv` Package in `R`

For fitting GAMs in the program `R`, Simon Wood implemented the package `mgcv` (Mixed GAM Computation Vehicle). An introduction to this package can be found in the book Wood (2006). The `mgcv` package contains many different functions in terms of fitting a GAM, but the most important one for our purpose is the function `gam()`. This function fits a generalized additive model to a given dataset. When fitting such a model, all properties of the model can be specified, like how the variables are included in the model, which member of the exponential family and what kind of link function is used and also the selection method of the smoothing parameter can be chosen. The function and its most important arguments are given by

```
> gam(formula, family=gaussian(link=identity), data=list(),
        method="GCV.Cp")
```

The first argument `formula` specifies the form of the GAM, by determining how the covariates are included in the model. This `formula` argument can be specified, for example, as

```
y ~ s(x, bs="cr", k=20) + z
```

In this formula, y defines the response variable and x and z the explanatory variables. In this example, the covariate x enters the model nonparametrically, since it is given as a function `s(x)`, whereas the covariate z is included linearly. In case of the nonparametric term, it is also possible to choose the basis functions for the smoother. This can be done with the argument `bs` in the function `s()`. The GAM is fitted with cubic regression splines when the argument is set to `"cr"` and with thin plate regression splines when it is set to `"tp"`. There are also other basis functions available, but these two are the most important ones for our purpose. If we consider multivariate smoothing, which means that the function `s()` takes several predictor variables as arguments, then there are two possible types of smooths. Within the function `s()`, the `bs` argument can be set to `"tp"` for thin plate splines or to `"te"` for tensor product splines. The theory about the above mentioned basis functions was already introduced in Chapter 3.

The last important argument in the nonparametric term of the above formula is the number of knots used for fitting the splines. This argument can be left out and then the default of $k = 10$ is used. To find a suitable basis dimension, one can start with a small number of knots and increase them in every step. There are sufficiently enough knots used for fitting, if the score of the chosen optimization criterion is not getting smaller any more when the knots are increased. The basis dimension also sets the maximum possible degrees of freedom allowed for each term, since the actual effective degrees of freedom are estimated from the data with an upper limit of $k - 1$.

The next argument of the function `gam()` is the `family` argument, which determines the member of the exponential family used for the model. This argument can be, amongst others, `gaussian()`, `binomial()`, `poisson()` or `Gamma()`. There are also other distributions which can be used here, but they are not important for our task. The argument in the exponential family member sets the chosen link function used for the model. The `data` argument refers to a data frame or a list containing the model response variable and all covariates. The argument `method` determines the smoothing parameter selection method. It can be set either to `"GCV.Cp"`, `"UBRE"`, `"AIC"` or to `"fREML"`. The selection of the smoothing parameter via GCV, UBRE and AIC was already discussed in the different sections of Chapter 3. The estimation with the fast stable restricted maximum likelihood (fREML) works with Newton-Raphson rather than Fisher scoring. Details about the algorithm can be read in Wood (2011) and will not be further discussed in this thesis.

Once the model is fitted, `mgcv` offers a few additional functions for obtaining further information about it. One example is the `summary()` or `summary.gam()` function, which returns the estimated parameters of the fitted model, the significance of the covariates and also the values for the GCV, UBRE, AIC or fREML score, respectively. In short, it gives a summary of the fitted model. Furthermore, an Analysis of Variance (ANOVA) can be done with the `anova()` or `anova.gam()` function, which returns the degrees of freedom of each term and also the corresponding p-values. The function `gam.check()` gives diagnostic information about the fitting procedure and results, which include residual plots and information about the convergence of the smoothness selection optimization. The function `plot.gam()` plots the component smooth functions of the model as well as the parametric model components on the scale of the linear predictor. Moreover, the function `vis.gam()` allows producing plots with model predictions from the fitted model, which can also be three-dimensional. When the dataset of interest is very large, the function `bam()` can be used instead of the function `gam()`, which works in the same manner as the function `gam()`. There exist also many other functions in terms of GAMs which will not be discussed here, but details about them can be read in the description of the `mgcv` package.

5.2 Derivation of a Hedonic House Price Model

Hedonic models regress the price of a house on a vector of physical and locational characteristics, whose prices are not independently observed. Thus, these models offer a flexible way to control for differences in quality and location (Waltl, 2016a). Decomposing property prices into different price determining components leads the hedonic price function. Additionally, time effects are used to measure the effect of the time the price announcement was made.

The exploratory analysis in Chapter 4 has shown that the renting and selling prices of houses in Sydney depend on the date of advertisement or the day of the sale, respectively, as well as on the number of bedrooms and bathrooms, the size of the land area, the location and the type of the house. With this information we can derive a hedonic model for the house price and for the weekly rents containing the just mentioned variables. As we have seen in Chapter 3, there are different ways of including a variable in a model and also of choosing a suitable link function and a proper member of the exponential family. In the next section, we begin with the simplest model, as we assume the rents and the sales to follow a normal distribution with an identity link.

5.2.1 House Price Model with Normally distributed Prices

As a first step, we assume that the renting and selling prices y_i of house i $(i = 1, \dots, n)$ can be described by a normal distribution. We suppose that the house prices and the rents change smoothly over time and also that the size of the land area can be described by a smooth function. Furthermore, there should be a bivariate smooth term depending on the longitude and latitude of the house, controlling the effect of the geographical position. In terms of the number of bedrooms and bathrooms, we try two different models. In the first model the variables bed and bath are included linearly, whereas in the second model they are included as categorical variables. Thus, we obtain the models mr_bed_bath_linear (for the rents) and ms_bed_bath_linear (for the sales) in (5.2.1), as well as the models mr_bed_bath_factor (for the rents) and ms_bed_bath_factor (for the sales) in (5.2.2). The models are therefore defined by

$$\eta_i = \beta_0 + f_1(\texttt{time}_i) + f_2(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) + \beta^{\texttt{bed}}\texttt{bed}_i + \beta^{\texttt{bath}}\texttt{bath}_i, \tag{5.2.1}$$

$$\begin{aligned}\eta_i &= \beta_0 + f_1(\texttt{time}_i) + f_2(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) \\ &\quad + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i),\end{aligned} \tag{5.2.2}$$

where $\eta_i = \mu_i$. Thus, the linear predictor equals the mean of the renting or selling price μ_i of house i. Consequently, the linear predictor η_i consists of a parametric part containing information about the number of bedrooms and bathrooms, as well as a nonparametric part consisting of the geographic location, the size of the land area and the date of advertisement or the day of the sale, respectively. The difference between the two models (5.2.1) and (5.2.2) is the parametric part. In the first model the number of bedrooms and bathrooms enters the model linearly. This means that there is one parameter β^{bath} and β^{bed}, respectively, which is multiplied with the number of bedrooms or bathrooms. In the second model, the number of bedrooms and bathrooms is modeled by a categorical variable, which means that in contrast to the first model, for each number of bedrooms and bathrooms there is a different parameter β^{bath} or β^{bed} and only the parameters corresponding to the actual number of bedrooms and bathrooms enter the model.

We first want to know which model, (5.2.1) or (5.2.2), has a better fit to the house rents and the house sales data. As argued in Chapter 4, we only use data from the years 2008 to 2014. The fitting of the models is done with the program R with the function bam() from the mgcv package. The call of the function bam() used for fitting the rents is as follows.

```
mr_bed_bath_linear <- bam(PRICE ~ s(TIME, bs="cr", k=20)
        + s(LONG,LAT, bs="tp", k=100)
        + s(AREA, bs="cr", k=20)
        + BED + BATH, data=houserents2008)
mr_bed_bath_factor <- bam(PRICE ~ s(TIME, bs="cr", k=20)
        + s(LONG,LAT, bs="tp", k=100)
        + s(AREA, bs="cr", k=20)
        + factor(BED) + factor(BATH), data=houserents2008)
```

The call for fitting the sales is analogue. The two datasets houserents2008 and housesales2008 contain the information about the rented and sold houses, respectively, from the years 2008 to 2014. The fitting procedure uses cubic regression splines for the functions depending on the time and on the size of the land area with 20 knots and thin

	link	**model**	**fREML**	**GCV**	**Deviance explained**
rents	identity	`mr_bed_bath_linear`	$1,49 \cdot 10^6$	$5,17 \cdot 10^4$	$66,0\%$
	identity	`mr_bed_bath_factor`	$1,48 \cdot 10^6$	$4,90 \cdot 10^4$	$67,9\%$
	log	`mrlog_bed_bath_factor`	$1,46 \cdot 10^6$	$4,01 \cdot 10^4$	$73,7\%$
	log	`mrlog_bed_bath_region_factor`	$1,46 \cdot 10^6$	$3,97 \cdot 10^4$	$73,9\%$
	log	`mrlog_bed_bath_region_factor_2`	$1,48 \cdot 10^6$	$4,86 \cdot 10^4$	$68,1\%$
	log	`mrlog_bed_bath_proptype_factor`	$1,46 \cdot 10^6$	$3,99 \cdot 10^4$	$73,8\%$
sales	identity	`ms_bed_bath_linear`	$3,57 \cdot 10^6$	$1,90 \cdot 10^{11}$	$63,8\%$
	identity	`ms_bed_bath_factor`	$3,56 \cdot 10^6$	$1,77 \cdot 10^{11}$	$66,4\%$
	log	`mslog_bed_bath_factor`	$3,52 \cdot 10^6$	$1,34 \cdot 10^{11}$	$74,6\%$
	log	`mslog_bed_bath_region_factor`	$3,52 \cdot 10^6$	$1,29 \cdot 10^{11}$	$75,5\%$
	log	`mslog_bed_bath_region_factor_2`	$3,57 \cdot 10^6$	$1,86 \cdot 10^{11}$	$64,5\%$
	log	`mslog_bed_bath_proptype_factor`	$3,52 \cdot 10^6$	$1,33 \cdot 10^{11}$	$74,6\%$

Table 5.1: Summary of the rounded GCV and fREML scores as well as the deviance explained for the different fitted models for the rents and for the sales

plate splines for the bivariate function depending on the longitude and latitude with 100 knots. In the above call of the function `bam()`, the default fREML procedure was used for fitting the model.

We can now compare the resulting deviance explained and the fREML score of the models. For stability in the model choice, we can also fit the models once again using the GCV as optimizing method. The corresponding scores of the GCV and the fREML and the deviance explained of the models are depicted in Table 5.1. It can be seen that for the rents as well as for the sales, the scores of the model with included factor variables for the number of bedrooms and bathrooms are smaller than for the model where they are included linearly. In addition, the corresponding deviance explained is higher for the models `mr_bed_bath_factor` and `ms_bed_bath_factor` than for the models `mr_bed_bath_linear` and `ms_bed_bath_linear`. From this it follows that model (5.2.2) is a better fit to both, the rents and the sales, than model (5.2.1).

Furthermore, we want to check, if the log link is a better choice than the identity link. Thus, the models `mrlog_bed_bath_factor` and `mslog_bed_bath_factor` are fitted, which are defined as the model (5.2.2) before, since we have seen that this model works better than model (5.2.1), but use the log link instead of the identity link. This means that the linear predictor η_i is defined as $\eta_i = \log(\mu_i)$, where μ_i is the mean renting or selling price, respectively. The basis functions and the number of knots used for the fitting procedure are chosen as before. The results are again depicted in Table 5.1. For the models with a log link, the fREML and also the GCV scores are considerably smaller than for the models with an identity link and also the deviance explained is quite higher. Thus, for the house rents as well as for the house sales data, the log link is a better choice than the identity link.

For the fitted models above, the geographic location of the house was included in the model as a bivariate function depending on the longitude and latitude, so we have assumed that the region of the house has no additional information on the house price. If this is true, then there is a smooth crossing of the price over the boundaries of the different regions. However, it would also be possible that there are regions which are more expensive than others and that there is a price jump at the boundaries. Consequently we have to check, if a model with included information about the region of each house is better than one without this information. To test this assumption, we add a new categorical variable for the region of the house to our models and obtain the models `mrlog_bed_bath_region_factor` for the rents and `mslog_bed_bath_region_factor` for the sales, defined by

$$\begin{aligned}\eta_i = \beta_0 + f_1(\texttt{time}_i) + f_2(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) \\ + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i) + \sum_{l=2}^{16} \beta_l^{\texttt{region}} \mathbb{1}_l(\texttt{region}_i). \qquad (5.2.3)\end{aligned}$$

Since we have seen above that the log link is better than the identity link, we fit these models with the log link, thus $\eta_i = \log(\mu_i)$. Furthermore, the variables `bed` and `bath` are included as factor variables. In this model, the categorical variable for the region regulates the possible jumps in the price at the boundaries of the regions. In Table 5.1 it can be seen that there is neither any great improvement in the GCV or fREML score nor in the deviance explained when information about the region is included, but the model complexity increases. It is also possible to test whether (5.2.3) fits better than model (5.2.2) via hypothesis testing. This can be done in R with the `anova.gam()` function. The call and the output of the function is as follows.

```
> anova.gam(mrlog_bed_bath_factor,
            mrlog_bed_bath_region_factor, test="F")

Analysis of Deviance Table

Model 1: PRICE ~ s(TIME, bs="cr", k=20)
            + s(AREA, bs="cr", k=20)
            + s(LONG, LAT, bs="tp", k=100)
            + factor(BED) + factor(BATH)
Model 2: PRICE ~ s(TIME, bs="cr", k=20)
            + s(LONG, LAT, bs="tp", k=100)
            + s(AREA, bs="cr", k=20)
            + factor(BED) + factor(BATH) + factor(REGION)

  Resid. Df Resid. Dev     Df Deviance      F    Pr(>F)
1    216832 8692670446
2    216817 8610807029 15.211 81863417 135.51 < 2.2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova.gam(mslog_bed_bath_factor,
        mslog_bed_bath_region_factor,test="F")

Analysis of Deviance Table

Model 1: PRICE ~ s(TIME, bs="cr", k=20)
        + s(AREA, bs ="cr", k=20)
        + s(LONG, LAT, bs="tp", k=100)
        + factor(BED) + factor(BATH)
Model 2: PRICE ~ s(TIME, bs="cr", k=20)
        + s(LONG, LAT, bs="tp", k=100)
        + s(AREA, bs="cr", k=20)
        + factor(BED) + factor(BATH) + factor(REGION)

  Resid. Df Resid. Dev     Df   Deviance      F    Pr(>F)
1    247529 3.3122e+16
2    247512 3.1898e+16 17.806 1.2241e+15 533.92 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The function anova.gam() tests, whether the more complex model of two nested models is a better choice than the simpler one. The above call of the function tests the hypothesis

$$H_0 : \beta_2^{\texttt{region}} = \cdots = \beta_{16}^{\texttt{region}} = 0$$
$$H_1 : \beta_2^{\texttt{region}}, \ldots, \beta_{16}^{\texttt{region}} \text{ arbitrary.}$$

It must be pointed out that the p-value in the output is only an approximation and thus has to be taken with care. The approximation is most accurate when the comparison is done for unpenalized terms, as in our case here. However, since we have a large dataset, it is better to look at the values of the deviance. The residual deviance compared to the residual degrees of freedom and the size of the dataset hardly decreases when adding the categorical variable region. This holds for both, the model for the house rents and the one for the house sales. Certainly, the residual deviance and the degrees of freedom are quite high for all models, escpecially for the models for the house sales, since we have big datasets. Thus, it is difficult to interpret, whether there is improvement in the deviance or not and consequently it is not advisable to use such a test in our case. Therefore, we can restrict ourselves to the use of the GCV and fREML scores for the model choice and conclude that the price of the house and the rent is smooth over the boundaries of the different regions and that there are no jumps. Hence, it is not necessary to include the region as an additional covariate in the model.

Moreover, it should be checked, whether the exact location of each house is at all necessary to explain the house price and the house rent or if the region of the house is sufficient. To check this speculation, we fit new models similar to the model (5.2.3), but the bivariate function depending on the longitude and latitude of the house is excluded. The models `mrlog_bed_bath_region_factor_2` and `mslog_bed_bath_region_factor_2` thus have the form

$$\begin{aligned}\eta_i = \beta_0 + f_1(\texttt{time}_i) + f_2(\texttt{area}_i) + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i)\\ + \sum_{l=2}^{16} \beta_l^{\texttt{region}} \mathbb{1}_l(\texttt{region}_i). \qquad (5.2.4)\end{aligned}$$

The corresponding scores of the models of interest are listed in Table 5.1. We can see, that the fREML and the GCV scores get much higher if we exclude the longitude and the latitude information and also the deviance explained gets lower. Hence, the exact location of the house is an important information for the house price and the weekly rents and only the region of the house is not sufficient for modeling the renting and selling prices.

As a last step, we want to analyse the influence of the type of the house on the house price as well as on the house rent. Chapter 4 has shown, that it is not clear whether the type of the house gives important information additional to the location of the house, since the different types of houses are located in specific regions. Thus, it might be sufficient, to just include the location of the house in the model. So we try a last model, named for the rents as `mrlog_bed_bath_proptype_factor` and for the sales as `mslog_bed_bath_proptype_factor`, which is defined as

$$\begin{aligned}\eta_i = \beta_0 + f_1(\texttt{time}_i) + f_2(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i)\\ + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i) + \sum_{l=2}^{5} \beta_l^{\texttt{proptype}} \mathbb{1}_l(\texttt{proptype}_i). \qquad (5.2.5)\end{aligned}$$

When comparing the scores of this model in Table 5.1 with the scores of the model without the categorical variable `proptype`, we can see that there is hardly any improvement in the goodness-of-fit if we include the variable `proptype` to the model, but the model complexity increases. Therefore, the location of the house is sufficient for modeling the house price and also the weekly rents and it is not necessary to additionally include information about the type of the house.

To sum up, we have seen that a model with a log link which includes the date of advertisement or the date of the sale of the house, the size of the land area as univariate smooth functions, the longitude and latitude as a bivariate smooth function and the number of bedrooms and bathrooms as categorical variables fits the given data about rented and sold houses best. Additional information about the region or the type of the house does not improve the goodness-of-fit. Thus, the best model compared to the others is the model `mrlog_bed_bath_factor` for the rents and `mslog_bed_bath_factor` for the sales. To this point, we do not know, if the normal distribution is a good choice for the GAM. In the next section, we are going to compare the results of normally distributed prices with models fitted with the gamma distribution.

5.2.2 House Price Model with Gamma distributed Prices

In the above section, we assumed the house prices and the weekly rents to be normally distributed. However, it is also possible, that they follow another member of the exponential family, for example the gamma distribution. In this section, we want to fit models using the gamma distribution instead of the normal distribution and compare the results with the ones before. We have already seen, that the log link is a better choice than the identity link and that we should include the variables `bed` and `bath` as categorical variables, the variables `time` and `area` as univariate smooth terms and the geographical position of the house, given as longitude and latitude, as bivariate smooth term. In the case of normally distributed house prices and rents, there was hardly any improvement in the goodness-of-fit when including additional information about the region or the type of the house, respectively. We now want to refit the models (5.2.2), (5.2.3) and (5.2.5) using the gamma distribution instead of the normal distribution. This yields the models `mrgamma_bed_bath_factor`, `mrgamma_bed_bath_region_factor` and `mrgamma_bed_bath_proptype_factor` and the analogue ones for the sales. The corresponding scores of the smoothing procedure and the deviance explained are shown in Table 5.2. When comparing the values in this table with the ones in Table 5.1 before, we can see that for each model, the gamma distribution is a better choice than the normal distribution, since the deviance explained of the models based on the gamma distribution is quite higher than the one for the models using the normal distribution. Furthermore, like before in the case of normally distributed prices, the additional information about the region and the type of the house, respectively, has hardly any improvement on the scores or on the deviance explained. Thus, the variables `proptype` and `region` do not need to be included in the model, since they only increase the model complexity and not the goodness-of-fit. The best model therefore is the model `mrgamma_bed_bath_factor` for the rents and the model `msgamma_bed_bath_factor` for the sales.

Until now it is not clear, whether there exists interaction between the variables and therefore, if it is necessary to include interaction terms in the model. It could be possible, for example, that the geographical behaviour of the price changes over time. To check this assumption, we fit the model `mrgamma_bed_bath_factor` and the model `msgamma_bed_bath_factor` two times, once for the years 2008 to 2010 and once for the years 2012 to 2014. With these models it is possible to test, if the spatial structure of the price is the same for the different years. If the spatial structure is the same, then there is no interaction between the geographical position and the time. To plot the spatial structure of the price and the rent under the model, the function `vis.gam()` can be used. This function allows us to produce perspective or contour plots of model predictions. The call of the function is

	model	fREML	GCV	Deviance explained
rents	mrgamma_bed_bath_factor	$-6,01 \cdot 10^3$	$5,52 \cdot 10^{-2}$	85,1%
	mrgamma_bed_bath_region_factor	$-7,64 \cdot 10^3$	$5,54 \cdot 10^{-2}$	85,2%
	mrgamma_bed_bath_proptype_factor	$-7,54 \cdot 10^3$	$5,44 \cdot 10^{-2}$	85,2%
sales	msgamma_bed_bath_factor	$-1,21 \cdot 10^4$	$5,29 \cdot 10^{-2}$	89,0%
	msgamma_bed_bath_region_factor	$-1,29 \cdot 10^4$	$5,26 \cdot 10^{-2}$	89,1%
	msgamma_bed_bath_proptype_factor	$-1,28 \cdot 10^4$	$5,47 \cdot 10^{-2}$	89,1%

Table 5.2: Summary of the rounded fREML and GCV scores as well as the deviance explained for the different models using the gamma distribution

```
> vis.gam(mr2008_2010, view=c("LONG","LAT"), n.grid=100,
        theta=-135, too.far=0.05)
```

The argument view in the function vis.gam() determines the two main effect terms of the model mr2008_2010 which are displayed on the x and y dimensions of the plot. The argument n.grid fixes the number of grid nodes in each direction used for calculating the plotted surface. The argument of theta is set to $-135°$ because then the coordinates are being rotated. In our case, we rotate the coordinates in order to have a better perspective on the spatial effect of the house price. The last argument, too.far, excludes grid nodes that are too far from the points defined by the variables. Another possibility for showing the spatial effect of the house price is the call of the function vis.gam() and setting the plot.type argument to "contour". In this case, the spatial structure of the price is plotted as a two-dimensional plot.

To see if the spatial strucure changes over time, we plot the renting and selling prices depending on the longitude and latitude for the models from 2008 to 2010 (mr2008_2010 and ms2008_2010) and for the models from 2012 to 2014 (mr2012_2014 and ms2012_2014) with the function vis.gam(), using the call of the function before. The two plots for the rents between 2008 and 2010 and from 2012 to 2014, respectively, can be found in Figure 5.1 and those for the sales in Figure 5.2. Since there is hardly any difference in the spatial structure of the different years neither for the rents nor for the sales, there is no evidence, that there is interaction between the time and the geographical position.

Furthermore, we can compare the estimated parameters for the models from 2008 to 2010 with the ones from 2012 to 2014 to see, if there is interaction between the number of bedrooms and bathrooms and the time. The relevant part of the output from the function summary() for each model is as follows.

```
> summary(mr2008_2010)

Parametric coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.628485   0.005234 1075.44   <2e-16 ***
factor(BED)2  0.343922   0.005566   61.78   <2e-16 ***
factor(BED)3  0.527003   0.005385   97.86   <2e-16 ***
factor(BED)4  0.676179   0.005714  118.34   <2e-16 ***
factor(BED)5  0.774080   0.006730  115.02   <2e-16 ***
factor(BED)6  0.786985   0.010983   71.66   <2e-16 ***
factor(BATH)2 0.154919   0.002076   74.64   <2e-16 ***
factor(BATH)3 0.393405   0.003584  109.76   <2e-16 ***
factor(BATH)4 0.640802   0.008271   77.47   <2e-16 ***
factor(BATH)5 0.870727   0.018313   47.55   <2e-16 ***
factor(BATH)6 1.013150   0.037996   26.66   <2e-16 ***

> summary(mr2012_2014)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.769947   0.004766 1210.67   <2e-16 ***
factor(BED)2  0.334690   0.005168   64.77   <2e-16 ***
factor(BED)3  0.521313   0.004925  105.86   <2e-16 ***
factor(BED)4  0.649659   0.005249  123.78   <2e-16 ***
factor(BED)5  0.744294   0.006257  118.96   <2e-16 ***
factor(BED)6  0.752961   0.010522   71.56   <2e-16 ***
factor(BATH)2 0.130956   0.002020   64.81   <2e-16 ***
factor(BATH)3 0.329142   0.003580   91.95   <2e-16 ***
factor(BATH)4 0.529367   0.008155   64.91   <2e-16 ***
factor(BATH)5 0.732827   0.018251   40.15   <2e-16 ***
factor(BATH)6 0.962926   0.035963   26.77   <2e-16 ***

> summary(ms2008_2010)

Parametric coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   13.054340   0.014351 909.677  < 2e-16 ***
factor(BED)2   0.096007   0.014526   6.609 3.89e-11 ***
factor(BED)3   0.197130   0.014411  13.680  < 2e-16 ***
factor(BED)4   0.292893   0.014533  20.153  < 2e-16 ***
factor(BED)5   0.345180   0.014793  23.333  < 2e-16 ***
factor(BED)6   0.335229   0.016046  20.892  < 2e-16 ***
factor(BATH)2  0.125556   0.002097  59.880  < 2e-16 ***
factor(BATH)3  0.300383   0.003196  93.992  < 2e-16 ***
factor(BATH)4  0.548116   0.006484  84.540  < 2e-16 ***
factor(BATH)5  0.814896   0.012668  64.328  < 2e-16 ***
factor(BATH)6  0.971672   0.027397  35.467  < 2e-16 ***

> summary(ms2012_2014)
```

```
Parametric coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)   13.347398   0.012366 1079.349  < 2e-16 ***
factor(BED)2   0.038345   0.012547    3.056  0.00224 **
factor(BED)3   0.123736   0.012423    9.960  < 2e-16 ***
factor(BED)4   0.213588   0.012516   17.066  < 2e-16 ***
factor(BED)5   0.265417   0.012666   20.955  < 2e-16 ***
factor(BED)6   0.252851   0.013374   18.906  < 2e-16 ***
factor(BATH)2  0.097272   0.001754   55.445  < 2e-16 ***
factor(BATH)3  0.237096   0.002526   93.854  < 2e-16 ***
factor(BATH)4  0.435769   0.004561   95.543  < 2e-16 ***
factor(BATH)5  0.683118   0.009394   72.718  < 2e-16 ***
factor(BATH)6  0.847676   0.018209   46.552  < 2e-16 ***
```

We want to compare the parameter estimates associated with the number of bedrooms and bathrooms for the different periods. When looking at the summary of the models for the rents, `mr2008_2010` and `mr2012_2014`, we can see that the parameter estimates for the factor variables `bed` and `bath` are nearly the same. Between the years 2008 and 2010 the rent of a house with six bedrooms was $\exp(0,787) = 2,197$ times more expensive than a house with one bedroom and between the years 2012 and 2014 $\exp(0,753) = 2,123$ times. Compared with a house with one bathroom, the rent of a house with two bathrooms was $\exp(0,155) = 1,168$ times more expensive in the years 2008 to 2010 and $\exp(0,131) = 1,14$ times in the years 2012 to 2014. The difference in the parameters of the models for the sales is also very small and thus, does not have much impact on the model output. As there is hardly any difference in the parameter estimates for the different years, we assume the variables `bed` and `time` as well as the variables `bath` and `time` to have no interaction, since an additional interaction term only deteriorates the complexity of the models.

Next we want to check, if there is interaction between the time and the size of the land area. To do so, we plot the dependence of the rent and the house price, respectively, on the size of the land area for the models from the years 2008 to 2010 and from 2012 to 2014. This can be done in `R` with the function `plot.gam()`. This function plots the component smooth functions of the models as well as the parametric model components on the scale of the linear predictor. The third plot we obtain is the renting or selling price, respectively, depending on the size of the land area. The corresponding plots are depicted in Figure 5.3. We can see that the behaviour of the sales in the years 2008 to 2010 depending on the size of the land area is essentially the same as in the years 2012 to 2014. The same holds for the weekly rents, but with a slight deviation. Since there is not much difference in the behaviour of the models for the different years, we can assume that there is no interaction between the variables `time` and `area`.

We further assume that there is no interaction between the remaining variables, since interaction complicates the model and we have already analysed the most important interactions which include the variable `time`. Thus, we have found our "best" model. The model for describing the weekly rents and the house prices has the form

$$\eta_i = \beta_0 + f_1(\texttt{time}_i) + f_2(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i), \tag{5.2.6}$$

where $\eta_i = \log(\mu_i)$ with the mean of the weekly rent or house price μ_i of house i, where the weekly rent and the house price follow the gamma distribution.

Since we have found our models for describing the trend of the renting and selling prices of houses in Sydney, we are now able to analyse and describe them depending on the covariates. Firstly, we want to get an overview of the estimated parameters of the models. The output of the function `summary()` for the model (5.2.6) fitted to the rents and to the sales is as follows.

```
> summary(mrgamma_bed_bath_factor)

Family: Gamma
Link function: log

Formula:
PRICE ~ s(TIME, bs = "cr", k = 20) + s(LONG, LAT, bs = "tp",
    k = 100) + s(AREA, bs = "cr", k = 20) + factor(BED)
    + factor(BATH)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.710912   0.003239 1763.15   <2e-16 ***
factor(BED)2  0.338494   0.003485   97.12   <2e-16 ***
factor(BED)3  0.523398   0.003342  156.60   <2e-16 ***
factor(BED)4  0.660181   0.003559  185.51   <2e-16 ***
factor(BED)5  0.755145   0.004229  178.54   <2e-16 ***
factor(BED)6  0.762948   0.007077  107.80   <2e-16 ***
factor(BATH)2 0.142552   0.001348  105.77   <2e-16 ***
factor(BATH)3 0.360280   0.002361  152.61   <2e-16 ***
factor(BATH)4 0.587739   0.005410  108.65   <2e-16 ***
factor(BATH)5 0.804727   0.012162   66.17   <2e-16 ***
factor(BATH)6 0.958933   0.024333   39.41   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Approximate significance of smooth terms:
              edf Ref.df       F p-value
s(TIME)     15.03  17.11 1492.41  <2e-16 ***
s(LONG,LAT) 97.89  98.98 2954.61  <2e-16 ***
s(AREA)     14.33  16.29   10.23  <2e-16 ***
```

```
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

R-sq.(adj) =  0.724   Deviance explained = 85.1%
fREML score = -6009.6  Scale est. = 0.055199  n = 216976

> summary(msgamma_bed_bath_factor)

Family: Gamma
Link function: log

Formula:
PRICE ~ s(TIME, bs = "cr", k = 20) + s(LONG, LAT, bs = "tp",
    k = 100) + s(AREA, bs = "cr", k = 20) + factor(BED)
    + factor(BATH)

Parametric coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)   13.203982   0.008796 1501.061   <2e-16 ***
factor(BED)2   0.062708   0.008914    7.035    2e-12 ***
factor(BED)3   0.156700   0.008835   17.737   <2e-16 ***
factor(BED)4   0.248540   0.008904   27.914   <2e-16 ***
factor(BED)5   0.299505   0.009033   33.158   <2e-16 ***
factor(BED)6   0.287754   0.009655   29.802   <2e-16 ***
factor(BATH)2  0.113074   0.001264   89.476   <2e-16 ***
factor(BATH)3  0.268954   0.001873  143.611   <2e-16 ***
factor(BATH)4  0.488596   0.003575  136.683   <2e-16 ***
factor(BATH)5  0.745636   0.007274  102.503   <2e-16 ***
factor(BATH)6  0.920806   0.014531   63.368   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Approximate significance of smooth terms:
              edf  Ref.df    F p-value
s(TIME)      17.03  18.50 4010  <2e-16 ***
s(LONG,LAT)  98.73  99.00 8128  <2e-16 ***
s(AREA)      17.73  18.72 1822  <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

R-sq.(adj) =  0.725   Deviance explained =   89%
fREML score = -12068  Scale est. = 0.052908  n = 247659
```

The output shows us that due to missing values in the two datasets, the number of data used for fitting the weekly rents was 216.976 and the number of data used for fitting the house sales was 247.659. We now want to analyse the impact of the number of bedrooms and bathrooms on the weekly rents. Compared to a house with only one bedroom, a house with two bedrooms is $\exp(0,338) = 1,402$ times more expensive and a house with six bedrooms even $\exp(0,763) = 2,145$ times. Moreover, the rent of a house with six bathrooms is $\exp(0,959) = 2,609$ times more expensive than the one of a house with only one bathroom. This analysis works similar for the sale prices. A house with two bedrooms has a $\exp(0,063) = 1,065$ times higher price compared to a house with only one bedroom and a house with two bathrooms is $\exp(0,113) = 1,12$ times more expensive than a house with one bathroom. This once again shows us that the variables `bed` and `bath` should not be included linearly, since the raise in the weekly rents and the house price is not linear in the number of bedrooms and bathrooms. For example, the increase of the weekly rent from a house with five bedrooms to a house with six bedrooms is less than the raise in the rent between a house with two bedrooms and one with three bedrooms. It can also be seen in the output that the significance of all covariates is very high. Furthermore, the function `summary()` shows us the smooth terms with corresponding p-values and effective degrees of freedom. It can be seen, that $k - 1$ is an upper bound for the effective degrees of freedom for each term, where k is the number of knots used for the splines. In the last two lines of the output, we can see the corresponding fREML scores and the deviance explained, which were already discussed before. Moreover, the estimated scale or dispersion parameter $\hat{\phi}$ is given in the last line. The meaning of this parameter was already discussed when the exponential family was introduced in Section 2.2. In case of the gamma distribution, the variance can be estimated as

$$\widehat{\text{var}}(y) = \hat{\phi}\hat{\mu}^2,$$

where $\text{E}(y) = \mu$. This means that the variance is proportional to the square of the mean. Thus, the standard deviation is $\sqrt{0.05} = 0.2236$ times the estimated mean prices.

It is also possible to produce some residual plots and to obtain further information about the fitting process. This can be achieved with the function `gam.check()` in `R`. With the following code we can get some diagnostic information about the fitting procedure and results for our two models.

```
> gam.check(mrgamma_bed_bath_factor)

Method: fREML   Optimizer: perf newton
full convergence after 7 iterations.
Gradient range [-5.04982e-06,4.812226e-06]
(score -6009.572 and scale 0.05519938).
Hessian positive definite, eigenvalue range [3.888066,108480.5].

Basis dimension (k) checking results. Low p-value (k-index<1)
may indicate that k is too low, especially if edf is close to k'.

                  k'    edf k-index p-value
s(TIME)      19.000 15.029   0.988    0.20
s(LONG,LAT) 99.000 97.893   0.859    0.00
s(AREA)      19.000 14.326   0.980    0.14

> gam.check(msgamma_bed_bath_factor)
```

```
Method: fREML   Optimizer: perf newton
full convergence after 9 iterations.
Gradient range [-5.173933e-07,5.100256e-07]
(score -12068.22 and scale 0.05290794).
Hessian positive definite, eigenvalue range [7.022519,123822].

Basis dimension (k) checking results. Low p-value (k-index<1)
may indicate that k is too low, especially if edf is close to k'.

                 k'    edf k-index p-value
s(TIME)      19.000 17.033   0.988    0.26
s(LONG,LAT)  99.000 98.734   0.730    0.00
s(AREA)      19.000 17.725   0.979    0.10
```

This output gives us information about the fitting procedure and the corresponding results. Furthermore, a test whether the basis dimension for a smooth term is adequate was done. This test works as follows: an estimate of the residual variance is calculated based on differencing residuals that are near neighbours according to the covariates of the smooth. The resulting estimate is then divided by the residual variance. This ratio is then called `k-index` and can be found in the output. Since all the values for the `k-index` are close to 1, it is not very likely that there is a missed pattern left in the residuals. The p-values of all smooth terms are very low and suggest, that every term is significant. In general, low p-values may indicate that the basis dimension k is too low, especially if the reported effective degrees of freedom are close to k', which is an upper bound for the effective degrees of freedom ($k' = k - 1$). The corresponding plots are depicted in Figure 5.4. The first plot at the upper left shows extreme tails deviating from the straight line of the QQ plot. This is due to the fact that we have very large datasets and thus, there will not be a perfect distributional assumption. Despite these deviations we will assume our house rents and house prices to follow a gamma distribution, since the earlier calculated fREML and GCV scores as well as the deviance explained of the models using the gamma distribution were quite good. The plot on the upper right suggests that the variance is approximately constant as the mean increases for both, the rents and the sales. Furthermore, the lower left histograms seem to be approximately consistent with normality. Last but not least, the lower right plot shows in both cases the relationship between the response and the fitted values. This relationship seems to be positive, which is not problematic.

Moreover, we can plot the component smooth functions of the models as well as the parametric model components on the scale of the linear predictor. In `R`, this can be done with the function `plot.gam()`. This function returns three different plots in our case. The first two plots for the rents are shown in Figure 5.5 and for the sales in Figure 5.7, whereas the third plot gives an two-dimensional overview of the price depending on the location in Sydney, which is not depicted here, since the later used `vis.gam()` function gives a better insight into the geographical behaviour. When looking at the top of Figure

5.5, we can see the trend of the weekly rents depending on the date of advertisement. The function is almost linear. Thus, we can assume the weekly rents to increase linearly over time. The plot at the bottom shows the dependence of the weekly rent prices depending on the size of the land area. We can see that there is a rise in the rent prices between a land area of about 100 to approximately 3.500 square meters. Beginning at a land area of about 3.500 square meters, there is a slight decrease in the rent prices up to a size of 5.000 square meters. This behaviour may be explained by the fact that houses with a larger land area are more frequently located in the cheaper, outer regions of Sydney. Furthermore, the area defines the land area of the house. Thus the actual size of the house may be much smaller than the land area. A very similar behaviour can be seen in Figure 5.7 in the case of house prices. Only the house prices depending on the time deviate slightly from a linear function. The trend of the house prices depending on the area shows the same behaviour as the weekly rents.

The output of the function `vis.gam()` is shown in Figure 5.6 for the rents and in Figure 5.8 for the sales. The plots for the rents and for the sales look very similar. The two plots at the top show the house price and the rent price, respectively, depending on the geographical position of the house. Care has to be taken, since the plot has been rotated. Thus the most expensive areas are the regions in the East at the coast of Sydney, whereas regions in the west of Sydney are less expensive. The same behaviour can be seen in the plots at the bottom. The lighter the colour of the plot, the more expensive the renting or selling price is. This behaviour of the house price and the weekly rents depending on the geographical position of the house was already depicted in Figure 4.4 in Chapter 4. Again care has to be taken, since the plots at the top and at the bottom of Figures 5.6 and 5.8 are not directly comparable, since the plots at the top are rotated.

When looking at the function depending on the size of the land area at the bottom of Figure 5.5 and Figure 5.8, we can see that the function nearly behaves like the logarithm function. Hence, we finally want to compare the above “best“ model with a model, where the smooth term depending on the size of the land area is replaced by the logarithm to simplify and stabilize the model. The models `mrgamma_bed_bath_factor_log_area` and `msgamma_bed_bath_factor_log_area` are defined by

$$\eta_i = \beta_0 + f_1(\mathtt{time}_i) + \beta^{\mathtt{area}} \log(\mathtt{area}_i) + f_{geo}(\mathtt{long}_i, \mathtt{lat}_i) + \sum_{j=2}^{6} \beta_j^{\mathtt{bed}} \mathbb{1}_j(\mathtt{bed}_i)$$

$$+ \sum_{k=2}^{6} \beta_k^{\mathtt{bath}} \mathbb{1}_k(\mathtt{bath}_i), \tag{5.2.7}$$

with the log link $\eta_i = \log(\mu_i)$. The corresponding AIC and deviance explained of the models (5.2.6) and (5.2.7) are depicted in Table 5.3. We can see, that the AIC for the model with the general smooth term for the variable `area` is hardly smaller than the one for the model with the logarithm term. Moreover, the deviance explained is nearly the same for both models. To sum up, we can conclude that model (5.2.7) fits the house rent

model	AIC	Deviance explained
mrgamma_bed_bath_factor	2.626.177	$85,1\%$
mrgamma_bed_bath_factor_log_area	2.626.346	$85,1\%$
msgamma_bed_bath_factor	6.593.123	$89,0\%$
msgamma_bed_bath_factor_log_area	6.594.721	$88,9\%$

Table 5.3: Summary of the AIC as well as the deviance explained for model (5.2.6) and model (5.2.7) for the rents and for the sales

and the house sale data as good as model (5.2.6), but is clearly simpler and thus to prefer.

5.2.3 Joint House Price Model for Renting and Selling Prices

In the section before, we have demonstrated that the renting and selling prices of houses in Sydney can be fitted with model (5.2.7). Thus, the chosen model for the rented and sold houses has the same structure, but they were fitted seperately and have different parameter estimates. When looking at Figures 5.6 and 5.8 we can see that the spatial structure for the rents and for the sales is nearly the same. Furthermore, in this section, we assume the parameters for the categorical variables `bed` and `bath` to be the same for the rents and the sales, except for the intercept, which is higher for the sales. The only function in the model that is different for the rents and the sales should be the function depending on the time, since we have seen in Figures 5.5 and 5.7 that the behaviour of the weekly rents over the time is not the same as for the house sales. We now want to fit model (5.2.7) again, using the house rents and the house sales data at once in one model. Thus, we only obtain one vector of estimated parameters for the rents and for the sales. We define our new model, based on all of the data, as

$$\begin{aligned}\eta_i = {}& \beta_0^{\texttt{rent}} \mathbb{1}_{\texttt{rent}}(\texttt{type}_i) + \beta_0^{\texttt{sale}} \mathbb{1}_{\texttt{sale}}(\texttt{type}_i) \\ & + f_1^{\texttt{rent}}(\texttt{time}_i) \mathbb{1}_{\texttt{rent}}(\texttt{type}_i) + f_1^{\texttt{sale}}(\texttt{time}_i) \mathbb{1}_{\texttt{sale}}(\texttt{type}_i) \\ & + \beta^{\texttt{area}} \log(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) \\ & + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i), \end{aligned} \tag{5.2.8}$$

where the linear predictor η_i is the logarithm of the mean price of the rent or house price of house i and the type of the house price can be either rent or sale. Thus, we assume the weekly rents and the house sales to have a different intercept $\beta_0^{\texttt{rent}}$ and $\beta_0^{\texttt{sale}}$, respectively, and the model also includes different functions depending on the time, namely $f_1^{\texttt{rent}}(\texttt{time}_i)$ and $f_1^{\texttt{sale}}(\texttt{time}_i)$. Moreover, we assume that the rents and the sales have a joint function depending on the geographical position and also that the factor variables `bed` and `bath` are the same, as well as the estimated parameter for the logarithm depending on the size of the land area.

To fit the model to all of the data, we first have to combine the datasets houserents2008 and housesales2008. This is done by first adding a new column named as TYPE, which is a categorical variable with labels "rent" and "sale", and then merging the datasets to the new dataset houses2008. With this new dataset, we can now fit our new model defined before in (5.2.8). The call in R is

```
m_all_data <- bam(PRICE ~ s(TIME, bs="cr", k=20, by=TYPE)
                + s(LONG,LAT, bs="tp", k=100) + TYPE
                + log(AREA) + factor(BED) + factor(BATH),
                family=Gamma(link=log), data=houses2008)
```

This call looks like the ones used before in Chapter 5, but now there is an additional argument in the function s() depending on the time, namely the by argument. This argument creates different smooth functions depending on the time for the rents and for the sales, like we defined the model before in (5.2.8) using the indicator function. Furthermore, we have an additional factor variable type, which enables the model to have different intercepts for the rents and for the sales. After fitting the model to the data, we can have a look at the output of the summary() function.

```
> summary(m_all_data)

Family: Gamma
Link function: log

Formula:
PRICE ~ s(TIME, bs = "cr", k = 20, by = TYPE) + s(LONG, LAT,
    bs = "tp", k = 100) + TYPE + log(AREA) + factor(BED)
    + factor(BATH)

Parametric coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.0184370  0.0072619  691.07   <2e-16 ***
TYPEsale      7.1262555  0.0007926 8991.04   <2e-16 ***
log(AREA)     0.1135285  0.0009964  113.94   <2e-16 ***
factor(BED)2  0.3326162  0.0035144   94.64   <2e-16 ***
factor(BED)3  0.4966282  0.0034037  145.91   <2e-16 ***
factor(BED)4  0.6122873  0.0035189  174.00   <2e-16 ***
factor(BED)5  0.6801033  0.0037918  179.36   <2e-16 ***
factor(BED)6  0.6743988  0.0049984  134.92   <2e-16 ***
factor(BATH)2 0.1214223  0.0010140  119.75   <2e-16 ***
factor(BATH)3 0.3052339  0.0016076  189.88   <2e-16 ***
factor(BATH)4 0.5539133  0.0032884  168.45   <2e-16 ***
factor(BATH)5 0.8243851  0.0069240  119.06   <2e-16 ***
factor(BATH)6 1.0148461  0.0138700   73.17   <2e-16 ***
---
```

```
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Approximate significance of smooth terms:
                   edf Ref.df    F p-value
s(TIME):TYPErent 15.82  17.72 1410  <2e-16 ***
s(TIME):TYPEsale 17.60  18.74 3069  <2e-16 ***
s(LONG,LAT)      98.70  99.00 9539  <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

R-sq.(adj) =  0.809   Deviance explained = 99.2%
fREML score =  29662  Scale est. = 0.066387  n = 464635
```

In the first lines of the estimated parametric coefficients, we can see that the intercept for the weekly rents is $\exp(5,018) = 151,11$, whereas the intercept for the house sales is $\exp(5,018) \cdot \exp(7,126) = 187.962,92$. The factor variables `bed` and `bath` are independent of the type, thus, for example, a house with two bedrooms is $\exp(0,333) = 1,395$ times more expensive than a house with one bedroom and a house with two bathrooms is $\exp(0,121) = 1,129$ times more expensive than one with just one bathroom. This holds for both, the rented houses as well as for the sold houses. Furthermore, the table shows us the summary of the smooth terms depending on the time and on the geographical position. As already mentioned before, there are different smooth terms depending on the time for the rented and for the sold houses, whereas there is one joint smooth term depending on the geographical position of the house. The last two lines of the output have to be interpreted carefully. The deviance explained of the model is 99,2% which is quite high, but that is due to the fact that we now have information about 464.635 houses.

The `plot.gam()` function allows us to plot the smooth terms depending on the time. The results are depicted in Figure 5.9. When comparing these plots with the plots in Figures 5.5 and 5.7 from before, where the rents and the sales where fitted seperately, we can see that the behaviour of the price depending on the time is the same, for the rents as well as for the sales. Moreover, we can have a look at the spatial structure of the model. The output of the `vis.gam()` function is illustrated in Figure 5.10. These plots can once again be compared with the plots in Figures 5.6 and 5.8. We can see that there is hardly any difference in the spatial structure of the joint model and the one for the models, where the rents and the sales were fitted seperately. To sum up, the results have shown us that the assumption of one joint model for the rented and the sold houses is hardly worse than if we assume the two datasets to behave like different models. With this result, we can now go on with our task of forecasting renting and selling prices for houses in Sydney in the next chapter.

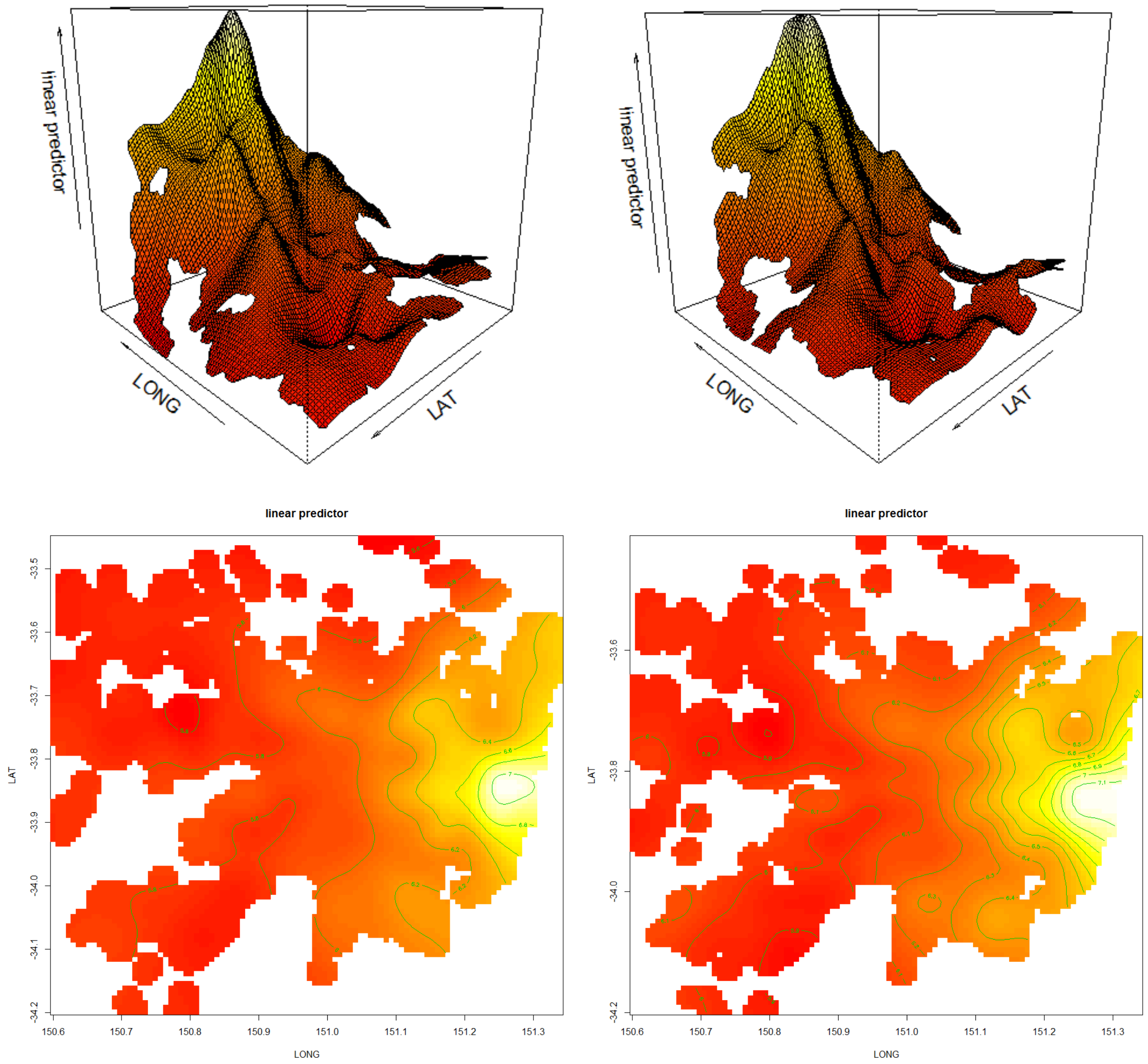


Figure 5.1: The spatial structure of the weekly rents from 2008 to 2010 (left) and from 2012 to 2014 (right)

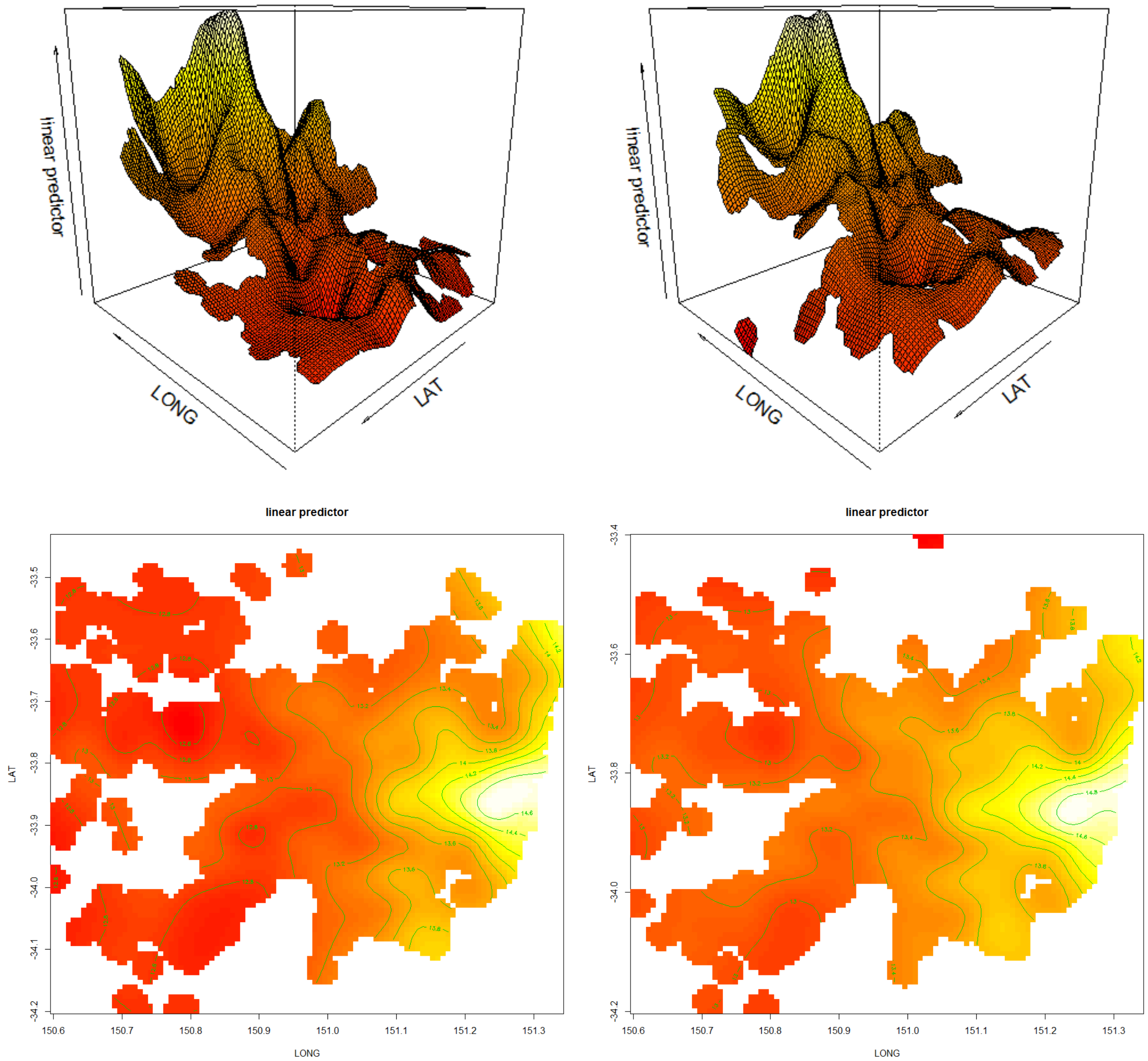


Figure 5.2: The spatial structure of the house prices from 2008 to 2010 (left) and from 2012 to 2014 (right)

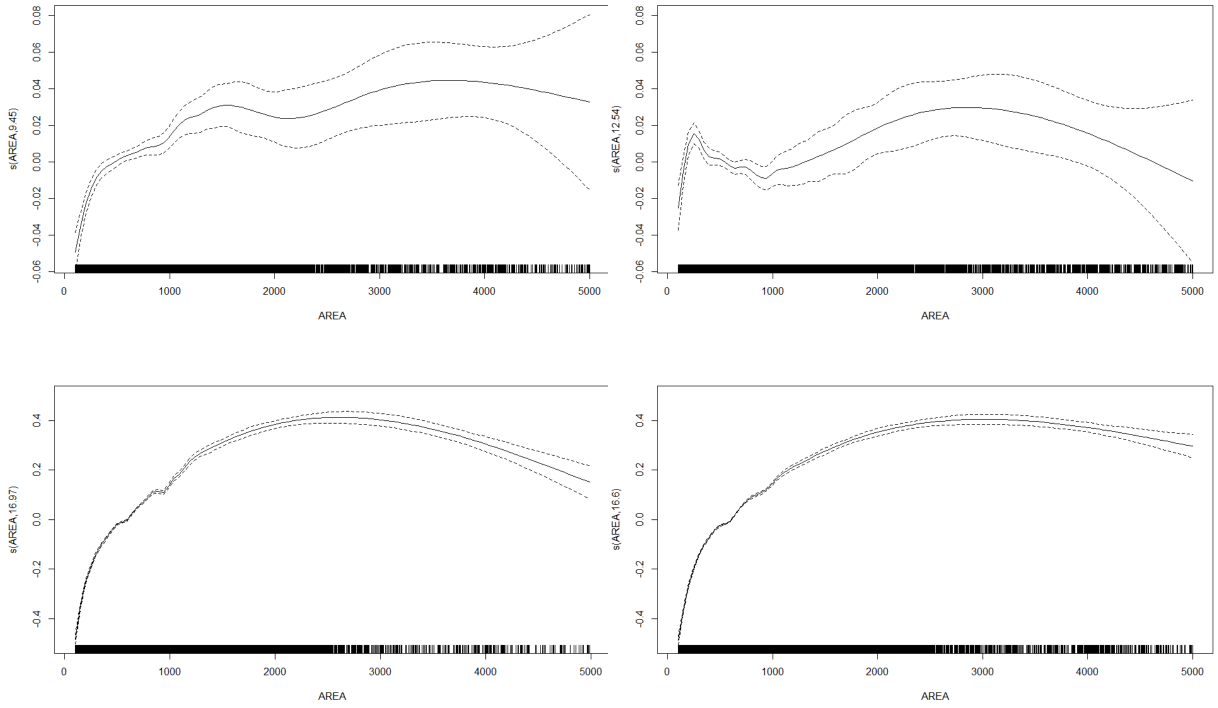


Figure 5.3: The dependence of the weekly rents on the size of the land area (top) and for the sales (bottom) for the years 2008 to 2010 (left) and for the years 2012 to 2014 (right)

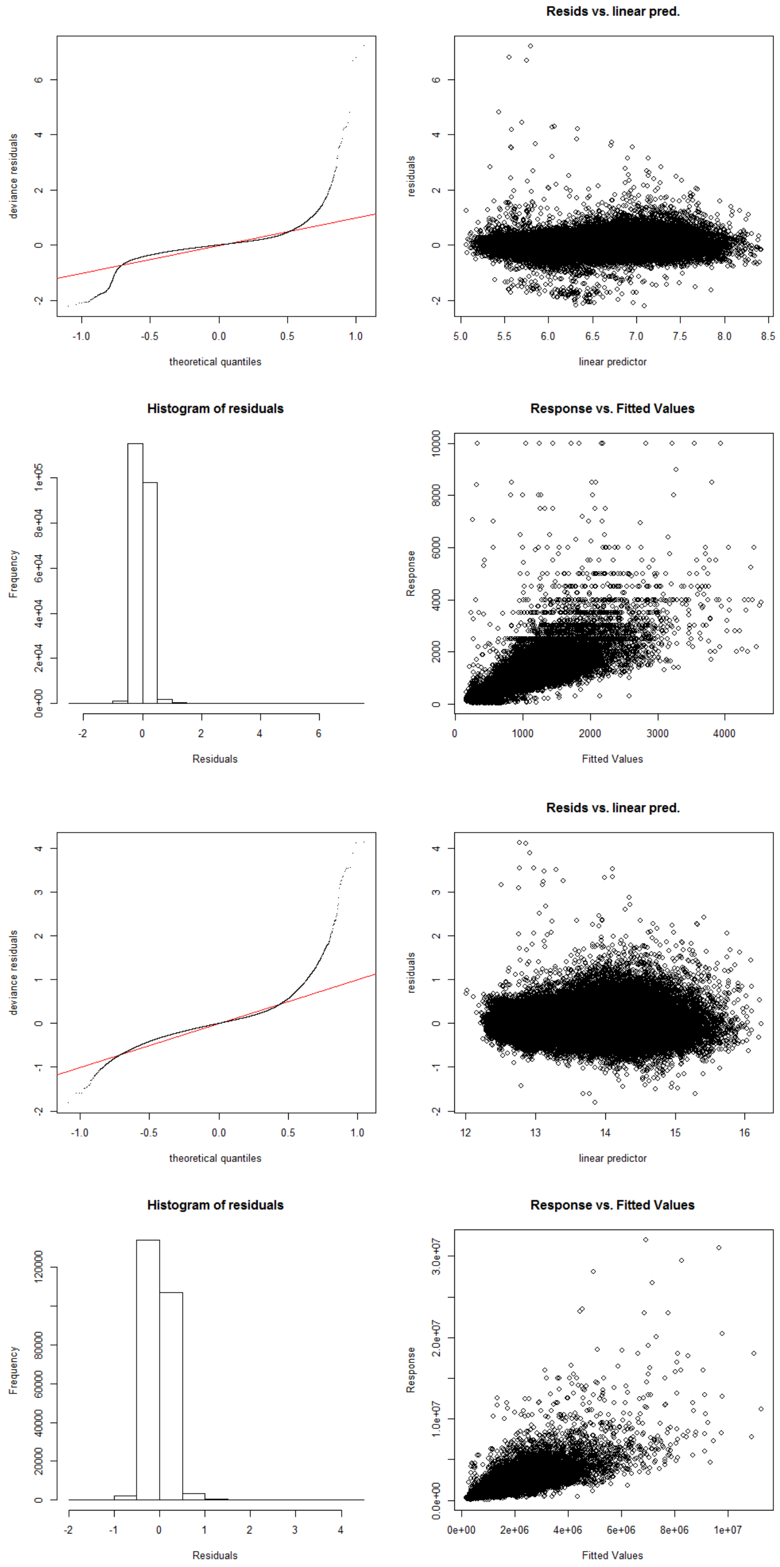


Figure 5.4: Some basic model checking plots for the models `mrgamma_bed_bath_factor` (top) and `msgamma_bed_bath_factor` (bottom)

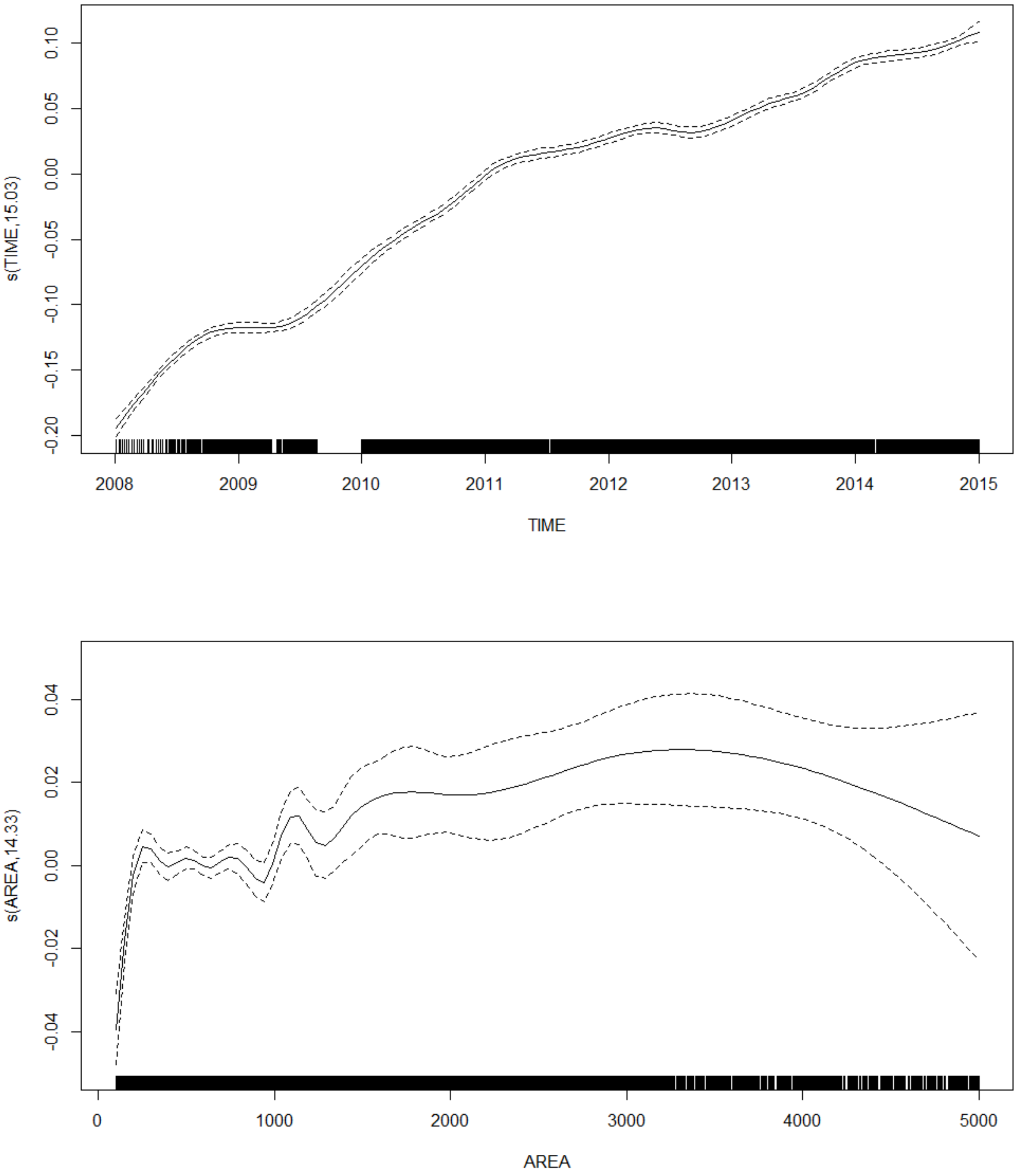


Figure 5.5: The trend of the weekly rents over the years 2008 to 2014 (top) and over the size of the land area (bottom) under the estimated model `mrgamma_bed_bath_factor`

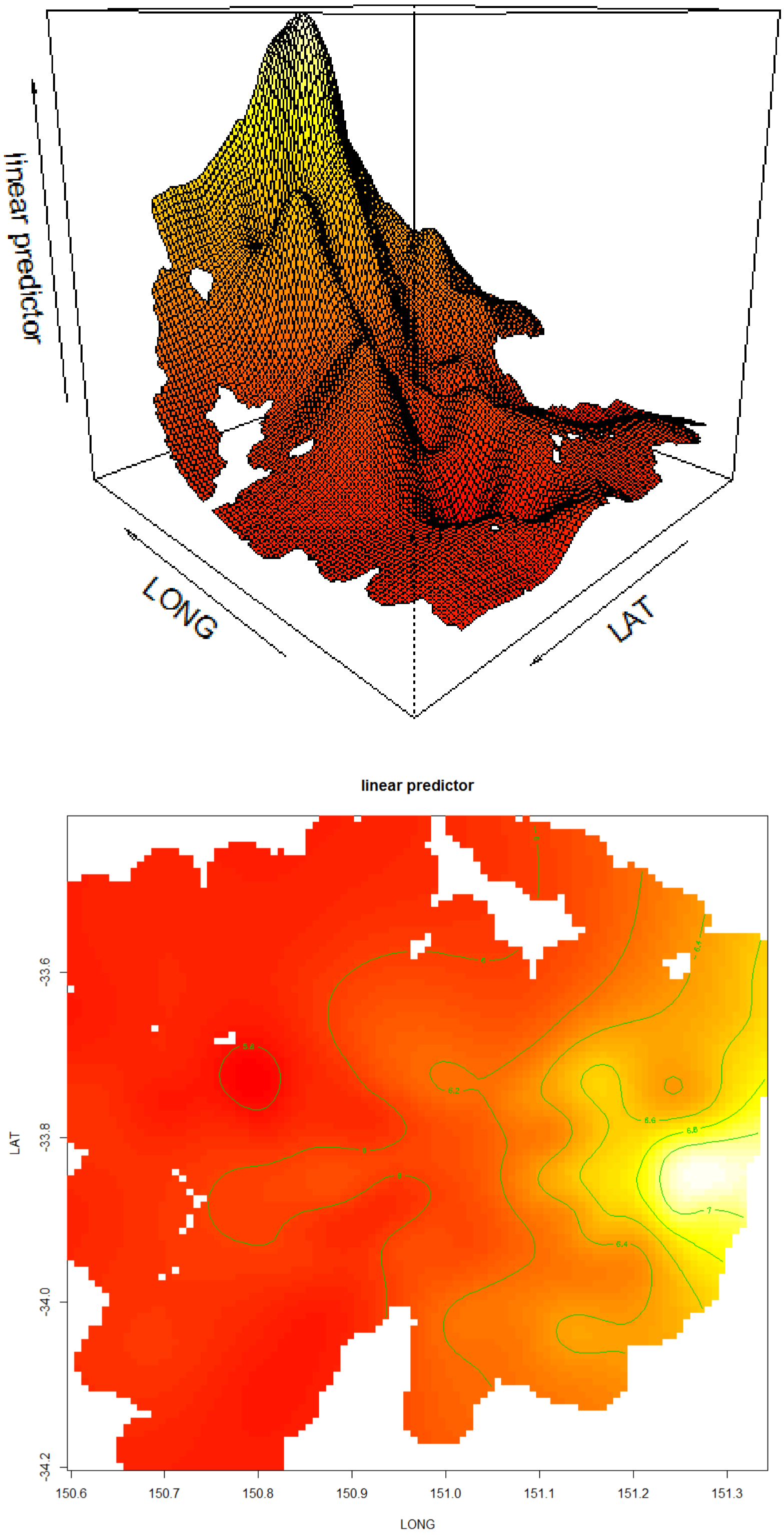


Figure 5.6: The price of the weekly rents over the region of Sydney under the estimated model `mrgamma_bed_bath_factor`

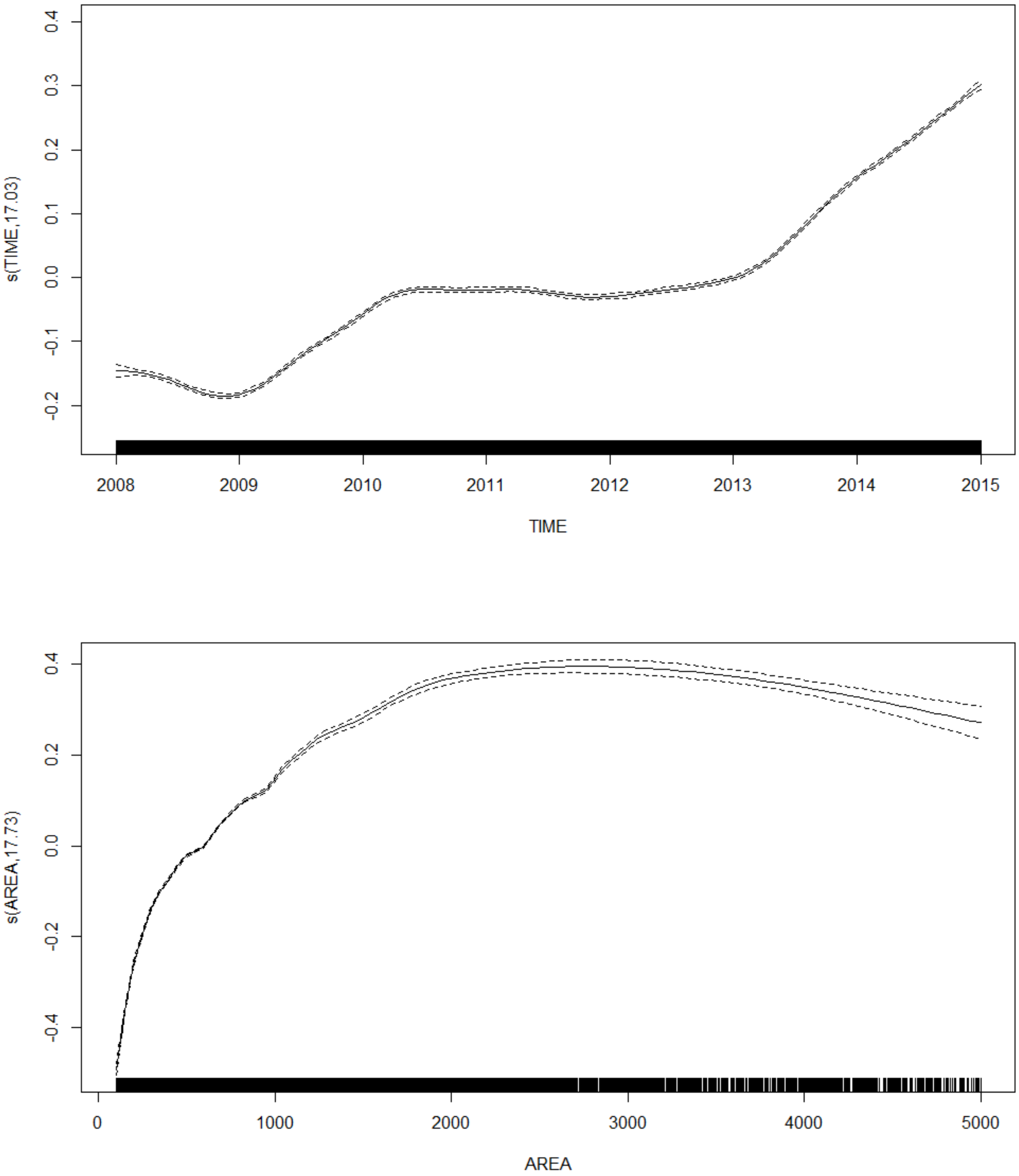


Figure 5.7: The trend of the house price over the years 2008 to 2014 (top) and over the size of the land area (bottom) under the estimated model `msgamma_bed_bath_factor`

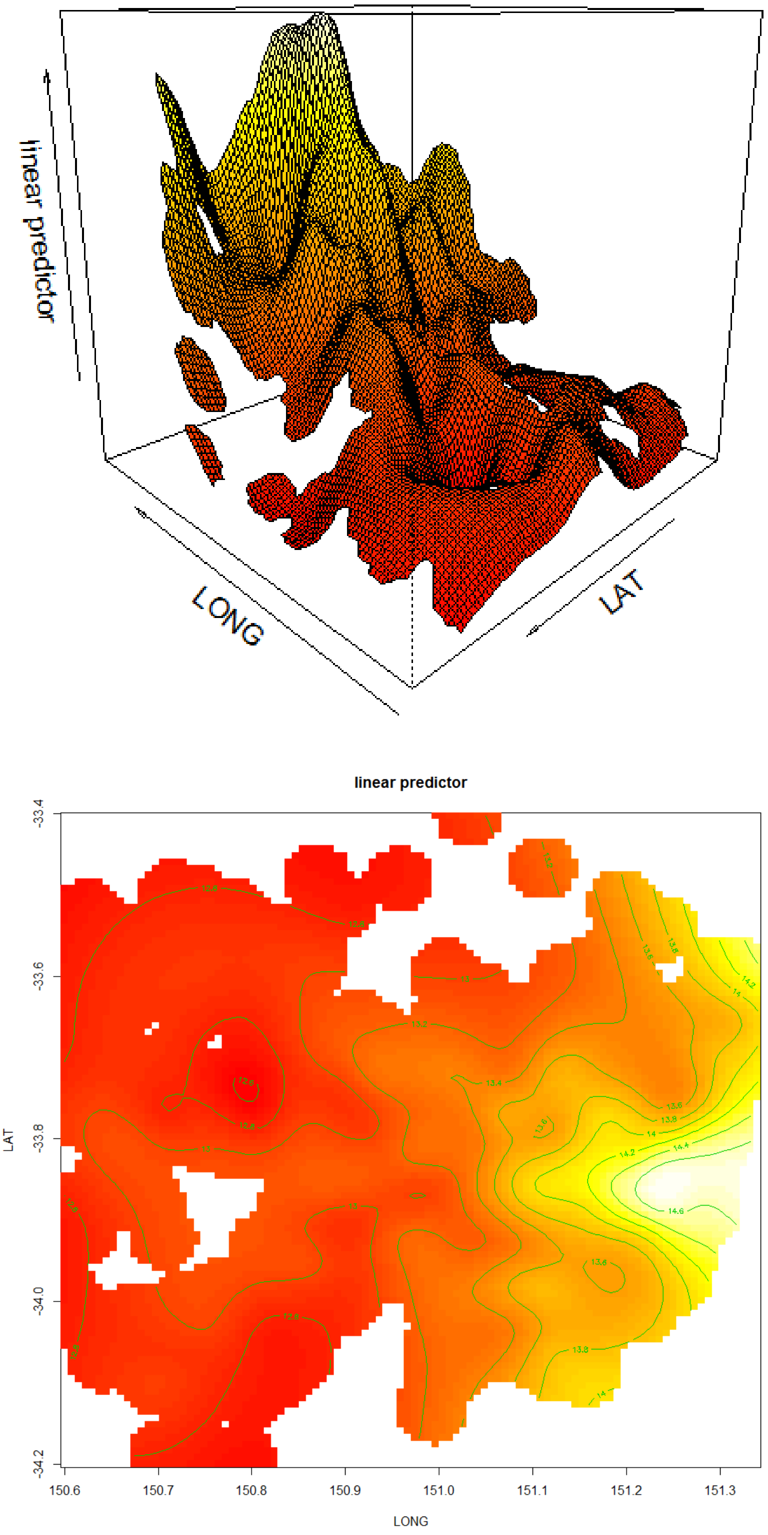


Figure 5.8: The price of the sales over the region of Sydney under the estimated model `msgamma_bed_bath_factor`

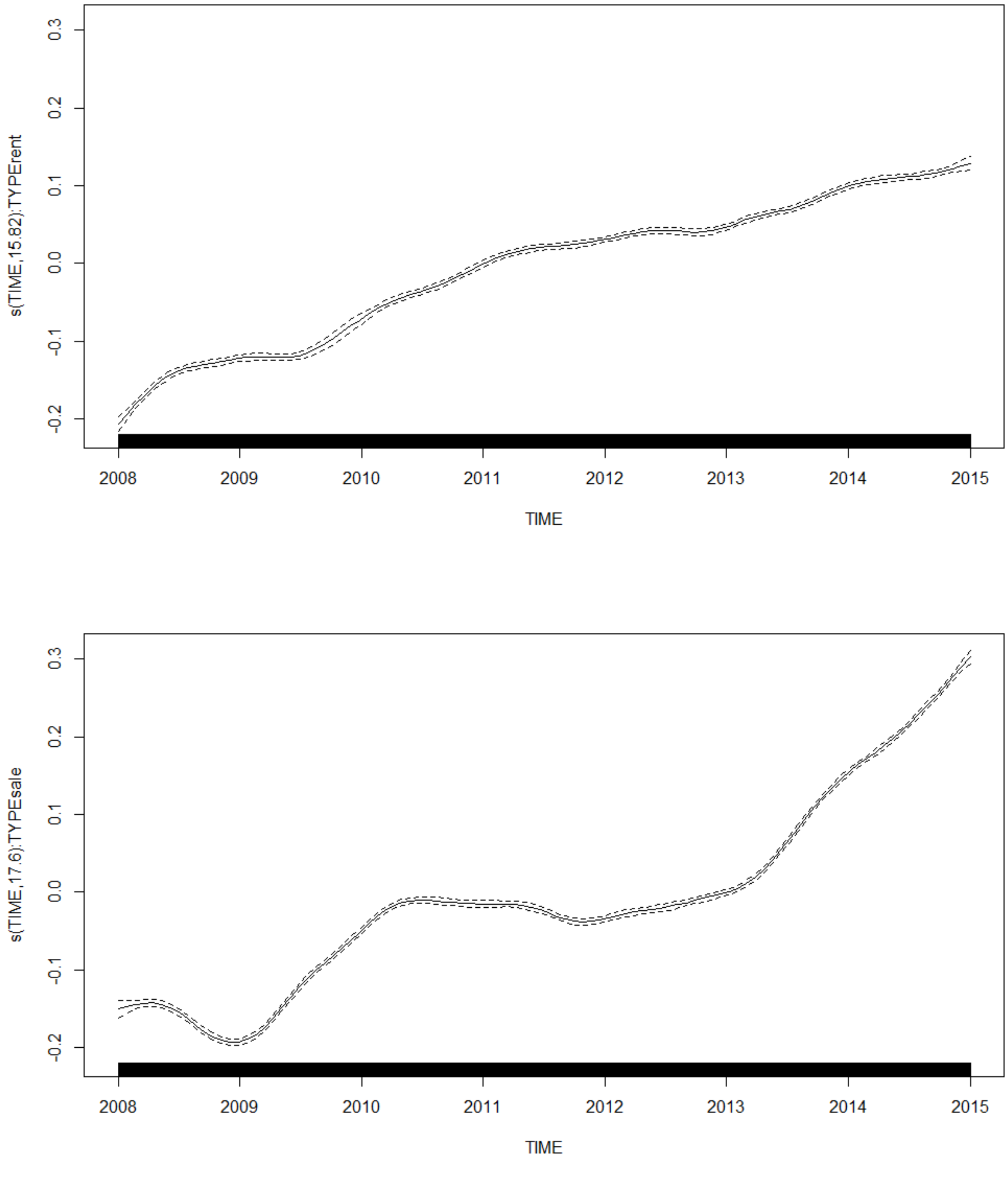


Figure 5.9: The price of the weekly rents (top) and the house sales (bottom) depending on the time under the joint model `m_all_data`

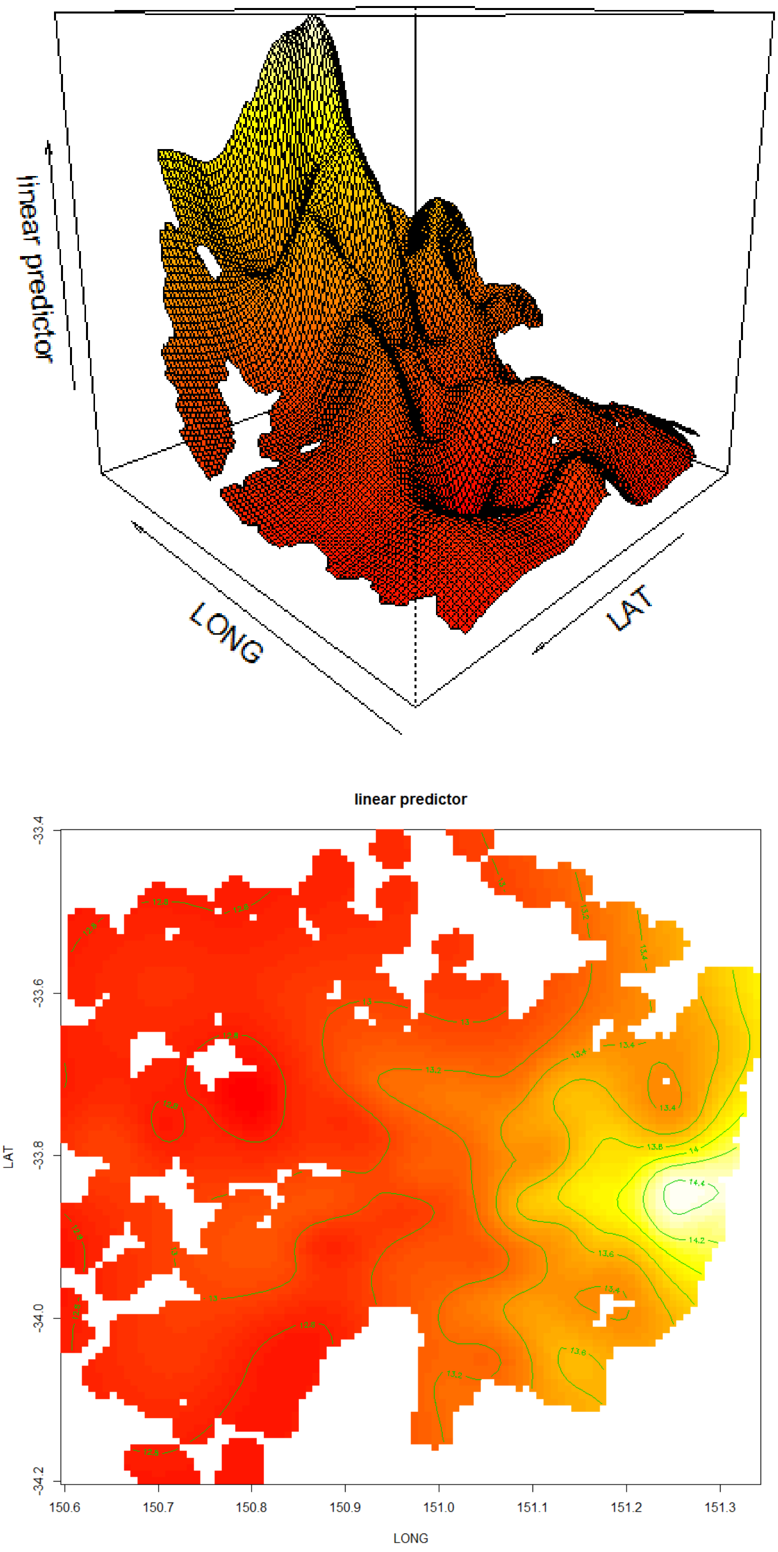


Figure 5.10: The spatial structure of the price of the weekly rents as well as for the house sales under the joint model `m_all_data`

CHAPTER 6

Forecasting House Prices for Sydney

As already mentioned before, the sudden and immense downturn in the U.S. house prices in the year 2006 triggered the global financial crisis in the year 2007 and consequently raised the interest in forecasting house prices for the near future for economic stability (Plakandaras et al., 2015). Therefore, many authors published papers about forecasting house prices. Most of them, for example Kholodilin and Siliverstovs (2014), work with a set of predictor variables divided into macroeconomic variables and confidence indexes. Housing lending rates, stock exchange price and performance indexes are classified as macroeconomic variables. an de Meulen et al. (2011) included indicators based on consumer surveys asking for example about the financial and economic situations of households as well as planned housing-related purchases and savings. They showed that the expected financial situation and the information on planned major purchase and intended savings are relevant for forecasting house prices. Other predictors like the unemployment rate, population and employment also play a decisive role in terms of forecasting, since income and employment variables give an insight into the ability of households to secure housing and thus reflect the housing demand (Rapach and Strauss, 2007).

The majority of the literature uses econometric models for forecasting. an de Meulen et al. (2011) for example, use Autoregressive Distributed Lag (ARDL) and Vector Autoregression (VAR) models and combinations of them, whereas Plakandaras et al. (2015) compare Bayesian Vector Autoregressive (BVAR) and Bayesian Autoregressive (BAR) models. Information about house characteristics are also included in the set of predictor variables when forecasting house prices. Thus, most of the literature derives models for forecasting a particular price index, which shows the average movement of an average home. Case and Shiller (1990) even found evidence of positive serial correlation in real housing price, which means that house prices are forecastable to a certain degree.

While most of the literature works with economic forecasting methods, in this thesis we want to forecast house prices with a generalized additive model derived in the previous chapter. This model allows us to forecast the house rent or the house price for a specific house with given properties and a particular location at a time of interest. Hence, in contrast to the above mentioned literature on forecasting, where only the average trend of the house price can be forecasted, macroeconomic variables such as stock exchange prices or unemployment rates are not included in the model. Hence the model assumes no major changes in the macroeconomic environment nor fundamental changes in expectations in the nearer future.

6.1 Forecasting Renting and Selling Prices

In Chapter 5 we have seen that it is possible to fit the rents and the sales in one joint model, which is defined by

$$\begin{aligned}\eta_i = {} & \beta_0^{\texttt{rent}} \mathbb{1}_{\texttt{rent}}(\texttt{type}_i) + \beta_0^{\texttt{sale}} \mathbb{1}_{\texttt{sale}}(\texttt{type}_i) \\ & + f_1^{\texttt{rent}}(\texttt{time}_i) \mathbb{1}_{\texttt{rent}}(\texttt{type}_i) + f_1^{\texttt{sale}}(\texttt{time}_i) \mathbb{1}_{\texttt{sale}}(\texttt{type}_i) \\ & + \beta^{\texttt{area}} \log(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) \\ & + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i),\end{aligned}$$

where $\eta_i = \log(\mu_i)$ and μ_i is the mean of the weekly rent or house sale of house i, respectively. Therefore, we assumed that the rents and the sales have a different intercept and a different smooth term depending on the time and that they have a joint smooth term for the spatial structure and the same parameters for the categorical variables `bed` and `bath`. In this section, we will use this model to forecast weekly rents and also house sales for the nearest future. Thus, we will predict the rent and the house price of a given house with specific properties and a particular location.

In Figure 5.9 we have seen that the increasing trend of the rents is nearly linear while the one for the sales is more wiggly. Nevertheless, the last increase of the price of sold houses between the years 2014 and 2015 is nearly linear. Therefore, forecasting can be done straighforward by extrapolating the price function depending on the time linearly. The prediction for the rents is now done for yearly rents, thus we multiply the given weekly rents by 52. The plot in Figure 6.1 is obtained with the following code.

```r
time <- seq(2008, 2016, by=0.05)
bed <- "1"
bath <- "1"
area <- 600
long <- 151
lat <- -33.8
rent.data <- expand.grid(TIME=time ,BED=bed, BATH=bath,
          AREA=area, LONG=long, LAT=lat, TYPE="rent")
pred.rents <- predict(m_all_data, newdata=rent.data,
        type="response", se=TRUE)
sale.data <- expand.grid(TIME=time, BED=bed, BATH=bath,
        AREA=area, LONG=long, LAT=lat, TYPE="sale")
pred.sales <- predict(m_all_data, newdata=sale.data,
        type="response", se=TRUE)
plot(time, (52*pred.rents$fit), type="l")
plot(time, pred.sales$fit, type="l")
```

We first define the time sequence up to the end of the calendar year 2016 and determine the characteristics of any given house of which we want to forecast the yearly rent and the house price. Thus, we have to define the number of bedrooms and bathrooms, the size of the land area and the geographic coordinates (longitude and latitude). It must be pointed out that the choice of the values is not relevant to the forecast, since the shape of the function stays the same for other given values, as the function is only shifted on the y-axis. Thus, we forecast the yearly rent and the house price of a house with one bedroom and one bathroom, a land area size of 600 square meters, which is located at a longitude of 151° and a latitude of $-33,8°$. The prediction for the rent and the house price is a linear increase up to the end of 2016, where only the height of the price on the y-axis is dependent on the chosen house characteristics.

Our forecast therefore predicts the rent and the house price of a specific house with particular properties for the nearest future. In contrast to that, most of the literature on forecasting house prices derived a house price index which shows the general movement of the house price. With this general house price index, it is not possible to forecast the price of a specific house as our model allows.

6.2 Trend of the Price-Rent Ratio

In the introduction of Chapter 5, the importance of the price-rent ratio was already mentioned. As described by Hill and Syed (2016), every durable good should follow the equilibrium condition

$$R_t = u_t P_t, \tag{6.2.1}$$

where R_t is the period t rental price and $u_t P_t$ the user costs of the good, which is the present value of buying it, using it for one period and then selling it. More precisely, the user costs consist of the purchase price P_t and the per dollar user costs u_t. The per dollar user costs can be calculated as

$$u_t = r_t + \omega_t + \delta_t + \gamma_t - g_t,$$

where r_t is the risk-free interest rate, ω_t the property tax rate, δ_t the depreciation rate for housing and γ_t denotes the risk premium of owning the house instead of renting it. Furthermore, g_t is the expected capital gain which is not directly observable. From Equation (6.2.1) it follows that

$$\frac{P_t}{R_t} = \frac{1}{u_t}. \tag{6.2.2}$$

This equilibrium condition assumes P_t and R_t to be calculated for properties of equivalent quality. The expected capital gain can then be computed as

$$g_t = r_t + \omega_t + \delta_t + \gamma_t - \frac{R_t}{P_t}. \tag{6.2.3}$$

Details about this procedure and results for Sydney over the period 2001 to 2009 can be read in Hill and Syed (2016).

For financial stability it is therefore also necessary to forecast the price-rent ratio for the nearest future. Thus, we have to express the ratio of the rents and the sales in our model. To derive the price-rent ratio, we write the linear predictor of the rents and the sales seperately, but both under the same joint model. Thus, the rents and the sales can be calculated as

$$\begin{aligned}\eta_i^{\texttt{rent}} &= \beta_0^{\texttt{rent}} + f_1^{\texttt{rent}}(\texttt{time}_i) + \beta^{\texttt{area}} \log(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) \\ &\quad + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i),\end{aligned} \tag{6.2.4}$$

$$\begin{aligned}\eta_i^{\texttt{sale}} &= \beta_0^{\texttt{sale}} + f_1^{\texttt{sale}}(\texttt{time}_i) + \beta^{\texttt{area}} \log(\texttt{area}_i) + f_{geo}(\texttt{long}_i, \texttt{lat}_i) + \sum_{j=2}^{6} \beta_j^{\texttt{bed}} \mathbb{1}_j(\texttt{bed}_i) \\ &\quad + \sum_{k=2}^{6} \beta_k^{\texttt{bath}} \mathbb{1}_k(\texttt{bath}_i),\end{aligned} \tag{6.2.5}$$

where $\eta_i^{\texttt{rent}} = \log(\mu_i^{\texttt{rent}})$ for the mean of the rent prices $\mu_i^{\texttt{rent}}$ and $\eta_i^{\texttt{sale}} = \log(\mu_i^{\texttt{sale}})$ for the mean of the sale prices $\mu_i^{\texttt{sale}}$. We are now interested in the behaviour of the mean rent and selling price of a specific house with j bedrooms and k bathrooms on an area of size `area` and at the location `long` and `lat`. The mean rent and selling price of this house is

$$\begin{aligned}\mu^{\texttt{rent}} &= \exp(\beta_0^{\texttt{rent}} + f_1^{\texttt{rent}}(\texttt{time}) + \beta^{\texttt{area}} \log(\texttt{area}) + f_{geo}(\texttt{long}, \texttt{lat}) + \beta_j^{\texttt{bed}} + \beta_k^{\texttt{bath}}), \\ \mu^{\texttt{sale}} &= \exp(\beta_0^{\texttt{sale}} + f_1^{\texttt{sale}}(\texttt{time}) + \beta^{\texttt{area}} \log(\texttt{area}) + f_{geo}(\texttt{long}, \texttt{lat}) + \beta_j^{\texttt{bed}} + \beta_k^{\texttt{bath}}).\end{aligned}$$

Thus the ratio of these two means is

$$\frac{\mu^{\texttt{sale}}}{\mu^{\texttt{rent}}} = \exp((\beta_0^{\texttt{sale}} - \beta_0^{\texttt{rent}}) + (f_1^{\texttt{sale}}(\texttt{time}) - f_1^{\texttt{rent}}(\texttt{time})). \tag{6.2.6}$$

The ratio of the mean rent and selling price of any specific house at any given location in our model therefore only depends on the time. However, this is not true in reality. As already mentioned before, Equation (6.2.2) assumes the rental and selling prices to be calculated for properties of equivalent quality. Hill and Syed (2016) therefore computed quality-adjusted price-rent ratios by ordering the rented and sold dwellings each year from the cheapest to the most expensive and computing the price-rent ratio for the lower quartile, median and the upper quartile sold houses and likewise for the lower quartile, median and upper quartile rented houses. They found out that the price-rent ratio increases from the lower to the upper end of the market. There are many different explanations for this behaviour. One reason is the fact that user cost may be lower at the high end of the market and thus the equilibrium price-rent ratio higher, as can be seen in Equation (6.2.2). These lower user cost are due to lower depreciation rate and lower risk premium at the high end of the market, but also the expected capital gain may be higher at the high end of the market. More details and explanations on this behaviour of the price-rent ratio can be read in Hill and Syed (2016).

Despite these contradictions to our assumption that the price-rent ratio only depends on the time, we will use Equation (6.2.6) to forecast the price-rent ratio for the nearest future. We have seen in Section 5.2.3 that the behaviour of the estimated parameters for the land area, the longitude and lattitude and also the number of bedrooms and bathrooms is nearly the same for the rented and sold houses. Therefore, we assume the price-rent ratio for reasons of simplification to only depend on the time.

The forecast of the price-rent ratio works in `R` like the one in the section before.

```
time <- seq(2008, 2016, by=0.05)
bed <- "1"
bath <- "1"
area <- 600
long <- 151
lat <- -33.8
rent.data <- expand.grid(TIME=time, BED=bed,BATH=bath,
        AREA=area, LONG=long, LAT=lat, TYPE="rent")
pred.rents <- predict(m_all_data, newdata=rent.data,
        type="response", se=TRUE)
sale.data <- expand.grid(TIME=time, BED=bed, BATH=bath,
        AREA=area, LONG=long, LAT=lat, TYPE="sale")
pred.sales <- predict(m_all_data, newdata=sale.data,
        type="response", se=TRUE)
plot(time, pred.sales$fit/(52*pred.rents$fit), type="l")
```

In this case, the defined characteristics of the specific house are not relevant to the price-rent ratio, since we have seen before that this ratio only depends on the time and not on any specific properties of the house. The trend of the price-rent ratio is plotted in Figure 6.2. While the house price at the beginning of the year 2014 was about 25 times of the corresponding yearly rent, the prediction shows us that at the beginning of the year 2016 one has to pay more than 32 yearly rents to cover the respective house price. From the beginning of the year 2012 onwards, there was an increase in the house prices compared to the house rents, since house prices increase faster than rents. Hence, over the years it has become more and more expensive to buy a house compared to renting a house.

Last but not least, we are interested in the standard errors of the index just derived to determine the accuracy of our prediction. To derive such standard errors we use the fact that every GAM has an underlying parametric representation, thus it is possible to obtain the so-called prediction matrix. When looking at model (6.2.4) for the weekly rents for example, the prediction matrix $\boldsymbol{X}^{\texttt{rent}}$ is the matrix which maps the estimated parameters $\hat{\boldsymbol{\beta}}^{\texttt{rent}}$ to the predictions of the linear predictor $\hat{\boldsymbol{\eta}}^{\texttt{rent}}$, which means that

$$\hat{\boldsymbol{\eta}}^{\texttt{rent}} = \boldsymbol{X}^{\texttt{rent}}\hat{\boldsymbol{\beta}}^{\texttt{rent}}.$$

The analogue relationship holds for model (6.2.5) for the sales, where the linear predictor $\hat{\boldsymbol{\eta}}^{\text{sale}}$ is obtained as

$$\hat{\boldsymbol{\eta}}^{\texttt{sale}} = \boldsymbol{X}^{\texttt{sale}}\hat{\boldsymbol{\beta}}^{\texttt{sale}}.$$

Since we are interested in the price-rent ratio, it follows under model (5.2.8) with joint parameter estimates vector $\hat{\boldsymbol{\beta}}$ that

$$\frac{\boldsymbol{\mu}^{\texttt{sale}}}{\boldsymbol{\mu}^{\texttt{rent}}} = \exp((\boldsymbol{X}^{\texttt{sale}} - \boldsymbol{X}^{\texttt{rent}})\hat{\boldsymbol{\beta}}),$$

where we assume the prediction matrices $\boldsymbol{X}^{\texttt{rent}}$ and $\boldsymbol{X}^{\texttt{sale}}$ to be the model matrices under the model over the time period from 2008 to 2016 for the rents and for the sales, respectively. According to Wood (2006), the prediction matrix is useful for the caluclation of variances for combinations of linear predictor values, like in our case the ratio of two linear predictors. To calculate the standard errors, we define the difference of the linear predictors as

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{X}^{\texttt{sale}} - \boldsymbol{X}^{\texttt{rent}})\hat{\boldsymbol{\beta}}.$$

We obtain the standard errors of these difference of linear predictors by calculating the square roots of the diagonal elements of

$$\widehat{\text{var}}(\hat{\boldsymbol{\delta}}) = (\boldsymbol{X}^{\texttt{sale}} - \boldsymbol{X}^{\texttt{rent}})\widehat{\text{var}}(\hat{\boldsymbol{\beta}})(\boldsymbol{X}^{\texttt{sale}} - \boldsymbol{X}^{\texttt{rent}})^{\mathsf{T}}.$$

Since we are interested in the variance of $\exp(\hat{\boldsymbol{\delta}})/52$ and not in the variance of $\hat{\boldsymbol{\delta}}$, we have to apply the delta method. This yields the linear approximation

$$\exp(\hat{\boldsymbol{\delta}})/52 = \exp(\boldsymbol{\delta})/52 + \frac{\partial \exp(\boldsymbol{\delta})/52}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

The derivative is calculated as

$$\frac{\partial \exp(\boldsymbol{\delta})/52}{\partial \boldsymbol{\beta}^{\mathsf{T}}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \frac{1}{52}\,\text{diag}(\exp(\boldsymbol{\delta}))(\boldsymbol{X}^{\texttt{sale}} - \boldsymbol{X}^{\texttt{rent}}).$$

Thus the variance of interest can be approximated by

$$\widehat{\text{var}}(\exp(\hat{\boldsymbol{\delta}})/52) = \frac{1}{52^2}\,\text{diag}(\exp(\hat{\boldsymbol{\delta}}))(\boldsymbol{X}^{\texttt{sale}} - \boldsymbol{X}^{\texttt{rent}})\widehat{\text{var}}(\hat{\boldsymbol{\beta}})(\boldsymbol{X}^{\texttt{sale}} - \boldsymbol{X}^{\texttt{rent}})^{\mathsf{T}}\,\text{diag}(\exp(\hat{\boldsymbol{\delta}})).$$

This result can now be used to plot the trend of the price-rent ratio with corresponding pointwise confidence intervals. The plot in Figure 6.2 is obtained with the following R-Code.

```
X.rent <- predict(m_all_data,newdata=r.data,type="lpmatrix")
X.sale <- predict(m_all_data,newdata=s.data,type="lpmatrix")
delta <- (X.sale-X.rent) %*% coef(m_all_data)
library(Matrix)
var.delta <- (X.sale-X.rent) %*% m_all_data$Vp %*%
        t(X.sale-X.rent)
diag.delta <- Diagonal(x=exp(delta))
sd.delta <- sqrt(diag(diag.delta %*% var.delta %*%
        diag.delta))/52
lines(time,exp(delta)/52+2*sd.delta,type="l", col="red")
lines(time,exp(delta)/52-2*sd.delta,type="l", col="red")
```

The function `predict()` in R returns the prediction matrices, if the `type` argument is set to `lpmatrix`. The estimated variance of the parameter vector $\hat{\boldsymbol{\beta}}$ under the model can be obtained with the call `m_all_data$Vp`. The rest of the input is a straightforward calculation of the above derived standard error. Finally, the plot in Figure 6.2 shows the prediction with corresponding 95% confidence intervals. The forecast for the year 2016 seems to be a good prediction, since the confidence interval is not that wide. To sum up, we have derived a model which shows a good fit to the data and also provides a good forecast for the nearest future. The prediction shows us that house prices in comparison to house rents will get more and more expensive in the future.

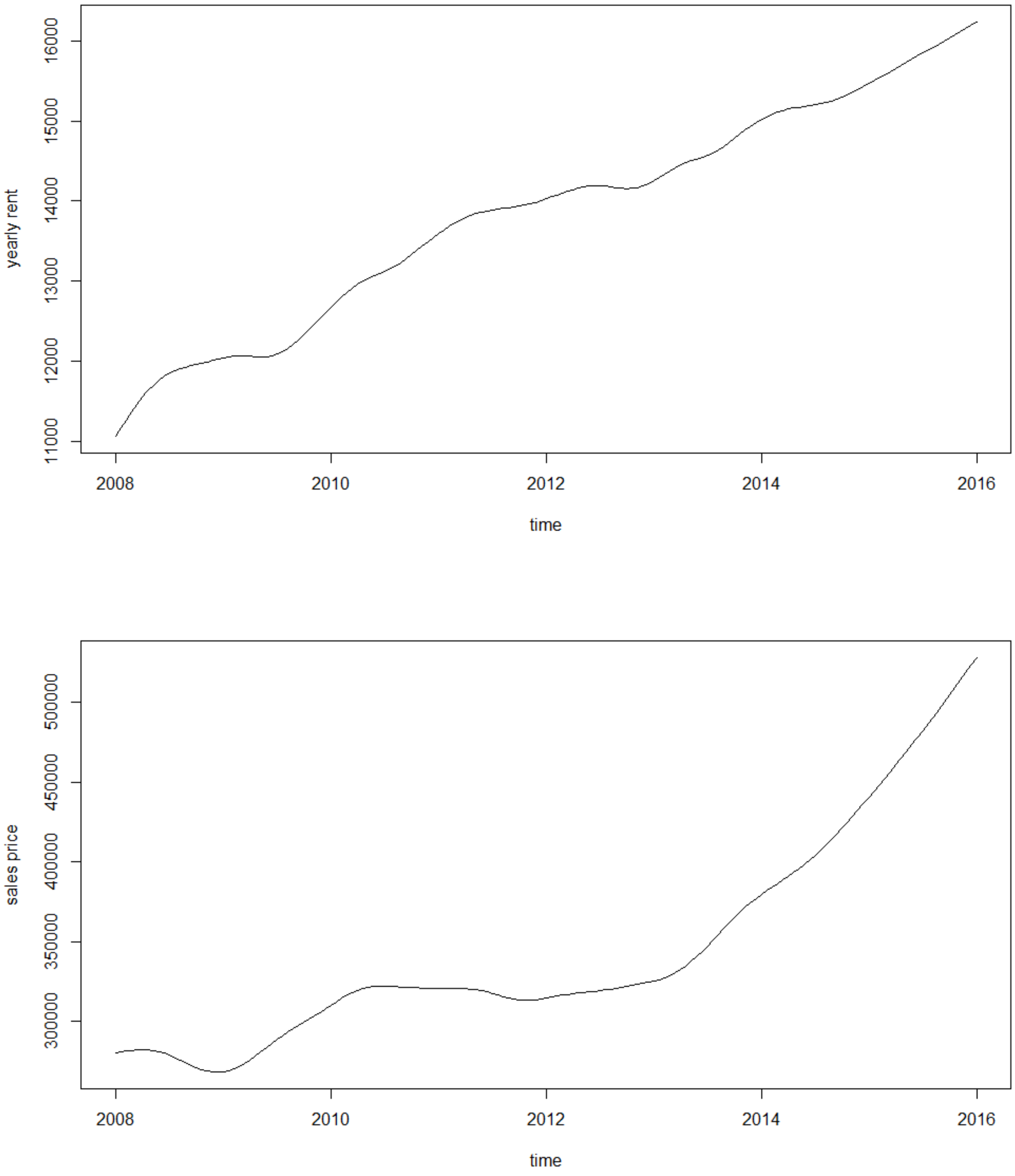


Figure 6.1: Prediction for the yearly rents and the house sales for the calendar year 2016

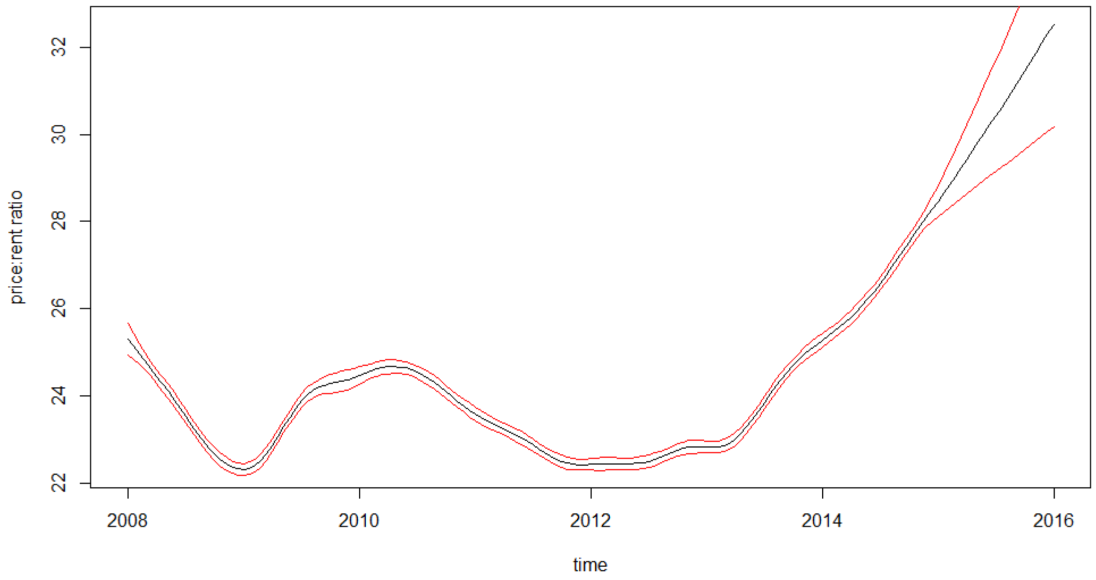


Figure 6.2: Prediction of the price-rent ratio for the calendar year 2016 with corresponding 95% confidence intervals

CHAPTER 7

Summary

This thesis provided insight into the large area of regression models. In the theoretical part of the thesis, the parametric and the nonparametric regression models were introduced and the different choices of basis functions used for the smoothing procedure presented. The theory of minimising the penalized least-squares criterion as well as the selection of the smoothing parameter was discussed in detail. Starting with the simplest univariate smoothing model, the bivariate and the multivariate smoothing model and finally the additive model were introduced. The additive model was further extended, so that it is possible to include geographical data and interactions between the covariates in the model. Last but not least, the generalized additive model, which allows to model a function of the response variable with an additive model and the response variable to follow the distribution of a member of the exponential family, was obtained. The chapter on nonparametric regression models finished by discussing the model choice and diagnostics in order to be able to check the adequacy of the chosen model.

The detailed theory on the nonparametric regression models was then used in the practical part. Two datasets with renting and selling prices of houses located in Sydney were used which also contained additional information about the houses' characteristics and the date of advertisement or the date of the sale, respectively. It started with a generalized additive model, assuming the house rent and the house price to follow a normal distribution. By trying different models and comparing their GCV and fREML scores it was found out that information about the date of advertisement or the date of the sale, respectively, the size of the land area, the location of the house given as longitude and latitude, as well as the number of bedrooms and bathrooms are relevant for explaining the renting and selling prices. In contrast, additional information about the region or the type of the house does not substantially improve the goodness-of-fit. While the date of advertisement or the date of the sale, respectively, and the size of the land area should each be included as a univariate smooth function, the location of the house given as longitude and latitude should be handled as a bivariate smooth function. Modeling the number of bedrooms and bathrooms as categorical factors showed a better fit to the data than including them as linear terms.

It was tried to fit the data using the identity link, but it was found that the log link is clearly a better choice. Furthermore, the assumption of gamma distributed renting and selling prices provides a better fit to the data than the assumption of a normal distribution. The most important interactions between the covariates were tried, namely the ones including the variable `time` and it showed that it is not necessary to consider interaction terms in the models. In conclusion, the univariate function depending on the size of the land area was exchanged by the logarithm function in order to simplify and stabilize the model.

The derivation of the price model was first done for the renting and selling prices seperately, yielding a model for the house rents and one for the house sales. These two models showed that the behaviour of the model terms is nearly the same for both models. The only exceptions are the term depending on the time and the intercept. This advised us to finally try a joint model for the house rents and the house prices. It showed that the assumption of one joint model for the rented and sold houses is hardly worse than if it is assumed that the renting and selling prices follow seperate models.

The derived model was further used in the last chapter to forecast the renting and selling prices as well as the price-rent ratio for the nearest future. The forecast for the renting and selling prices was just a linear extrapolation of the price curve, yielding that the trend of the house rents and house prices is a linear increase in the calendar year 2016. When looking at the ratio of the mean renting and selling prices under the estimated model, it showed that this ratio only depends on the time. Thus, we were able to forecast the price-rent ratio by extrapolating linearly. Finally, we derived pointwise 95% confidence intervals and concluded that our forecast seems to be a good prediction for the year 2016. The linear trend of the price-rent ratio tells us that house prices in comparison to the house rents will get more and more expensive. This factor more precisely ranges from about 25 in the year 2014 up to 32 in 2016, thus one has to pay 32 yearly rents in 2016 to cover the corresponding house price.

References

an de Meulen, P., Micheli, M., and Schmidt, T. (2011). Forecasting house prices in Germany. *Ruhr Economic Papers*, *294*.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, *17*, 453-510.

Case, K. E., and Shiller, R. J. (1990). Forecasting prices and excess returns in the housing market. *Real Estate Economics*, *18 (3)*, 253-273.

Craven, P., and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31 (4)*, 377-403.

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89-102.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Berlin Heidelberg: Springer.

Friedman, J., and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, *31 (1)*, 3-21.

Green, P. J., and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

Gu, C., and Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal on Scientific and Statistical Computing*, *12 (2)*, 383-398.

Gupta, R., and Kabundi, A. (2010). Forecasting real US house prices: principal components versus Bayesian regressions. *International Business and Economics Research Journal (IBER)*, *9 (7)*, 141-152.

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hatzvi, E., and Otto, G. (2008). Prices, rents and rational speculative bubbles in the Sydney housing market. *Economic Record*, *84 (267)*, 405-420.

Hill, R. J., and Scholz, M. (2014). Incorporating geospatial data in house price indexes: A hedonic imputation approach with splines. *Graz Economics Papers*, *5*.

Hill, R. J., and Syed, I. A. (2016). Hedonic price-rent ratios, user cost, and departures from equilibrium in the housing market. *Regional Science and Urban Economics*, *56*, 60-72.

Hutchinson, M. F., and de Hoog, F. R. (1985). Smoothing noisy data with spline functions. *Numerische Mathematik*, *47 (1)*, 99-106.

Kholodilin, K. A., and Siliverstovs, B. (2014). Business confidence and forecasting of housing prices and rents in large German cities. *DIW Berlin Discussion Paper, 1360.*

Kim, Y. J., and Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society B, 66 (2)*, 337-356.

Kneib, T., Hothorn, T., and Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics, 65 (2)*, 626-634.

Leamer, E. E. (2007). Housing is the business cycle. *National Bureau of Economic Research.*

Mallows, C. L. (1973). Some comments on C_p. *Technometrics, 15 (4)*, 661-675.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). Chapman and Hall, London.

Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A, 135*, 370-384.

Plakandaras, V., Gupta, R., Gogas, P., and Papadimitriou, T. (2015). Forecasting the US real house price index. *Economic Modelling, 45*, 259-267.

Rapach, D. E., and Strauss, J. K. (2007). Forecasting real housing price growth in the eighth district states. *Federal Reserve Bank of St. Louis. Regional Economic Development, 3 (2)*, 33-42.

Stiglitz, J. E. (1990). Symposium on bubbles. *Journal of Economic Perspectives, 4 (2)*, 13-18.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: Siam.

Waltl, S. R. (2016a). A hedonic house price index in continuous time. *International Journal of Housing Markets and Analysis, 9 (4)*, 648-670.

Waltl, S. R. (2016b). Variation across price segments and locations: A comprehensive quantile regression analysis of the Sydney housing market. *Real Estate Economics, doi:10.1111/1540-6229.12177.*

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society B, 62 (2)*, 413-428.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society B, 65 (1)*, 95-114.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association, 99*, 673-686.

Wood, S. N. (2006). *Generalized Additive models: An Introduction with R.* CRC/Chapman and Hall.

Wood, S. N. (2011). Fast stable REML and ML estimation of semiparametric GLMs. *Journal of the Royal Statistical Society B, 73 (1)*, 3-36.

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