Three studies on semi-mixed effects models

Dissertation

presented for the degree of Doctor of Philosophy

at the Faculty of Economic Sciences

of the Georg-August-Universität Göttingen

by Duygu Savaşcı

from Nazilli, Türkiye

Göttingen, 2011

1. Examiner: Prof. Dr. Stefan Sperlich
2. Examiner: Prof. Dr. Martin Schlather
3. Examiner: Prof. Dr. Inmaculada Martínez-Zarzoso

Disputation: 28.09.2011

1. Supervisor: Prof. Dr. Stefan Sperlich
2. Supervisor: Prof. Dr. María José Lombardía

To my family

Contents

List of Tables

List of Figures

Abbreviations and notations

The following list contains the most used abbreviations.

AIC	Akaike information criterion
ANC	Acid neutralizing capacity
EMAP	Environmental Monitoring and Assessement Program
EPA	Environmental Protection Agency
BLUE	Best linear unbiased estimator
BLUP	Best linear unbiased predictor
FEM	Fixed effects model
GAM	Generalized linear model
GAMM	Generalized linear mixed effects model
GCV	Generalized cross validation
GDP	Gross domestic product
GLM	Generalized linear model
GLS	Generalized least squares
GLMM	Generalized linear mixed effects model
HUC	Hydrologic unit codes
MEM	Mixed effects model
OCV	Ordinary cross validation
OLS	Ordinary least squares

P-splines	Penalized splines
PPML	Poisson pseudo maximum likelihood
REML	Restricted maximum likelihood
SMEM	Semi-mixed effects model
TPS	Thin plate spline

The following list contains the most used notations in Chapter 2 and 3.

$\boldsymbol{y}$	response vector
$\boldsymbol{X}, \boldsymbol{Z}$	design matrices
$\boldsymbol{\beta}$	vector of regression coefficients
$\boldsymbol{\epsilon}$	vector of errors
$\boldsymbol{u}$	vector of random effects
$g(.)$	link function
$(x)_+^p$	the function $x^p \boldsymbol{I}_{\{x>0\}}$
p	degree of the spline
λ	penalty term
n	number of responses
D	number of small areas
n_d	number of subjects in area d
q	dimension of $\boldsymbol{X}$
$m(.)$	smooth function
$\boldsymbol{F}, \boldsymbol{W}$	matrices of individual and area covariates
r_1, r_2	dimensions of $\boldsymbol{F}$ and $\boldsymbol{W}$
γ, η	additive smooth functions
$\boldsymbol{C}_i$	geographical coordinates for observation i

The following list contains the most used notations in Chapter 4.

T_{ijt}	export from country i to country j at time t
D_{ijt}	binary information
z_{ij}	non-binary time invariant information
v_{ijt}, η_{ij}	unexplained heterogeneity
u_{ij}	unobserved random effect
β, γ, δ	vectors of unknown coefficients
β_{yi}, β_{yj}	unknown scalar coefficients
w_{ij}	time-invariant variables
$\psi(.)$	additive function

Acknowledgements

I would like to convey my sincere thanks to my supervisor Prof. Dr. Stefan Sperlich of University of Geneva, it has been an honor to be his Ph.D. student. I also wish to express my appreciation to the other members of my committee, Prof. Dr. Martin Schlather and Prof. Dr. Inmaculada Martínez-Zarzoso of the University of Göttingen, for their collaboration during my studies. Special thanks are due to Prof. Dr. María José Lombardía of the University of A Coruña for assisting in my thesis.

Furthermore, I would like to thank Asst. Prof. Isabel Proença of the Technical University of Lisbon for her contribution to my studies. I wish to thank the members of the Institute for Statistics and Econometrics and the members of the Centre for Statistics of the University of Göttingen for their support they provided me throughout my studies. Thanks are especially due to Prof. Dr. Bernhard Brümmer and Prof. Dr. Walter Zucchini for the time and effort spent in Centre for Statistics Program. I also would like to thank Jean-Philippe Dupras for the proofread of my thesis.

Finally, I would sincerely like to thank my family. This dissertation would not have been possible without their love, understanding, continuous support, and dedication in all stages of my research.

1 Introduction

Today, applications of non- and semiparametric models are found in nearly all fields of empirical research. Since nonparametric methods do not have restrictive assumptions about the distribution of the observations or functional forms of the underlying data generating process, they are attractive methods when other necessary assumptions cannot be assured. However, nonparametric methods might be limited in practice due to other questions like the Bellman's curse of dimensionality or the true underlying degrees of freedom. In other words, certainly they cannot overcome the classical problem in statistics to find the optimal bias-variance trade-off. More specifically, for cases involving only a moderate sample size but many variables, we suffer from the curse of dimensionality. Then, by introducing partial parametric components, that may allow us to match structural conditions, such as linearity in some variables, the semiparametric modeling compromises between flexibility and simplicity in statistical procedures. More information about non- and semiparametric see, for example, Härdle, Müller, Sperlich and Werwatz (2004). One may consider, as the basis for many semiparametric models, the well known generalized linear model (Nelder and Wedderburn, 1972), given by $E(\boldsymbol{y}|\boldsymbol{X}) = g(\boldsymbol{X}^T\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is the parameter vector to be estimated and g is the link function. This model can be generalized in many ways like extending the index to be nonlinear or the link function to be nonparametric.

A different but also popular extension is still parametric but nowadays is often used as a bridge between parametric and non- or semiparametric models; adding some random effects in this generalized linear model leads us to a generalized linear mixed effects model (see Breslow and Clayton, 1993). Parametric mixed effects models are being widely used in many areas. One of the most common uses is in small area estimation, other examples are longitudinal studies, panel econometrics, multi-level models, repeated measurements in biometrics, etc. In other words, any statistical model that provides an intuitive clustering

which is modeled via random cluster effects imposing a new variance-covariance structure of the response data. Unfortunately, the application of these models is put into practice under the independence assumption between random effects and the covariates. We say unfortunately, because this clearly does not accurately reflect real world dependencies.

Like in other contexts, we will see that semiparametric modeling can be used to overcome this problem in an else purely parametric modeling context. In this dissertation, we aim to relax the independence assumption by introducing semi-mixed effects models. The inclusion of non- or semiparametric functions shall help to filter out possible dependence between random cluster (or level) effects and the covariates of interest. This helps us to establish the above mentioned but in practice else often unrealistic independence assumption. For the practical implementation, the semiparametric modeling is done by using splines. Note that, so far, this idea has only been introduced via kernels and was applied to a particular problem of a two level estimation problem.

We will first introduce the idea along with the model referring mainly to Lombardía and Sperlich (2011). Our main focus is on the extension to P-splines popularized by Eilers and Marx (1996), the different case-specific splines implementations including radial or spatial splines (see, for example, Green and Silverman, 1994), choice of penalizing term, automatic estimation of variances, and other practical questions. After a detailed introduction, and an intensive discussion, we include some simulation studies and implementation in the statistical software package R. Note that we extend the so far existing methods also toward additive modeling (see Deaton and Müllbauer, 1980), the inclusion of further nonparametric parts, nonlinear link functions, etc. After the discussion of implementation and general functioning, we continue with two real data problems. These have been chosen from the maybe most typical application areas in econometrics and economic statistics, i.e. empirical economics; small area estimation (see Rao, 2003 or Jiang and Lahiri, 2006) and panel data econometrics (see Baltagi, 2005 or Arellano, 2003).

Our first study, the one in the field of small area estimation, is an application that will address an environmental small area problem. Here, to overcome the dependencies between the random effects and the covariates, we include area-specific effects plus the information of location in the model. Therefore, we estimate the nonparametric functions in our semiparametric model by using P-splines and thin plate splines. Thin plate splines are isotropic smoothers and thus especially appropriate for spatial coordinates, i.e. handling

the effect of location (see, for example, Duchon, 1977 or Wood, 2003). P-splines are easy to implement and allow for additivity where we can model the explanatory variables non-parametrically (see, for example, Ruppert, Wand, and Carroll, 2003). This was necessary for the other covariates, except the location, to handle the double curse of dimensionality; the statistical one referring to the slow rate of convergence, and the practical one referring to the interpretation.

As previously mentioned, another common area where mixed effects are being frequently used is the econometric analysis of panel data. Panel data combines features of both cross section and time series data and have become widely used as a means to control for unobserved cross-section heterogeneity (see Mátyás, 1997, or Baltagi, Egger and Pfaffermayr, 2003). In our dissertation, we present an application with the gravity model to explain panel bilateral country trade flows. We apply our new semiparametric approach to panel gravity model via adding a nonparametric term in the transformed (via a known link function) conditional mean, which depends on observable proxy variables, in order to capture the dependency between the explanatory variables and the unobserved individual heterogeneity term. For this application, we use the generalized additive mixed effects model, which is an additive extension of generalized mixed effects model. This is again to avoid the curse of dimensionality. Note that, this panel data gravity model is an extension to the former application, as we have now complex link functions involved. Some other new aspects will be the question of model selection, in particular we refer to variable selection, and the selection of adequate software. Note that we are theoretically able to use commands provided in the statistical software package **R** and (or) the widely used econometric software package `Stata`.

Thus, the aim of this dissertation is to focus on semiparemetric estimation using mixed effects models in panel data and small area estimation where it is intended to relax the independence assumption. This independence assumption can presently be considered as a main challenge in the use of mixed effects models in practice after for two decades the main focus was directed toward the also quite crucial question of relaxing distributional assumptions. Recall that the (linear) mixed effects models are mainly used by maximum likelihood based estimation procedures. We first translated the idea of Lombardía and Sperlich (2011) from kernel smoothing to splines. We are aware of the fact that this is at the cost of losing the nice *slider* interpretation they have as now we have no parameter that in its two extremes includes fixed effects models and random effects models (without

further level effect modeling). But, we gain in different issues of practicability; simpler implementation up to the use of already implemented software (in **R**, `Stata` and SAS at least),

1. simpler automatic choice of the slider (see Lombardía and Sperlich, 2011) as a fraction of variances of error variance and the splines' pseudo-variance,
2. faster calculation since splines are computationally much less expensive than kernels,
3. the possibility of local smoothing via heteroskedasticity of the splines' pseudo-variance,
4. simple extension to additive modeling of the semiparametric filter of dependence,
5. straight forward extension to semiparametric modeling of the other covariates' impacts,
6. a well studied extension to the generalized linear mixed effects models (which for kernels has only been studied in theory to our knowledge),
7. etc.

We performed and studied different implementations where

1. we allowed for partly choosing and partly fixing the smoothness of the nonparametric parts (especially the *slider*, see Lombardía and Sperlich, 2011),
2. we tried different alternative estimation methods for the variance of the random effects and the pseudo-variances of the splines,
3. we analyzed possible extensions to allow for heteroskedasticity of the residuals (for the random effects this has been studied already in detail, see, for example, Foulley and Quaas, 1995 or Robert-Granié, Heude and Foulley, 2002),
4. compared semi-mixed effects model (where only the filter is nonparametric), additive semi-mixed, and semiparametric mixed effects models (i.e. additive partial linear models plus semiparametric filter),
5. etc.

We also compared

1. own different implementations in **R** and Fortran 90,

2. the different tools offered by the mgcv package in **R**, see Wood (2006),
3. the competing recent implementations in `Stata` 11 (`Stata` 10 partly did not offer alternatives),
4. applications to different real data sets,
5. etc.

Nevertheless, in this thesis we only present a very small selection of results for the following reasons;

1. in parallel to our thesis, this topic has been studied world-wide intensively such that most of our findings have already been studied and published somewhere,
2. unfortunately but maybe not that surprisingly, many of our implementations turned out to not work well in practice,
3. where we compared different methods, software or commands, we present here only the outcome of the most reliable ones,
4. and for the sake of brevity.

In the next sections, the basic information for the applications will be introduced. This will then be followed firstly by our small area estimation problem for environmetrics, and finally by the panel data analysis application. All sections are mainly independent from each other, so can be read individually. However, the information given in the following sections will be the main guidance for all following sections. To make the sections somehow independent, some repetitions are unavoidable.

2 Mixed effects models

Certainly, like many other types of statistical regression models, a mixed effects model (MEM) describes simply a relationship between a response variable and the covariates that have been measured or observed along with the response. What distinguishes the mixed effects model from the others is that we have a natural clustering given in the data, may it be due to repeated measurements for (almost) each subject or individual, a regional, climatic, social, ... area. Actually, this can even lead to a nested clustering; we then speak of multilevel models (with more than just two levels). The idea is then to allow for random deviations from the general mean but driven by the affiliation to a specific cluster. These deviations are marked in either just the intercept which may *randomly* vary over the different clusters or even in deviations from the general slopes (coefficients) in a (generalized) linear mixed effects model. To understand the basics of the mixed models, we start by examining linear models and the extended case of mixed effects model.

A statistical model is explained as a mathematical relationship between the explanatory variables and the response variable. The response variable y is the one whose content is modeled with other variables, namely the explanatory variables $x_1, x_2, ..., x_n$. One can start with modeling the variables using a simple *linear model*

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i \tag{2.1}$$

for $i = 1, ..., n$, unknown coefficients β_j, $j = 0, 1, ..., k$, and a random deviation (the error term) e_i. We can always rewrite the model in matrix notation, i.e.

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\boldsymbol{y}$ is the response vector of length n, $\boldsymbol{X}$ is the vector of covariates of dimension $(1 + k) \times n$ with a first column of ones referring to the constant term β_0. Then, $\boldsymbol{\beta}$ is the

This chapter is a joint work with Sperlich, S. and Lombardía, M. J..

vector of regression coefficients of dimension $(1+k)$, and $\boldsymbol{\epsilon}$ is the vector of errors of length n. Often, not necessarily but typically for the extensions to random effects models, it is assumed that the distribution of the error term is known, for example $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_n)$. For details see the classic book of Searle (1971).

Linear models, which are fully determined up to a parameter, have been used in econometrics and statistics for decades, since they were easy to implement and can easy to interpret. With parametric modeling, the estimation procedure is easy as long as the underlying assumptions are accurate. However, the estimates can be inconsistent and give misleading information if the assumptions are violated, which leads to several extensions, nonparametric regression being the most flexible one among all, see for example Härdle, Müller, Sperlich, and Werwatz (2004).

Generalized linear models allow for a different extension; now the response variable can also follow a discrete distribution, which is also possible for the linear model but definitely not reasonable for different obvious reasons. The link functions are typically assumed to belong to the exponential family such as Binomial, Poisson, etc. and they allow for non-linear structures in the model:

$$g(\mu_i) = \boldsymbol{X}_i \boldsymbol{\beta}, \tag{2.3}$$

where $\mu_i \equiv E(y_i)$, and g is a the (typically assumed to be known) link function. See McCullagh and Nelder (1989) for the classical generalized linear model and its implementation, as well as some basics about the exponential family. Then, in Fahrmeir and Tutz (2001) and McCulloch and Searle (2001) we get already introduced to the extension and transition to linear and generalized linear mixed effects models which we will consider next.

2.0.1 Linear mixed effects model

Often, the data to be analysed is clustered, grouped, or otherwise hierarchically organized. Mixed effects models include additional random effects for the particular clusters. These models have turned out to be much more appropriate for representing these types of data. The fact that we add random and not fixed effects is a statistical trick but does not necessarily follow a deeper interpretation idea. It does not mean that the cluster

effect is random but this cluster effect is not further modeled, for example due to the lack of sufficient information. Note that this randomness has consequences for the conditional distribution of the response y; in particular, observations belonging to the same cluster are dependent if one does not condition on the (unobserved) cluster effect. This allows to make use of the idea of generalized or weighted least square estimation, or the inclusion of an additional variance modeling in maximum likelihood procedures. This yields more efficient estimates but at the price of possible misspecification as it assumes independence between the unobserved cluster effect and the observed information gathered in the covariates vector X.

To summarize, for a vector of response $\mathbf{y}$ with known $\mathbf{X}$ (including the intercept) and $\mathbf{Z}$ (typically a subvector of X, often just the intercept) design matrices, a *linear mixed effects model* is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{e}, \tag{2.4}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices, $\boldsymbol{\beta}$ is the fixed effects vector, and the random effect vector $\mathbf{u}$ is independent from $\mathbf{X}$ and $\boldsymbol{e}$. For the sake of presentation, let us assume for a moment that $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$ and $\boldsymbol{e} \sim N(\mathbf{0}, \mathbf{R})$. The covariance matrices $\boldsymbol{D}$ and $\boldsymbol{R}$ may depend on a set of unknown variance components. Here are some properties: $\mathbf{u} \sim N(\mathbf{0}, \mathbf{D})$ means that the expected value of the random effects are 0, $E(\boldsymbol{u}) = 0$ and the variance is $\boldsymbol{D}$, $Var(\boldsymbol{u}) = \boldsymbol{D}$. Variance of $\boldsymbol{y}$, is specified as $Var(\boldsymbol{y}|\boldsymbol{u}) = \boldsymbol{R}$ since $Var(\boldsymbol{u}) = \boldsymbol{D}$. The model in (2.4) can also be written as;

$$E(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{Z}, u) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \tag{2.5}$$

where $\boldsymbol{y}$ follows a distribution with mean $\sim \boldsymbol{X}\boldsymbol{\beta}$, and since $E(\boldsymbol{u}) = 0$, the variance of $\boldsymbol{y}$ is

$$\boldsymbol{V} = Var(\boldsymbol{y}) = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{'} + \boldsymbol{R}. \tag{2.6}$$

For known variance components, the *best linear unbiased estimator* (BLUE) of $\boldsymbol{\beta}$ is then

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{'}\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{V}^{-1}\boldsymbol{y} \tag{2.7}$$

and the *best linear unbiased predictor* (BLUP) of $\boldsymbol{u}$ is

$$\hat{\boldsymbol{u}} = \boldsymbol{D}\boldsymbol{Z}^{'}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{2.8}$$

For details, see for example, Henderson, Kempthorne, Searle and von Krosigk (1959), Robinson (1991) or McCulloch and Searle (2001).

Clearly, if the variances are known - something often assumed in earlier publications - then it is obvious how a generalized or weighted least squares estimator can be applied to obtain consistent and even efficient estimates for $\boldsymbol{\beta}$. The first difficulty occurs when these variance components have to be estimated. Usually, estimates for the variance components depend on $\boldsymbol{\beta}$, but $\boldsymbol{\beta}$ depends on these variance components, such that an iterative estimation procedure seems to be necessary. In case of normality and homoscedasticity for both the error and the random effect, there exists a linear transformation so that it is possible to estimate the variance components without knowing $\boldsymbol{\beta}$. This procedure used is known as *restricted maximum likelihood* (REML) which maximizes the likelihood of linear combinations of elements of $\boldsymbol{y}$, see Patterson and Thompson (1971).

2.0.2 Generalized linear mixed effects models and other extensions

As in a linear mixed effects model, a *generalized linear mixed effects model* (GLMM) includes a vector of observations $\mathbf{y}$, design matrices $\mathbf{X}$ and $\mathbf{Z}$, fixed effects, $\boldsymbol{\beta}$ and random effects $\mathbf{u} \sim (\mathbf{0}, \mathbf{D} = \sigma_u^2 \boldsymbol{I})$ with typically known distribution. See, for example, Breslow and Clayton (1993), McCulloch and Searle (2001). The extension is done by introducing an inverse link function $g(\boldsymbol{.})$. The fixed and random effects are combined to form a *linear predictor*

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} = \boldsymbol{\eta} \tag{2.9}$$

where $\boldsymbol{\mu}$ is the vector of the conditional mean of $\boldsymbol{y}$ given both, the observed covariates $\boldsymbol{X}$, and the (unobserved) random effects and is linked to the parameter by the mentioned function $g(\boldsymbol{.})$. We introduced here $\boldsymbol{\eta}$ just for simplification; it often refers to the *linear predictor*, cf. Table 2.1 gives a short list of some of the common link functions for various distributions of response $\boldsymbol{y}$, see also Härdle, Müller, Sperlich, and Werwatz (2004) for further details.

Not surprisingly, the selection of the (inverse) link function is based on the distributions of the error term and the random effect(s). For example, for the linear mixed model, the inverse link function is the identity function $g(\boldsymbol{\eta}) = \boldsymbol{\eta}$.

Distribution	Link	Inverse Link
Normal	Identity	η
Binomial	Logit	$e^\eta/(1+e^\eta)$
Poisson	Log	e^η
Gamma	Inverse	$1/\eta$

Table 2.1: Common link functions for various distributions.

Hastie and Tibshirani (1990) introduced many semiparametric generalizations of the GLM in their seminal book, but gave little emphasis to the inclusion of random effects. This was quite different in Fahrmeir and Tutz (2001) or McCulloch and Searle (2001). There exists indeed a vast non- and semiparametric literature on additive and generalized additive models and it is not our aim to repeat this here; see for example Sperlich (1998).

Our focus is directed toward the extensions which include explicitly random effects: A *generalized additive mixed model* (GAMM) is a special form of a GLMM where the originally linear predictor is now specified in terms of a smooth function or functions of the covariates (see, for example, Lin and Zhang, 1999 or Wood, 2006). More specifically, a GAMM has the form

$$g(\mu_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \sum_j f_j(x_{ji}) + \boldsymbol{Z}\boldsymbol{u} \tag{2.10}$$

where $u \sim (0, \boldsymbol{D})$ with some known distribution, and $y_i \sim exponential\ family(\mu_i, \boldsymbol{R})$. Lin and Zhang (1999) proposed an approximate inference in GAMMs using smoothing splines and marginal quasi-likelihood. The advantage compared to the GLMM is just the allowance for more flexibility of the functional form of the (originally) linear predictor. The idea is to relax the assumption to know in advance how the covariates $\boldsymbol{X}$ enter the regression model. Without completely depreciating this idea, our interest will be rather to explore the possibility of arbitrary flexibility in order to filter potential dependencies between the random effects and the covariates, which otherwise will render the estimation inconsistent.

Before doing so in the next subsection, let us briefly mention the main advances that have been achieved in recent years. The important point here is that the contributions in this area are not just papers and books but in most cases accomplished by software

packages. This explains why these models are quite popular in practice in biometrics and environmetrics, today. **R** library package MASS provides *glmmPQL* function which fits GLMMs and works by repeated calls to lme that is provided in package nlme. The mgcv package provides *gamm* function which fits GAMMs by a call to lme in the normal errors identity link case, otherwise by a call to *gammPQL*, which is a modification of glmmPQL. Further details can be found in Package mgcv documentation (2011) or in Wood (2006). In this dissertation, the mgcv package is used for small area estimation problem for environmetrics. For the panel data study, we studied in both **R** and `Stata`. `Stata` 11 provides the command *xtmepoisson* that fits mixed effexts models for count responses. For details, see `Stata` 11 documentation.

2.0.3 The semi-mixed effects model

Recently, Lombardía and Sperlich (2011) proposed a new model that allows to change from MEM, without area specific covariates, to a *semi-mixed effects model* (SMEM) with a smooth area specific mean and a random effect, up to a *fixed effects model* (FEM).

Mixed effects models allow for efficient estimation of the fixed parts in the model. Also, they treat the small area effects as random effects, and make use of the random effects for prediction. In the moment of prediction, one adds the predicted random effect to the total prediction. The additional variance of the prediction caused by assuming this effect to be random is only slightly larger than the variance of a fixed effect estimate based on small samples, but the modeling of the new variance structure allows for a more efficient estimation of the coefficients. It might improve prediction in the mean, but under the assumption of independence between random effects and the covariates. Clearly, the independence assumption is not shared by the fixed effects models given by

$$Y_{id} = X_{id}\boldsymbol{\beta} + u_d + \epsilon_{id}, \tag{2.11}$$

where $d = 1, ..., D$ and $i = 1, ..., n_d$ with u_d being the area specific fixed effect without independence assumption, meaning without being independent from the individual effects, X_{id}. FEM provides an unbiased estimate of $\boldsymbol{\beta}$ depending on the method but it also contains $D + 1$ parameters (D intercept and a slope) which leads to a large covariance for all estimators.

With the smooth transition from MEM to SMEM and FEM, one can model the area effect

and relax the independence assumption. The SMEM is, then, defined as

$$Y_{id} = X_{id}\boldsymbol{\beta} + \eta_v(\boldsymbol{W}_d) + u_d + \epsilon_{id}, \tag{2.12}$$

where $\eta_v : I\!R^q \to I\!R$ is a nonparametric function with a given slider v. If one sets $v = 0$, then SMEM model turns into a FEM but if $v = \infty$ is set, one obtains the MEM with u_d being a random effect. The estimation procedure of a SMEM is as calculating a partial linear mixed effects model (see e.g. Lombardía and Sperlich, 2008 or Opsomer, Claeskens, Ranalli, Kauermann, and Breidt, 2008). We later provide a section (Section 2.2.3) where we combine the semi-mixed effects modeling idea with the spline implementation.

2.1 Typical applications of mixed effects models

Mixed effects models are widely used in many fields of empirical research. As indicated above, they are especially appropriate if we face data with intuitively clustered data. This is typically the case for small area statistics where *area* may refer to geographical, administrative, political, climatic, topographic, etc. areas, see, for example, Rao (2003). Other examples are; repeated measurements (see e.g. Davidian and Giltinan, 1995) and longitudinal data (see e.g Verbeke and Molenberghs, 2009), as we have them most frequently in biometrics (medicine). Not to forget the hierarchical models in social science, including economics, and finally panel data econometrics (see e.g. Baltagi, 2005). In econometrics, they are also quite commonly used for data mapping (see e.g. Davis, 2003) and data matching (see e.g. Elbers, Lanjouw and Lanjouw, 2003).

For our applications, we focus on small area statistics in an environmental context, and panel data analysis for an econometric modeling problem. Before we speak of the implementation of mixed effects models with splines, the approach we have chosen for given the practical advantages, let us briefly review these two fields.

2.1.1 Small area estimation and environmetrics

Small area is the term that is used to refer, generally, to a small geographical area, though it may also refer to an isolated particular demographic. If a survey has been carried out for a whole population, a problem arises when trying to generate accurate estimates relative

to any particular small area within this population, because this area may be too small. While design-based inference methods may be appropriate for the overall survey sample size, one has to rely on alternative methods, namely model-based, for small domains where population level auxiliary information is available. In these circumstances, the statistical techniques involving the estimation of parameters are simply called *small area estimation*.

Models based on random area-specific effects that account for area variations are called *small area models* so that the indirect estimators based on small area models are consequently called the *model-based estimators*. Small area models can be classified in two types: aggregate level (or area level) models and unit level models. Fay and Herriot (1979) were the first to use an area level model for estimating per capita income for small areas in U.S.A. and proposed an empirical Bayesian method. Unit level models are relevant for continuous $\boldsymbol{y}$ response variables and these models may be regarded as special cases of linear mixed effects models. In the case of binary response, the logistic mixed effects model is used and, in the case of count response, the loglinear mixed effect model is used where both models are the specific cases of generalized linear mixed effects models. Battese, Harter and Fuller (1988) used the unit level model to estimate county crop areas using survey and satellite data and constructed an empirical best linear unbiased predictor for the small area means. For further details about small area models, see, for example, Small Area Estimation by Rao (2003).

During the last few decades, mixed effects models have been widely used in small area statistics. See, for example, Jiang and Lahiri (2006), Opsomer, Claeskens, Ranalli, Kauermann, and Breidt (2008), Lombardía and Sperlich (2011). For combining information from various sources and explaining different sources of errors, these models offer great flexibility and are well suited to solving many problems in small area estimation. The most frequent argument is that direct estimates use too little information; then, imposing a common model that deviates say *randomly* from one area to another is a way to borrow information from all the other areas. Note that this argument is particularly valid if one is just interested in a particular area-specific information (a macro-parameter) like the area mean. However, for a consistent model-parameter estimation and prediction, the independence assumption turns out to be crucial. In practice, this independence is often questionable and renders not just the point prediction but moreover the inference and interval prediction invalid. Another crucial point to be mentioned is the common use of strong distributional assumptions. These have been the focus of lively discussion and re-

search for more flexible methods, especially in biometrics. The former problem, however, i.e. the independence assumption, is still an untouched nimbus in small area statistics, maybe because of the unknown consequences of what would happen if it fails to hold. See, for example, Jiang and Lahiri (2006) for further details on mixed model estimation in small area context.

We will consider a problem of environmental small area estimation where we try to relax the independence assumption between random effects and the covariates. To overcome the dependencies between the random effects and the covariates, we include area-specific effects semiparametrically in the model. We estimate the nonparametric functions in our model by using P-splines and thin plate splines. As indicated, and as will be discussed more in detail, the thin plate splines shall help us to incorporate the geographical location but is thus especially vulnerable to the independence assumption. Here, a filter is unavoidable to make the outcome interpretable. We will be using, as a case study, a survey of lake water quality in North-eastern states of U.S.A. conducted by the Environmental Monitoring and Assessement Program (EMAP) of the Environmental Protection Agency (EPA) (Opsomer, Claeskens, Ranalli, Kauermann, and Breidt, 2008).

2.1.2 Panel data analysis

A common use of mixed effects models is in the analysis of *panel data* (or *longitudinal data*) (see e.g. Diggle, Heagerty, Liang and Zeger, 2002). A panel contains observations for each subject over multiple time periods. The common feature of panel data sets are that the sample of individuals is typically relatively large while the number of time periods is generally small. The main advantage mentioned in economics or econometrics is the chance to overcome the problem of unobserved heterogeneity leading to endogeneity of covariates and thus to inconsistent estimates. Other advantages are the possibility of estimating dynamic effects, the increase of efficiency, etc.

So, it has the potential to solve problems neither cross section methods nor pure time series methods can solve (see e.g. Hsiao, 2003). The reasons for favoring a panel data approach can be that panel data source grants the ability to control for individual fixed effects and to model temporal effects without aggregation bias. Therefore, panel data estimation methods have become increasingly popular in both theoretical and applied micro- as well

as macro- economics; but this is also as a consequence of the increased available data of this type.

There are several panel regression models; some include individual or subject specific effects, others time specific effects, sometimes both, etc.. A main distinction is to separate them into the *fixed effects panel data models*, where the model includes an individual effect that is constant over time, and the *random effects panel data models*, which basically coincide with our mixed effects models having a random effect for the individuals (which does not change over time). So the individual effects are considered as random rather than fixed constants. The simplest approach to the estimation is the *pooled ordinary* or *weighted least squares* estimation. For the model we are interested in, i.e. the random effects panel model, the notation for the errors might be set to $u_{it} = \alpha_i + \epsilon_{it}$ where α_i are the individual effects. The errors of the same cross-section unit are then correlated and the *generalized least squares* is thus used to estimate the model. For detailed descriptions of these estimation methods, see, for example, Green (2003) or Wooldridge (2002).

The gravity model of trade has been widely used in economics due to its ability to explain trade flows among countries. Tinbergen (1962) was the first to use the gravity model in that context. The gravity model has generally been estimated using cross-sectional data. However, this might generate inefficient results since heterogeneity among countries cannot be controlled for in an adequate manner. To address this problem, the gravity model is now being estimated using panel data, which have the advantage that they allow for more general types of heterogeneity (see e.g. Westerland and Wilhelmsson, 2009). The common procedure to estimate gravity equations with panel data is based on the ordinary (or weighted if we model possible time dependence) least squares estimation of the transformed log-linear specification including fixed effects to control for country unobserved heterogeneity. This may lead to a lack of efficiency due to the great number of parameters to be estimated, but the problem which make it less attractive is the difficulties in estimating the effects of time-invariant variables. Unfortunately, if we use random effects we again face the problem of the independence assumption which has always been in the center of controversy discussions in econometrics.

In this dissertation, we introduce a nonparametric component in the gravity panel equation that captures country unobserved heterogeneity dependent on the explanatory variables without compromising the estimate of the effect of time invariant variables. Additionally,

to address the criticism of Santos Silva and Tenreyro (2006) we transform the panel gravity model to an additive mixed effects model with a Poisson link function. Then, we have to estimate a generalized additive semi-mixed effects model. There, the introduction of gamma distributed (by assumption) random effects extends the conditional distribution of the response to a negative binomial one which is thus much more flexible than the Poisson one. For details see the corresponding Sections in our thesis. We will analyze the trade flows among the EU25 countries from 2004 till 2007.

2.2 Non- and semiparametric regression

Many books have been written about the topic of non- and semiparametric estimation, even when just concentrating on some regression problems, see, for example, Ruppert, Wand, and Carroll, (2003), Härdle, Müller, Sperlich and Werwatz (2004) or Horowitz (2009).

It is by no means our aim to summarize even only parts of it. Instead, we directly focus on the ideas we will apply in the following for our modeling and filtering. Let us start very simply and consider the two regression examples below;

Example 1. $E(y|x) = \beta_0 + \beta_1 x$, a parametric model, and

Example 2. $E(y|x) = m(x)$, a nonparametric model with $m(\cdot)$ smooth but not further specified.

Parametric models are determined up to a finite number of parameters. If the underlying assumptions are correct, the estimations and predictions of these models are done easily unless the assumptions are violated. Let $m(x)$ be a smooth function that is unknown but to be estimated. The objective is to estimate m by means of a function that both fits the data well and is sufficiently smooth. In other words, nonparametric models avoid the restrictive assumptions on the functional form (example 1). In order to meet this, nonparametric regression estimators have to be rather flexible. However, some problems become uncontrollable as the number of the variables increase. In the literature, this fact is know as the curse of dimensionality (Bellman, 1957). Other problems are the lack of interpretability, the choice of smoothness, etc. A main criticism is the lack of modeling. To address the curse of dimensionality, interpretability, and the modeling idea, an accepted

compromise is found by the semiparametric modeling.

The methods which try to overcome the dimensionality problem by combining some of the parametric and the nonparametric techniques are known as the *semiparametric methods*. The basis for many semiparametric models is the generalized linear model (Nelder and Wedderburn, 1972) which is given by $E(\boldsymbol{y}|\boldsymbol{X}) = g(\boldsymbol{X}^T\boldsymbol{\beta})$ where $\boldsymbol{\beta}$ is the parameter vector to be estimated and g is the link function, see our discussions above. This model can be generalized in many ways. If we consider an unknown smooth link function, the model then leads to a single index model (see e.g. Ichimura, 1993). If we assume a nonparametric additive argument of g, this leads to a generalized additive model (Hastie and Tibshirani, 1990). If we assume a combination of additive linear and nonparametric components in the g argument, then this model leads us to a generalized partial linear model. See, for example, Severini and Staniswalis (1994), Lin and Carroll (2001). If there is no link function, we get an additive model (see Friedman and Stuetzle, 1981). If we also add random effects to these models we get mixed effects models, etc. We discussed some of the extensions of these models earlier.

These models are particularly popular in econometrics; see Yatchew (2003), Härdle, Müller, Sperlich, and Werwatz (2004), Sperlich, Härdle, and Aydinli (2006) or Horowitz (2009) to mention only few of the many books on this field.

The main challenge is, typically, to estimate the parametric part of the model at the parametric convergence rate, namely $O(\sqrt{n})$. A second challenge is, then, the implementation. While for splines it seems to be much harder to elaborate and derivate exact mathematical theory, and therefore it is often done for series estimators or kernel methods, the implementation seems to be most attractive via splines. Among them, P-splines have nowadays attracted most of the attention for various reasons. This is why we now concentrate on them, starting with the following section that tries to describe in detail some of the main ideas and implementation.

2.2.1 Semiparametric modeling with penalized splines

Mixed effects models have been famous recently in semiparametric statistics. Bayesian approaches, feasible algorithms using spline methods and kernel based smoothing methods are all used in mixed effects models by several researchers. For Bayesian semiparametric

approaches see, for example, Fahrmeir and Lang (2001) or Kneib and Fahrmeir (2007). For approaches using splines see Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008) or Claeskens, Krivobokova and Opsomer (2009). For kernel based approaches see, for example, Lombardía and Sperlich (2008).

We are interested in using the penalized splines in our studies since they are typically implemented in different software packages like **R** and `Stata`. Moreover, they allow for additive modeling, automatic choice of smoothing parameters and further extensions. Certainly, we also can model some of the explanatory variables nonparametrically if wanted. In the Section 2.2.2, we will discuss the basics of penalized splines approaches and their implementations. The following section is especially aimed at showing how to combine nonparametric regression and mixed models with penalized splines. The only danger of misinterpretation is that one might be tempted to mix up the interpretation of the 'real' random effects and the pseudo random effects which are only generated to calculate the nonparametric spline estimates. Although we will abstract from our original model at the moment of implementation, we should always be aware of the differences between our deterministic parts of the model, the filtering part (which is also deterministic), and the real random part. Recall that the latter one is separated into residuals and a random (level or) area effect.

2.2.2 Penalized spline regression

Penalized spline regression, often referred to as P-splines, are popularized by Eilers and Marx (1996). P-splines are an attractive smoothing method because of their flexibility and are also a natural candidate for constructing nonparametric small area estimation. An advantage of the spline based approach is that it allows easily for additivity and is easy to calculate. Finally, even the extension to generalized models, i.e. including additional nonlinear though known link functions seems to be relatively manageable compared to competing methods.

Although Hastie and Tibshirani (1986) pushed splines forward a lot, they sold it mainly under the name of generalized additive modeling (GAMs) and backfitting methods. It should be said that in particular contributions of Wahba (1990) and Gu (2002), among others, heavily influenced the advances in that field. When looking at splines which played

a key role for the practical advances, recall that Duchon (1977) invented thin plate splines which we will use later on, in our first application. Penalized regression splines go at least partly back to Wahba (1980), but were given real impetus by Eilers and Marx (1996) and in a GAM context by Marx and Eilers (1998), always with a special emphasis on implementation and its practical use. (In fact, speaking about mathematical statistics we would have to mention a rather different literature, contributors and authors). Wood (2006) comprised in his book, *Generalized Additive Models: An Introduction with R*, the main results, and gave it a real push ahead with the mgcv package implementation in R.

The main target of our presentation is to show how the implementation of P-splines and (generalized) linear mixed effects models are related to each other - what basically has made up their popularity. Consider now the relatively simple model written in matrix notation,

$$\boldsymbol{Y} = \boldsymbol{\eta}(\boldsymbol{F}) + \boldsymbol{\epsilon}, \tag{2.13}$$

where $\boldsymbol{\epsilon}$ is a vector of independent random variables with mean zero. Let us assume them, for a moment, to be normally distributed with mean zero and variance σ_e^2. Further, $\eta(\boldsymbol{F})$ is an unknown (for the sake of interpretation and to avoid the curse of dimensionality) additive function such that $\eta(\boldsymbol{F}) = \sum_{j=1}^{J} \eta_j(\boldsymbol{F}_j)$. The latter one will certainly be estimated using a P-spline. In order to do so, note that the model can be approximated adequately well by

$$\begin{gathered} \tilde{\eta}(\boldsymbol{F}) = F_1\eta_1 + F_2\eta_2, \;\; \eta_1 \in I\!R^{pJ}, \;\; \eta_2 \in I\!R^{JK}, \\ F_1 = \left(\boldsymbol{F} \vdots \boldsymbol{F}^2 \vdots \ldots \vdots \boldsymbol{F}^p\right), \\ F_2 = \left((\boldsymbol{F} - \boldsymbol{\tau}_1)_+^p \vdots (\boldsymbol{F} - \boldsymbol{\tau}_2)_+^p \vdots \ldots \vdots (\boldsymbol{F} - \boldsymbol{\tau}_K)_+^p\right), \end{gathered} \tag{2.14}$$

where p is the degree of spline, $(x)_+^p$ denotes the function $x^p \boldsymbol{I}_{\{x>0\}}$ and $\tau_1 < ... < \tau_K$ is a set of previously fixed knots. In practice, one can take each tenth ordered observation of the particular covariate.

In P-spline regression, K is typically taken to be large, e.g. with 1 knot every 4 or 5 observations (Opsomer, Claeskens, Ranalli, Kauermann, and Breidt, 2008). Higher values of p, the power of spline, may lead to smoother spline functions. For what is considered to be a reasonable number of knots, the degree of the spline basis usually has little influence on the fitted spline at the knot points, although interpolation between the knots will take

the form of the underlying basis. Reducing the number of knots reduces the flexibility of the fitted spline. There is also a need for minimizing the number of knots to avoid overfitting. On the other hand, usage of the penalty term avoids the overfitting and lets one use the sufficient number of knots.

A substantial question is how much η_2 is allowed to vary. Note that if its variation is arbitrary, then our model (2.14) is over-parameterized. This can and should be avoided by a penalty term. For a given sample, this is done by defining the regression estimators as the minimizers over $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ of

$$(\boldsymbol{Y} - \tilde{\eta}(\boldsymbol{F}))^t(\boldsymbol{Y} - \tilde{\eta}(\boldsymbol{F})) + \lambda \boldsymbol{\eta}_2^t \boldsymbol{\eta}_2, \tag{2.15}$$

where λ is the penalty term or smoothing parameter which controls the bias-variance ratio. Note that if it is zero, we have no bias but large variance; if it is large we have large bias but low variance. So we know already intuitively that it must be proportional to the ratio of the variance of ϵ divided by the variance of η_2. Under these circumstances, estimating the smoothness for the model is now estimating the smoothing parameter λ or the variances of the error and η_2. If λ is too high, the data will be over-smoothed. If it is too low, then the data will be under-smoothed. In either case, the spline estimate $\tilde{\eta}$ will not be close to the true function.

With this in mind, the penalty for the $\boldsymbol{\eta}(\boldsymbol{F})$ function can be estimated via $\boldsymbol{\lambda} = \hat{\sigma}_e^2 / \hat{\boldsymbol{\sigma}}_{\eta_2}^2$ with $\boldsymbol{\lambda} = (\lambda_{1_1}, ..., \lambda_{1_J})'$. Another solution will be to set the λ and therefore fix the smoothness. Note that this corresponds to some extend to the inverse of the slider in Lombardía and Sperlich (2011). There are three different methods, namely ordinary cross validation (OCV), generalized cross validation (GCV), and assuming distributions for ϵ and η_2 and estimating their variances. GCV has computational advantages over OCV (Wahba, 1990). We will concentrate on the last method based on the idea of linear mixed effects model estimation (see e.g. Ruppert, Wand, and Carroll, 2003).

2.2.3 Extended model description and mixed effects model representation

Combining now the semi-mixed effects modeling idea of Lombardía and Sperlich (2011), see our Section 2.0.3 with the spline implementation of above, we consider the following

two-level model

$$Y_{id} = X_{id}\boldsymbol{\beta} + \gamma(F_{id}) + \eta(W_d) + u_d + \epsilon_{id}, \tag{2.16}$$

where $d = 1, .., D$ are the indices for the area and $i = 1, ..., n_d$ the indices for the subjects or individuals in area d, i.e. the index d runs over the small areas and i runs over the elements of each areas. If we consider a panel data study, i may refer to time and d may refer to the individual. Let $\boldsymbol{Y} \in I\!R^n$ be the vector of $n = \sum_{d=1}^{D}(n_d)$ responses, $\boldsymbol{X} \in I\!R^{n\times q}$ and $\boldsymbol{F} \in I\!R^{n\times r_1}$ matrices containing, respectively, q and r_1 covariates for the n individuals. $\boldsymbol{X}$ contains also one column of ones for the constant, say β_0. Let further $\boldsymbol{W} \in I\!R^{D\times r_2}$ indicate the matrices of the regional covariates, and $\boldsymbol{Z} \in I\!R^{n\times D}$ a matrix of ones and zeros indicating in what area the individual lives, $\boldsymbol{u} \in I\!R^D$ random area effects, and the remaining unobserved individual effects $\boldsymbol{\epsilon} \in I\!R^n$, where $\boldsymbol{u} \perp \boldsymbol{\epsilon}$, i.e. independence is assumed. Let $\boldsymbol{\beta} \in I\!R^{q+1}$ be a fixed effect, $\boldsymbol{\gamma} : I\!R^{r_1} \to I\!R$ and $\boldsymbol{\eta} : I\!R^{r_2} \to I\!R$ nonparametric unknown but smooth functions that have to be estimated. Then, in matrix notation we can rewrite the model as follows:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\gamma}(\boldsymbol{F}) + \boldsymbol{\eta}(\boldsymbol{W}) + \boldsymbol{Zu} + \boldsymbol{\epsilon}. \tag{2.17}$$

Recall that for consistent estimation, using classical methods, we further need the independence between covariates and area effects. In our model, area effects are separated into a controlled (say deterministic) effect $\boldsymbol{\eta}(W)$ and the random one $\boldsymbol{u}$. The idea is that an appropriate choice of $\boldsymbol{\eta}$ filters possible dependence between the covariates and the area remainder $\boldsymbol{u}$.

If $\boldsymbol{\gamma}(\boldsymbol{F}) \neq 0$ and $\boldsymbol{\eta}(\boldsymbol{W}) \neq 0$, then combining the P-spline approximation (2.14) with the model (2.16), we can rewrite the model as

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{S\theta} + \boldsymbol{Gh} + \boldsymbol{M\delta} + \boldsymbol{Lv} + \boldsymbol{Zu} + \boldsymbol{\epsilon}, \tag{2.18}$$

where we defined the following matrices,

$$\boldsymbol{S} = \begin{bmatrix} F_{11} & \cdots & F_{11}^p & |\cdots| & F_{r_1 1} & \cdots & F_{r_1 1}^p \\ & \vdots & & & & \vdots & \\ F_{1n} & \cdots & F_{1n}^p & |\cdots| & F_{r_1 n} & \cdots & F_{r_1 n}^p \end{bmatrix},$$

$$\boldsymbol{G} = \begin{bmatrix} (F_{11}-\rho_{11})_+^p & \cdots & (F_{11}-\rho_{1k_1})_+^p & |\cdots| & (F_{r_1 1}-\rho_{11})_+^p & \cdots & (F_{r_1 1}-\rho_{1k_1})_+^p \\ & \vdots & & & & \vdots & \\ (F_{1n}-\rho_{11})_+^p & \cdots & (F_{1n}-\rho_{1k_1})_+^p & |\cdots| & (F_{r_1 n}-\rho_{11})_+^p & \cdots & (F_{r_1 n}-\rho_{1k_1})_+^p \end{bmatrix},$$

$$\boldsymbol{M} = \begin{bmatrix} W_{11} & \cdots & W_{11}^p & |\cdots| & W_{r_21} & \cdots & W_{r_21}^p \\ & \vdots & & & & \vdots & \\ W_{1n} & \cdots & W_{1n}^p & |\cdots| & W_{r_2n} & \cdots & W_{r_2n}^p \end{bmatrix},$$

$$\boldsymbol{L} = \begin{bmatrix} (W_{11}-\tau_{11})_+^p & \cdots & (W_{11}-\tau_{1k_2})_+^p & |\cdots| & (W_{r_21}-\tau_{11})_+^p & \cdots & (W_{r_21}-\tau_{1k_2})_+^p \\ & \vdots & & & & \vdots & \\ (W_{1n}-\tau_{11})_+^p & \cdots & (W_{1n}-\tau_{1k_2})_+^p & |\cdots| & (W_{r_2n}-\tau_{11})_+^p & \cdots & (W_{r_2n}-\tau_{1k_2})_+^p \end{bmatrix},$$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{r_1})' \in I\!R^{pr_1}$ being a fixed parameter with $\boldsymbol{\theta}_{r_1} = (\theta_{r_11}, ..., \theta_{r_1p})'$, $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_{r_2})' \in I\!R^{pr_2}$ being a fixed parameter with $\boldsymbol{\delta}_{r_2} = (\delta_{r_21}, ..., \delta_{r_2p})'$.

Further, we have $\boldsymbol{h} = (\boldsymbol{h}_1, ..., \boldsymbol{h}_{r_1})' \in I\!R^{k_1r_1}$ being the first spline's pseudo random effect with $\boldsymbol{h}_{r_1} = (h_{r_11}, ..., h_{r_1k_1})' \in N(\boldsymbol{0}, \boldsymbol{I}\boldsymbol{\sigma}_h^2)$, $\boldsymbol{\sigma}_h^2 = (\sigma_{h_1}^2, ..., \sigma_{h_{r_1}}^2)'$, and $\boldsymbol{v} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_{r_2})' \in I\!R^{k_2r_2}$ being the second spline's pseudo random effect with $\boldsymbol{v}_{r_2} = (v_{r_21}, ..., v_{r_2k_2})' \in N(\boldsymbol{0}, \boldsymbol{I}\boldsymbol{\sigma}_v^2)$, $\boldsymbol{\sigma}_v^2 = (\sigma_{v_1}^2, ..., \sigma_{v_{r_2}}^2)'$. The remaining terms have already be defined before.

Then, for the sake of implementation, the model can be rewritten in matrix notation as

$$Y = \boldsymbol{T}\boldsymbol{\alpha} + \boldsymbol{C}\boldsymbol{\xi} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{2.19}$$

by merging fixed to fixed and pseudo-random to random parts, where $\boldsymbol{T} = [\boldsymbol{X}\,\boldsymbol{S}\,\boldsymbol{M}]$, $\boldsymbol{\alpha} = [\boldsymbol{\beta}\,\boldsymbol{\theta}\,\boldsymbol{\delta}]'$, $\boldsymbol{C} = [\boldsymbol{G}\,\boldsymbol{L}]$ and $\boldsymbol{\xi} = [\boldsymbol{h}\,\boldsymbol{v}]'$. Finally, $\boldsymbol{u} \sim N(0, \boldsymbol{\Sigma}_u = \sigma_u^2\boldsymbol{I}_D)$, $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}_\epsilon = \sigma_e^2\boldsymbol{I}_n)$ and $\boldsymbol{\xi} \sim N(0, \boldsymbol{\Sigma}_\xi = diag[\boldsymbol{I}\sigma_h^2, \boldsymbol{I}\sigma_v^2])$.

Then, we define $\boldsymbol{\Sigma}_y = \boldsymbol{C}\boldsymbol{\Sigma}_\xi\boldsymbol{C}' + \boldsymbol{Z}\boldsymbol{\Sigma}_u\boldsymbol{Z}' + \boldsymbol{\Sigma}_\epsilon$. If the variances of the random and pseudo-random components (i.e. in the latter case the smoothing parameters) were known, the standard results from BLUP theory (McCulloch and Searle, 2001) guarantee that given the model specifications

$$\hat{\alpha} = (\boldsymbol{T}'\boldsymbol{\Sigma}_y^{-1}\boldsymbol{T})^{-1})\boldsymbol{T}'\boldsymbol{\Sigma}_y^{-1}\boldsymbol{Y} \tag{2.20}$$

is the BLUE (best linear unbiased predictor), and consequently

$$\hat{\boldsymbol{\xi}} = \boldsymbol{\Sigma}_\xi\boldsymbol{C}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{Y} - \boldsymbol{T}\hat{\alpha}),\ \hat{\boldsymbol{u}} = \boldsymbol{\Sigma}_u\boldsymbol{Z}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{Y} - \boldsymbol{T}\hat{\alpha} - \boldsymbol{C}\hat{\boldsymbol{\xi}}) \tag{2.21}$$

are the BLUPs (best linear unbiased predictors).

Alternatively, we can estimate the variance components simultaneously by ML method with or without a correction, see Rao (2003). For notational convenience, we write estimation of σ_h^2, σ_v^2 on the one hand but σ_u^2 on the other hand separately to distinguish the random part from the splines' pseudo random parts:

$$\hat{\sigma}_u^2 = (\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}} + \hat{\sigma}_u^2\, trace(\boldsymbol{T}_{ii}^*))/D,$$

$$\hat{\boldsymbol{\sigma}}_h^2 = (\hat{\boldsymbol{h}}'\hat{\boldsymbol{h}} + \hat{\boldsymbol{\sigma}}_h^2\, trace(\boldsymbol{T}_{ii}^{*1}))/k_1,\ \boldsymbol{\sigma}_h^2 = (\sigma_{h_1}^2, ..., \sigma_{h_{r_1}}^2)',$$
$$\hat{\boldsymbol{\sigma}}_v^2 = (\hat{\boldsymbol{v}}'\hat{\boldsymbol{v}} + \hat{\boldsymbol{\sigma}}_v^2\, trace(\boldsymbol{T}_{ii}^{*2}))/k_2,\ \boldsymbol{\sigma}_v^2 = (\sigma_{v_1}^2, ..., \sigma_{v_{r_2}}^2)',$$

$$\hat{\sigma}_e^2 = (\boldsymbol{Y}'\hat{\boldsymbol{\epsilon}})/n$$

with $\boldsymbol{T}_{ii}^* = (\boldsymbol{I} + \boldsymbol{Z}'(\hat{\sigma}_e^2)^{-1}\boldsymbol{Z}\hat{\sigma}_u^2)^{-1}$ (here, each i runs over D), $\boldsymbol{T}_{ii}^{*1} = (\boldsymbol{I} + \boldsymbol{G}'(\hat{\sigma}_e^2)^{-1}\boldsymbol{G}\hat{\boldsymbol{\Sigma}}_h)^{-1}$ (here, each i runs over $r_1 * k_1$), $\boldsymbol{T}_{ii}^{*2} = (\boldsymbol{I} + \boldsymbol{L}'(\hat{\sigma}_e^2)^{-1}\boldsymbol{L}\hat{\boldsymbol{\Sigma}}_v)^{-1}$ (here, each i runs over $r_2 * k_2$), and $\hat{\boldsymbol{\Sigma}}_y = \boldsymbol{C}\hat{\boldsymbol{\Sigma}}_\xi\boldsymbol{C}' + \boldsymbol{Z}\hat{\sigma}_u^2\boldsymbol{Z}' + \hat{\sigma}_e^2\boldsymbol{I}_n$ where $\boldsymbol{\Sigma}_\xi = diag[\boldsymbol{I}\sigma_h^2, \boldsymbol{I}\sigma_v^2]$. Note that this made easy with additivity.

As can be seen, an iteration is necessary if REML (restricted maximum likelihood estimation) cannot be applied for the estimation of variance components. The iteration runs over estimating the fixed effects, predicting the random effects, and finally estimating the variance components to restart with the fixed effects estimation and so on. We typically stopped if the fixed effects vectors did not change more than 1 percent compared to the last iteration's outcome.

Further alternatives have been implemented to account for the possibility of only estimating the variances of the truly random parts $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ but fix the smoothness of function γ or pre-determine the slider for our dependence filter function η. It turned out that (a) to distinguish between random and pseudo-random only makes sense if we want to assume different distributions, and even then it is not evident what numerically happens, (b) if one wants to fix the smoothness - you can also speak of pre-setting the λs - it is better from an implementation point of view to simply fix σ_h^2 and/or σ_v^2. We conclude that one of these implementations is sufficient. Other extensions were implemented to account for possible heteroskedasticity of either the random effects, the pseudo-random effects, or the residuals. Note that heteroskedasticity of σ_h^2 and σ_v^2 simply cause locally different smoothness that might be wanted or not. In contrast, heteroskedasticity of $\boldsymbol{u}$ or $\boldsymbol{\epsilon}$ have a quite different interpretation.

Allowing for local smoothing, or equivalently from an implementational point of view, for the (pseudo-) random effects has been studied in different works (see e.g Brumback and Rice, 1998 or Ruppert, Wand, and Carroll, 2003). Our implementation for the heteroskedasticity of the residuals did work but not very well. We basically followed White's well known approach in econometrics by simply using the squares of residuals on the variance matrices. A much more successful extension to incorporate simultaneously heteroskedasticity for the error term (recall that this effects the smoothness parameter and, in our case certainly also the random effects prediction) can be found in Crainiceanu, Ruppert, Carroll, Joshi, and Goodner (2007), Krivobokova, Crainiceanu, and Kauermann (2008) or Wiesenfarth, Krivobokova, Klasen, and Sperlich (2011).

For the construction uniform confidence bands and inference tests in our type of models and methods, see Sperlich and Lombardía (2010) which use kernel smoothing and bootstrap based inference, or Wiesenfarth, Krivobokova, Klasen, and Sperlich (2011) who used P-splines for estimation and the the volume-of-tube formula for inference.

3 An application in environmetrics related to small area estimation

3.1 Abstract

Mixed effects models allow for efficient estimation of the fixed parts in the model. Also, they treat the small area effects as random effects, to account for the between-area variations beyond that which is explained by the variations in model covariates, and they make use of the random effects for prediction. In the moment of prediction, one adds the predicted random effect to the total prediction. The additional variance of the prediction caused by assuming this effect to be random is only slightly larger than the variance of a fixed effect estimate based on small samples, but the modeling of the new variance structure allows for a more efficient estimation of the coefficients. It might improve prediction in the mean, but under the assumption of independence between random effects and the covariates. In this study, we try to relax this assumption with a semiparametric modeling using splines approach. We carry out an application that will address an environmental small area problem. To overcome the dependencies between the random effects and the covariates, we include area-specific random effects in the model. We estimate the nonparametric functions in our semiparametric model by using P-splines and thin plate splines. Thin plate splines, introduced to geometric design by Duchon (1977), are isotropic smoothers and are appropriate for spatial coordinates. Penalized spline regression, or P-splines, were popularized by Eilers and Marx (1996). P-splines are attractive as a smoothing method because of their flexibility, and also a natural candidate for constructing nonparametric small area estimation. They are easy to implement and allow for additivity where we can model the explanatory variables nonparametrically.

This chapter is a joint work with Lombardía, M. J. and Sperlich, S..

3.2 Introduction

As in the last chapter we have already introduced some details of spline estimation and of the semi-mixed effects model basics, we concentrate here rather on what is new in this Chapter:

Small area estimation. We briefly mentioned this topic in the earlier chapter as motivating the use of mixed effects models. What we did not discuss are the main statistical challenges when using mixed effects model estimation and prediction.

In our environmental context, the notation of small areas will indeed refer to what one is intuitively thinking of: small geographical areas. For environmental questions, this automatically entails the question: to what extent geographical correlation does matter.

"Small area" is the term that is used to refer, generally, to a small area or population, though it may also refer to an isolated particular demographic. If a survey has been carried out for a whole population, a problem arises when trying to generate accurate estimates relative to any particular small area within this population, because the sample in many of the areas may be (moreover, typically is) too small. While design-based inference methods may be appropriate for the overall survey sample size, one has to rely on alternative methods, namely model-based, when looking at small domains. The idea is based on the availability of that population level auxiliary information. In those circumstances, the statistical techniques involving the estimation of parameters are simply called "small area estimation" (see, for example, Rao, 2003). During the last few decades, mixed effects models have attracted a lot of attention and have consequently been widely used in small area estimation (see e.g. Lombardía and Sperlich, 2008 or Opsomer, Claeskens, Ranalli, Kauermann, and Breidt, 2008). For combining information from various sources and explaining different sources of errors, these models offer great flexibility and are well suited for solving most problems in small area estimation.

To be more specific: in small area statistics, one often is not that much interested in the model even though in practice it might then be used and interpreted. What is of interest are area level parameters, say macro-parameters for each area. Typical examples are: average income in a municipality, average added value produced by firms inside a county, average tourist expenditures for each province, percentage of unemployment for certain areas, percentage of poverty in certain regions, etc. As can be guessed from these examples,

small area statistics are an important issue for statistical institutes and administrative statistics in general. See, for example, Ghosh and Rao (1994), Lahiri and Rao (1995), Ugarte, Goicoa, Militino and Durbán (2009) or Lombardía and Sperlich (2011).

A problem is that, given the available data, these numbers have to be estimated or predicted. Although the point estimator might be the most interesting number for the politician or many decision makers, we know in statistics, econometrics and biometrics that this number, especially if we think of small samples and prediction, can be quite uninformative and even misleading unless reliable confidence and prediction intervals are provided. Given the used statistical model and estimator, these might be quite complicated to obtain.

To reduce the variability inherited by the nature of being 'small' (small area, small sample), the model based approach is the most popular one, and when thinking of 'very small areas' we assume to have many areas (say D) with few observations (say n_d, $d = 1, \ldots, D$). Then, it is more reasonable to work with random area effects than with fixed ones given the particular mixed effects model. Depending on the estimation method, one may use distributional assumption on the random effects and error terms or not. Also, the model choice is of certain importance even though the model itself may not be of further interest, but further inference is model based and therefore only valid if the model is correct. One crucial assumption, then, is the mutual independence between covariates, random effects and error terms.

To conclude, in small area statistics we face the following sequence of statistical problems: estimate or predict the area-parameters though the sample is too small for many areas; use a model based approach and choose an adequate model where 'adequate' includes the problematic independence assumption; depending on the estimation and prediction method, one may have to choose or model the distributions of the random effects and error term; from the model estimates, calculate the wanted parameters for each area and construct a prediction interval. The last point is equivalent to the estimation of the prediction squared error. The latter point has maybe been the most studied point in small area statistics for about few decades. (see e.g. Prasad and Rao, 1990, Lahiri, 2003, Das, Jiang and Rao, 2004 or Lombardía and Sperlich, 2011)

We will consider here not an econometric problem like the given examples above but an environmental one. Not surprisingly, also in environmetrics, small area estimation has become a more and more popular tool (see, for example, Militino, Ugarte and Goicoa,

2006 or Pratesi, Ranalli, and Salvati, 2008). We consider data from 1991 to 1996 of 334 lakes in the north-eastern states of the U.S.A.. The figure of interest is the acid neutralizing capacity of the water, say $\boldsymbol{Y}$. As we will use three continuous explanatory variables ($\boldsymbol{F}_1$, $\boldsymbol{F}_2$), an intercept plus a continuous variable that is 'known' to have a linear impact ($\boldsymbol{X}$), and the random small area effects $\boldsymbol{u}$. When looking at geographical areas and environmental data, one would intuitively suppose to face spatial correlation in the response that could pass to a spatial correlation in the random area effects; see Gosh, Natarajan, Stroud and Carlin (1998). This, however, does ignore the chance to have correlation between the covariates and random effects as probably both are related to where they have been observed geographically. Therefore we prefer to include the geographical location in the deterministic (fixed) part of the model. This will be done via thin plate spline (TPS) regression of the coordinates ($\boldsymbol{C}$); see Wood (2003). Finally, we include two area-specific variables ($\boldsymbol{W}_1$, $\boldsymbol{W}_2$) to filter possible dependence between $\boldsymbol{u}$ and the ($\boldsymbol{F}_1, \boldsymbol{F}_2$). We now concentrate on the estimation of our model.

Summarizing, we will work with the following model:

$$Y_{id} = X_{id}\beta + \delta(C_i) + \gamma_1(F_{1id}) + \gamma_2(F_{2id}) + \eta_1(W_{1d}) + \eta_2(W_{2d}) + u_d + \epsilon_{id}, \tag{3.1}$$

where $d = 1, .., D$ and $i = 1, ..., n_d$. The index d runs over the small areas and i runs over the elements of each area. In matrix notation this is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\delta}(\boldsymbol{C}) + \boldsymbol{\gamma}_1(\boldsymbol{F}_1) + \boldsymbol{\gamma}_2(\boldsymbol{F}_2) + \boldsymbol{\eta}_1(\boldsymbol{W}_1) + \boldsymbol{\eta}_2(\boldsymbol{W}_2) + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}. \tag{3.2}$$

As said, $\boldsymbol{Y} \in I\!R^n$ is the vector of n responses, $\boldsymbol{X} \in I\!R^{n\times q}$ a matrix containing q covariates for n individuals including one column of ones for the constant β_0. Let $\boldsymbol{C}_i = (c_{1i}, c_{2i})$ denote the geographical coordinates for the observations, $\boldsymbol{F}_1 \in I\!R^{n\times r_1}$, $\boldsymbol{F}_2 \in I\!R^{n\times r_1}$ and $\boldsymbol{W}_1 \in I\!R^{D\times r_2}$, $\boldsymbol{W}_2 \in I\!R^{D\times r_2}$ the matrices of the individual and the area covariates. $\boldsymbol{Z} \in I\!R^{n\times D}$ is a matrix of ones and zeros indicating in what area the observation is located. Finally, we have $\boldsymbol{u} \in I\!R^D$, the random area effects, and the remaining heterogeneity $\boldsymbol{\epsilon} \in I\!R^n$, where $\boldsymbol{u} \perp \boldsymbol{\epsilon}$ independent.

Then, we have to estimate $\boldsymbol{\beta} \in I\!R^{q+1}$, an unknown fixed vector. $\boldsymbol{\delta} : I\!R^2 \to I\!R$, $\boldsymbol{\gamma}_1 : I\!R^{r_1} \to I\!R$, $\boldsymbol{\gamma}_2 : I\!R^{r_1} \to I\!R$ and $\boldsymbol{\eta}_1 : I\!R^{r_2} \to I\!R$, $\boldsymbol{\eta}_2 : I\!R^{r_2} \to I\!R$ are the smooth functions that have to be estimated by thin plate and cubic splines.

So, the aim of this chapter is to avoid possible dependence between the random effects and the covariates with a semiparametric modeling approach using splines. We do this along

a case study where we analyze a survey of lake water quality in the north-eastern states of the U.S.A.. The survey was conducted by the Environmental Monitoring and Assessment Program (EMAP) of the Environmental Protection Agency (EPA) from 1991 to 1996. In the next section, we reconsider the P-splines. Afterwards, we introduce briefly the basic ideas of thin plate spline regression to better account for spatial smoothing. Thanks to this property, they should play a more important role in small area estimation. We close that section with a short revision of how the prediction mean squared error can be calculated for our spline estimation approach. A small Monte Carlo simulation study shall illustrate the estimation performance of the proposed model in our context, see Section 3.4. Then, in Section 3.5, we will carry out the case study based on the aforementioned data with the introduced methods.

3.3 Spline models, bases and prediction error estimation

Given the huge amount of literature on spline regression, here we try to be brief and just indicate the basic ideas of the methods we will apply in the following sections. For the sake of presentation, we will separate the discussion for the cubic splines which will be used for estimating one dimensional additive functions γ_1, γ_2, η_1, η_2 in model (3.1), and thin plate splines used for estimating the two-dimensional function δ. We further simplify the notation by considering a one-dimensional nonparametric regression problem; the simple model

$$y_i = m(x_i) + \epsilon_i, \qquad i = 1, \ldots, n \tag{3.3}$$

with ϵ_i being independent and identically distributed (i.i.d.) random terms independent of x_i and among themselves.

While some bases are numerically more stable and allow computation of a model fit with better accuracy, the change in basis does not, in principle, change the fit. Aside from numerical stability, one may consider ease of implementation and interpretability in order to select one basis over another. We start with a typical simple basis for P-spline regression.

For understanding the underlying mechanics of spline-based regression, one can study the truncated power bases. These bases can be applied easily in practice as long as the knots are selected carefully or a penalized fit is used. For given knots $\kappa_1, \cdots, \kappa_K$ consider the

truncated power basis of degree p

$$1, x, ..., x^p, (x - \kappa_1)_+^p, ..., (x - \kappa_k)_+^p, \tag{3.4}$$

where the + signs indicate that the bracket will be zero if the expression inside is negative. Then, the pth-degree P-spline model can be written as

$$m(x; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)_+^p. \tag{3.5}$$

A commonly used modification of the cubic spline model is the natural cubic spline basis. The truncated cubic basis for the knots $\kappa_1, ..., \kappa_k$ is

$$1, x, x^2, x^3, (x - \kappa_1)_+^3, ..., (x - \kappa_k)_+^3. \tag{3.6}$$

A *cubic smoothing spline*, m, minimizes the residual sum of squares and the penalty on the integral of the squared second derivative $(m'')^2$

$$\sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int (m(x)'')^2 dx \tag{3.7}$$

where λ is the penalty term that plays crucial role in determining the smoothness of the model. We have seen in the first chapter that one can write this like a linear mixed effects model estimation problem where the coefficients of the $(x - \kappa_1)_+^3, ..., (x - \kappa_k)_+^3$ are treated like random coefficients orthogonal to $1, x, x^2, x^3$. In practice, the knots are a subsequence of the ordered observations x_i. For further details see Ruppert, Wand and Carroll (2003).

Since both the P-spline and the small area estimation are random effects models, it is not difficult to combine both into a semiparametric small area estimation of an environmental problem based on mixed effects model regression. Furthermore, there are cases where x is multivariate, other bases might be preferable to the truncated polynomials. For our study we combine cubic truncated power basis with thin plate splines, as the latter are most appropriate for the spatial coordinates, cf. model (3.1).

Again, consider the problem of estimating the smooth function $m(x_i)$ in the model (3.3). Let us fix w as the highest order of derivative of m we plan to control to measure wiggliness. Considering geographical coordinates, x is now a two-dimensional variable. Then, thin plate spline smoothing estimates function m by finding the function $\hat{f}$ that minimizes

$$\|y - f\|^2 + \lambda J_w(f), \tag{3.8}$$

where y is a vector of y_i data and $f = [f(x_1), ..., f(x_n)]'$. $J_w(f)$ is a penalty functional measuring the wiggliness of f, and λ is a smoothing parameter boosting or shrinking multiplicatively the penalty done by

$$J_w = \int\int \sum_{v_1+v_2=w} \frac{w!}{v_1!v_2!} \left(\frac{\delta^w f}{\delta x_1^{v_1} \delta x_2^{v_2}} \right)^2 dx_1 dx_2,$$

where $2w > d_x$ =dimension of x. It is obvious how this can be extended to higher dimensions. It can be shown that the function minimizing (3.8) has the form

$$\hat{f}(x) = \sum_{i=1}^{n} \delta_i \eta_w(\|x - x_i\|) + \sum_{j=1}^{M} \alpha_j \phi_j(x), \tag{3.9}$$

where $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ are vectors of the coefficients to be estimated, $M = \binom{w+d_x-1}{d_x}$, and ϕ_j are linearly independent polynomials spanning the space of polynomials of degree less than w. Next, δ is subject to the linear constraints that $\boldsymbol{T}'\boldsymbol{\delta} = 0$ where $T_{ij} = \phi_j(x_i)$, and η_w is defined as

$$\eta_w(r) = \begin{cases} \frac{(-1)^{w+1+d_x/2}}{2^{2w-1}\pi^{d_x/2}(w-1)!(w-d_x/2)!} r^{2w-d_x} log(r) & \text{if } d_x \text{ even} \\ \frac{\Gamma(d_x/2-w)}{2^{2w}\pi^{d_x/2}(w-1)!} r^{2w-d_x} & \text{if } d_x \text{ odd} \end{cases}$$

Defining matrix $\boldsymbol{E}$ by $E_ij \equiv \eta_w(\|x - x_i\|)$, the thin plate spline fitting becomes

$$minimize \quad \|\boldsymbol{y} - \boldsymbol{E\delta} - \boldsymbol{T\alpha}\|^2 + \lambda \boldsymbol{\delta}' \boldsymbol{E\delta} \quad \text{subject to } \boldsymbol{T}'\boldsymbol{\delta} = \boldsymbol{0}\ , \tag{3.10}$$

with respect to $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$.

Penalized regression splines simply use fewer spline basis functions. There are two alternatives: knot based and Eigen based splines. In knot based splines, we choose a subset of our data, namely the knots, and create the basis as if we are only smoothing that data. In Eigen based, we choose the number of basis functions (which are simply given by the Eigenvectors of $\boldsymbol{E}$) to construct the spline that will optimally approximate a full spline. In that sense thin plate splines are easy to implement, but we still have to choose the K, the number of basis dimension, for thin plate splines. One has to decide roughly how large the basis dimension is. The number of basis dimension for the model has to be fairly certain enough to provide flexibility. For further details we refer to Green and Silverman (1994) or Wood (2003).

At the end of this section we give the formulas to estimate the variance of the model based predictors of $\bar{y}_d$. In order to do so, we first have to extend our model (3.3). It has already

been shown in the last chapter but can also be seen from the book of Ruppert, Wand and Carroll (2003) that we can write our spline model (with also real $\boldsymbol{u}$ and with pseudo random effects, say $\boldsymbol{\xi}$) in terms of

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{G}\xi + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{3.11}$$

where $\boldsymbol{X}$ now refers to the non-truncated spline part, $\boldsymbol{G}$ in case of just one dimensional cubic splines would consist of the $(x_i - \kappa_1)_+^3, ..., (x_i - \kappa_k)_+^3$, and $\boldsymbol{Z}$ as usual indicating in which area each observation is located. An obvious candidate to estimate for $\bar{y}_d$ is

$$\hat{y}_d = \bar{\boldsymbol{x}}_d\hat{\beta} + \bar{\boldsymbol{g}}_d\hat{\xi} + \boldsymbol{e}_d\hat{u}, \tag{3.12}$$

where the bars indicate true means for each dimension (power) of our variables, the hats indicate estimates or predictors, and $\boldsymbol{e}_d$ is a zero-vector with a 1 at its d-th position.

Set $Var[\boldsymbol{Y}] = \boldsymbol{\Sigma}_y$, $\boldsymbol{\Sigma}_\xi$, $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\epsilon$ being the variance-covariance matrices of the (pseudo) random terms, $\boldsymbol{W} = (\boldsymbol{G}, \boldsymbol{Z})$, $\boldsymbol{\omega} = (\boldsymbol{\xi}', \boldsymbol{u}')$, $\bar{w}_d = (\bar{g}_d, \boldsymbol{e}_d)$, and

$$\boldsymbol{\Sigma}_\omega = \begin{pmatrix} \boldsymbol{\Sigma}_\xi & 0 \\ 0 & \boldsymbol{\Sigma}_u \end{pmatrix}$$

Then for $\boldsymbol{c}_d = \bar{x}_d - \bar{w}_d\boldsymbol{\Sigma}_\omega\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{X}$ we have

$$\hat{y}_d - \bar{y}_d = \boldsymbol{c}_d(\hat{\beta} - \boldsymbol{\beta}) + \bar{w}_d\left(\boldsymbol{\Sigma}_\omega\boldsymbol{W}^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{Y} - \boldsymbol{X}\beta) - \boldsymbol{\omega}\right) \tag{3.13}$$

We know from Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008) that we can estimate the prediction mean square error via

$$\hat{c}_d(\boldsymbol{X}'\hat{\Sigma}_y^{-1}\boldsymbol{X})^{-1}\hat{c}_d' + \bar{w}_d\hat{\Sigma}_\omega\left(\boldsymbol{I} - \boldsymbol{W}'\hat{\Sigma}_y^{-1}\boldsymbol{W}\hat{\Sigma}_\omega\right)\bar{w}_d' + 2(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})'\hat{S}'\hat{I}^{-1}\hat{S}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{3.14}$$

with $\boldsymbol{I}$ being the identity matrix, and $\boldsymbol{S}$ being a matrix with rows

$$\boldsymbol{S}_j = \bar{w}_d\left(\frac{\boldsymbol{\delta}\boldsymbol{\Sigma}_\omega}{\boldsymbol{\delta}(\boldsymbol{\sigma}^2)_j}\boldsymbol{W}'\boldsymbol{\Sigma}_y^{-1} + \boldsymbol{\Sigma}_\omega\boldsymbol{W}'\frac{\boldsymbol{\delta}\boldsymbol{\Sigma}_y^{-1}}{\boldsymbol{\delta}(\boldsymbol{\sigma}^2)_j}\right)\ ,\ j = 1, 2, 3$$

where $(\boldsymbol{\sigma}^2)_j$ refers to the variances of $\boldsymbol{\xi}$, u and $\boldsymbol{\epsilon}$ respectively. Where we put hats, the variances were replaced by their estimates. Further, $\hat{I}$ is a 3×3 matrix with elements ij

$$\frac{1}{2}trace(\boldsymbol{P}\boldsymbol{B}_i\boldsymbol{P}\boldsymbol{B}_j)\ , \quad \text{with } \boldsymbol{P} = \hat{\Sigma}_y^{-1} - \hat{\Sigma}_y^{-1}\boldsymbol{X}(\boldsymbol{X}'\hat{\Sigma}_y^{-1}\boldsymbol{X})^{-1};$$

further $\boldsymbol{B}_1 = \boldsymbol{G}\boldsymbol{G}'$, $\boldsymbol{B}_2 = \boldsymbol{Z}\boldsymbol{Z}'$ and $\boldsymbol{B}_3$ the identity matrix of rank n.

An alternative to this linear approximation is doing bootstrap; see, for example, Lombardía and Sperlich (2008).

3.4 A small illustrative simulation

As the cubic spline smoothing has been studied intensively in many simulations and applications over decades (see e.g. Wahba, 1990, Gu, 2002 or Ruppert, Wand and Carroll, 2003), we concentrate here exclusively on the much less known thin plate spline performance. Moreover, we will give a special emphasis on the performance of the variance estimation. Here, with variances are also meant the parameters indicating the variation of our pseudo-random effects. So these are rather smoothing parameter estimates than variance estimates. To this aim we will consider models somewhat more complex than those considered in the last section but coming closer to the model we plan to study in our small area environmental study.

A thin plate spline is an isotropic smoother, which is appropriate for spatial coordinates (see e.g. Duchon, 1977). For construction of our small area estimation model, we will be using the coordinates of each lake. To demonstrate the practical performance of the thin plate splines, we carry out a simulation study. For the simulation consider the following test function

$$\begin{aligned} f(c_1, c_2) &= 1.2exp(-(c_1 - 0.2)^2/\sigma_{c_1}^2 - (c_2 - 0.3)^2/\sigma_{c_2}^2) \\ &\quad +0.8exp(-(c_1 - 0.7)^2/\sigma_{c_1}^2 - (c_2 - 0.8)^2/\sigma_{c_2}^2) \end{aligned}$$

where $\sigma_{c_1}^2 = 0.3$ and $\sigma_{c_2}^2 = 0.4$. Let us first consider the simple model

$$y_{id} = \beta_0 + X_{id}\beta_i + f(c_{1id}, c_{2id}) + \upsilon_d + e_{id} \tag{3.15}$$

where we generate X_{id} from a normal distribution with variance 1 and mean function $0.8 + c_{1id}^2 + c_{2id}^2$ to correlate our data as otherwise the estimation problem would be too trivial. Set the basis dimension (number of knots, see above) to $K = 80$, and let the random effects be $\boldsymbol{u} \sim N(0, \sigma_u^2 = 0.05)$, and error term $\boldsymbol{e} \sim N(0, \sigma_e^2 = 0.1)$. We set the number of small areas to $D = 86$ with $n_d = 6$ observations in each resulting in a sample size of $n = 516$.

To fit the model we use the command gamm() of the mgcv package with identity link; see Wood (2003). Figure 3.1 shows the example of an estimate when data are generated with f_1 and estimated using the thin pate splines. As we consider here two dimensional functions, the figures are given as contour plots. Comparing the true data generating

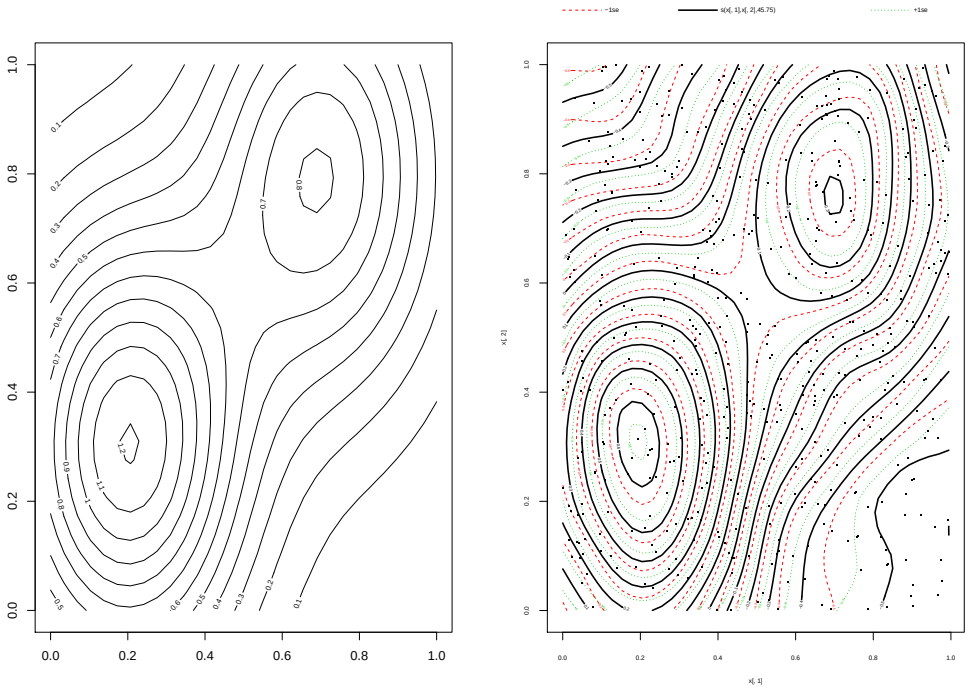

Figure 3.1: Left; true function, right; TPS term of the fit using function f.

process and its estimate, we see how thin plate splines can capture geographical or spatial features very well.

Now we use the test function f for further studies and consider the following model, i.e. the data generating process

$$\begin{aligned} y_{id} &= \beta_0 + X_{id}\boldsymbol{\beta} + f(c_{1i}, c_{2i}) + \sum_{r_1=1}^{r_1}[sin(a * F_{1r_1d}) + b] \\ &\quad + \sum_{r_2=1}^{r_2}[sin(a * F_{2r_2d}) + b] + u_d + \epsilon_{id} \end{aligned} \tag{3.16}$$

where now the function

$$X_{id} = 0.8 + \sum_{r=1}^{r} c_{1rd}^2 + \sum_{r=1}^{r} c_{2rd}^2 + \sum_{r_1=1}^{r_1} F_{1r_1d}^2 + \sum_{r_1=1}^{r_1} F_{2r_2d}^2 \tag{3.17}$$

is used to correlate the data. We generate $F_{r_1d} \in U_{[0,2]}$, $F_{r_2d} \in U_{[0,2]}$ where $r = r_1 = r_2 = 1$, $d = 1, ..., 86$, with $a = 2.5$, $b = 0$, $D = 86$. $n_d = 6$ and $n = \sum_{d=1}^{D}(n_d) = 516$. Here f is

still the two dimensional smooth function that has to be estimated with thin plate splines. The number of Monte Carlo iterations is set to 500.

The variance and the bias of the estimated $\hat{\sigma_u}$, $\hat{\sigma_e}$ are given in the Table 3.1. It can be seen that the automatic choice of smoothing parameter as well as the estimation of the true variances of our random terms work well for our context.

σ_e^2	σ_u^2	$Bias(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.10	0.05	-0.001035	-0.003289
	0.25	-0.002508	-0.003393
σ_e^2	σ_u^2	$Var(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.10	0.05	0.000109	0.000049
	0.25	0.001721	0.000049
σ_e^2	σ_u^2	$(1/m \sum_{i=1}^m \hat{\sigma}_{u_i}, 1/m \sum_{i=1}^m)\hat{\sigma}_{e_i})$	
0.10	0.05	0.048965	0.096711
	0.25	0.247492	0.096607
σ_e^2	σ_u^2	$Bias(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.50	0.05	-0.002820	-0.008778
	0.25	-0.005066	-0.009266
σ_e^2	σ_u^2	$Var(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.50	0.05	0.000455	0.001200
	0.25	0.002720	0.001207
σ_e^2	σ_u^2	$(1/m \sum_{i=1}^m \hat{\sigma}_{u_i}, 1/m \sum_{i=1}^m)\hat{\sigma}_{e_i})$	
0.50	0.05	0.047180	0.491222
	0.25	0.244934	0.490734

Table 3.1: d=86,n=516, first model.

Now consider the following model where we add the averages of the individual effects over

small areas

$$\begin{aligned} y_{id} &= \beta_0 + X_{id}\boldsymbol{\beta} + f_1(c_{1i}, c_{2i}) + \sum_{r_1=1}^{r_1}[sin(a * F_{1r_1n}) + b] \\ &\quad + \sum_{r_2=1}^{r_2}[sin(a * F_{2r_2n}) + b] + \sum_{r_2=1}^{r_2}[sin(a * \overline{W_{1r_2d}}) + b] \\ &\quad + \sum_{r_2=1}^{r_2}[sin(a * \overline{W_{2r_2d}}) + b]u_d + \epsilon_{id} \end{aligned} \tag{3.18}$$

where the following function

$$X_{id} = 0.8 + \sum_{r=1}^{r} c_{1rd}^2 + \sum_{r=1}^{r} c_{2rd}^2 + \sum_{r_1=1}^{r_1} F_{1r_1d}^2 + \sum_{r_1=1}^{r_1} F_{2r_1d}^2 + \sum_{r_2=1}^{r_2} W_{1r_2d}^2 + \sum_{r_2=1}^{r_2} W_{2r_2d}^2 \tag{3.19}$$

is used to correlate the data where $r = r_1 = r_2 = 1$. The variance and the bias of the estimated $\hat{\sigma_u}$, $\hat{\sigma_e}$ are given in the Table 3.2. The number of Monte Carlo iterations is set to 500.

Before we come to the application, we should also mention here that we did several studies for checking the quality of the curve estimation and the reliability of the confidence bands automatically provided in R. While the estimation works very well it turned out that, from a frequency interpretation, the confidence bands seem to be a little bit liberal. Nonetheless, in general we seem to have a quite powerful and reliable tool at hand to perform our environmental case study.

3.5 The environmental application

As mentioned earlier, the auxiliary information in small area estimation is often used via linear mixed model regression to improve the precision of survey estimators of finite population means and totals through linear or generalized regression estimation techniques. It is stated and nowadays accepted that the resulting estimators have good theoretical and practical properties. However, as Breidt, Opsomer, Johnson and Ranalli (2007) showed, it is not always clear that ratio or linear models are good approximations of the true relationship between the auxiliary variables, in our case $\boldsymbol{X}$, $\boldsymbol{C}$, $\boldsymbol{F}_1$ and $\boldsymbol{F}_2$, and the variable of interest in the survey. This results in a serious efficiency loss when the model is not appropriate. In the article of Breidt, Opsomer, Johnson and Ranalli (2007), it is explained

σ_e^2	σ_u^2	$Bias(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.10	0.05	-0.001978	-0.003323
	0.25	-0.008631	-0.003432
σ_e^2	σ_u^2	$Var(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.10	0.05	0.000117	0.000049
	0.25	0.001792	0.000049
σ_e^2	σ_u^2	$(1/m\sum_{i=1}^m \hat{\sigma}_{u_i}, 1m\sum_{i=1}^m)\hat{\sigma}_{e_i})$	
0.10	0.05	0.048022	0.096677
	0.25	0.241369	0.096568
σ_e^2	σ_u^2	$Bias(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.50	0.05	-0.005187	-0.009028
	0.25	-0.013827	-0.009432
σ_e^2	σ_u^2	$Var(\hat{\sigma}_u, \hat{\sigma}_e)$	
0.50	0.05	0.000453	0.001203
	0.25	0.002754	0.001200
σ_e^2	σ_u^2	$(1/m\sum_{i=1}^m \hat{\sigma}_{u_i}, 1/m\sum_{i=1}^m)\hat{\sigma}_{e_i})$	
0.50	0.05	0.044813	0.490972
	0.25	0.236173	0.490568

Table 3.2: d=86,n=516, model with averages.

how non- and semiparametric regression estimation can be extended in simple and more complicated designs. These much more flexible models maintain all the good theoretical and practical properties of the linear models, but they are better able to capture complicated relationships between variables. This should and often does result in substantial efficiency gains. Therefore, we do not only use semiparametric modeling to filter possible dependence between the auxiliary variables and random effects, but also allow for more flexibility concerning the impact of our covariates.

The Environmental Monitoring and Assessment Program of the US Environmental Pro-

tection Agency surveyed 334 lakes out of a population of 21026 in the north-eastern states of the U.S.A. between the years 1991 and 1996. Some of them were visited several times during the study period, amounting to a total of 551 measurements. For a description of the Environmental Monitoring and Assessment Program and the north-eastern lakes survey, see Whittier, Paulsen, Larsen, Peterson, Herlihy, and Kaufmann (2002).

Variable	N	mean	Std.Dev.	min	max
HUC	551	2009409.7314	1213979.5160	1010001.0000	5010002.0000
ANC	551	385.20526	547.5779	-72.2000	3371.0000
ELEV	551	321.78221	196.1327	4.0000	807.0000
LONG	551	-72.78855	2.3735	-78.9789	-67.3006
LAT	551	43.43670	1.4502	39.4508	47.1998
CO3	551	5.69828	20.2739	0.0000	203.3600
OH	551	0.62625	2.6479	0.0000	38.9000

Table 3.3: Descriptive statistics

In the data set, 113 small areas are defined by eight-digit "Hydrologic Unit Codes" (HUC) within the region of interest, including 27 that have no sample observations. HUCs divide all U.S.A. land based on the individual drainage basins according to a nested arrangement from largest (region) to smallest (unit). They are often used in surveys of natural resources as a way to delineate areas. The "acid neutralizing capacity" (ANC), also called the acid binding capacity, measures the buffering capacity of water against negative changes in pH-value (see Wetzel, 1975) and is often used, in water resource surveys, as an indicator of acidification risks in bodies of water.

As HUC boundaries follow watershed drainage areas, the lakes contained within these boundaries can be expected to present the same hydrological features, and thus HUCs make meaningful subdivisions of a region. In other words, lakes in the same HUC are expected to be more similar than two lakes in different HUCs. On the other hand, factors affecting the ANC go across HUCs, so overall spatial trends may be useful in predicting the ANC. Therefore, an HUC prediction model has the potential to capture most of the patterns in the data. We also used CO3 and OH levels of the lakes' water, which are highly correlated, see Figure 3.2. Unknown smooth impacts of its averages are used to filter out

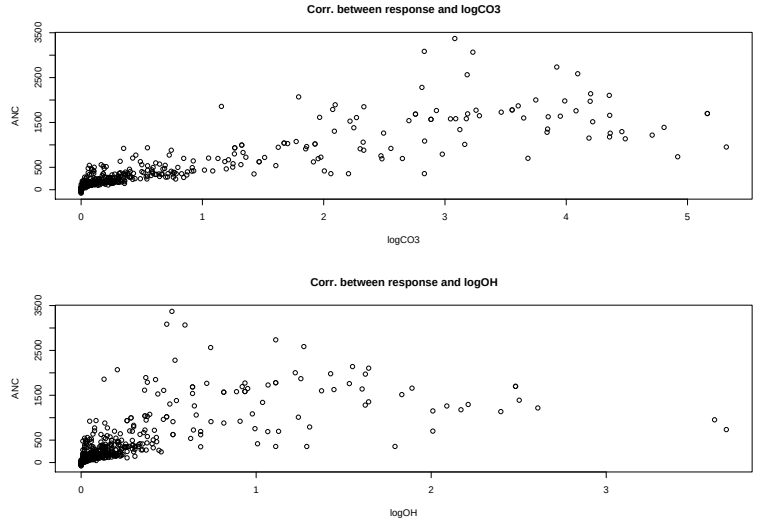

Figure 3.2: Correlation between the response and selected and logged variables.

possible dependence. As discussed, additionally we used the geographical co-ordinates of the centroid of each lake in the construction of our small area model in this environmental problem.

If these control functions, namely δ, η_1, and η_2 were jointly insignificant, then a classic mixed effects model for small area estimation and prediction would be adequate. If they were significant, in the classic model the independence assumption would definitely be violated and all further inference, estimation, prediction and calculation of the prediction mean squared error would therefore be wrong. Note that also bootstrap would fail here as it is model based, too.

We should say here that we studied former empirical results and so far used models, and also did some prior model selection studies. The model we have finally come up with is comparable for example to the studies of Breidt, Opsomer, Johnson and Ranalli (2007) as well as that of Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008), just that they all ignored the dependency problem.

The model we finally fitted was

$$Y_{id} = X_{id}\boldsymbol{\beta} + \delta(c_i) + \gamma_1(lnF_{1id}) + \gamma_2(lnF_{2id}) + \eta_1(lnW_{1d}) + \eta_2(lnW_{2d}) + u_d + \epsilon_{id}, \quad (3.20)$$

where $d = 1, .., 86$ and $i = 1, ..., 551$. The index d runs over the small areas and i runs over the elements of each area. $\boldsymbol{Y}$ is a vector of responses, the variable ANC, $\boldsymbol{X}$ is the elevation

of the lakes, $ln\boldsymbol{F}_1$ is a vector of the logged CO3 values and $ln\boldsymbol{F}_2$ is a vector of the logged OH values. $ln\boldsymbol{W}_1$ and $ln\boldsymbol{W}_2$ are the vectors of the averages of the variables logCO3 and logOH over small areas, respectively, $\overline{logCO3}_{ij}$, with $\overline{logCO3}_{ij} = (1/D)\sum_{d=1}^{D} logCO3_{ijd}$ and $\overline{logOH}_{ij}$, with $\overline{logOH}_{ij} = (1/D)\sum_{d=1}^{D} logOH_{ijd}$, $\boldsymbol{u}$ random area effects (HUCs), and the individual effects $\boldsymbol{\epsilon} \in IR^n$. γ_1,γ_2, η_1, η_2 (cubic splines) and δ (TPS) are the smooth functions and $\boldsymbol{c}_i = (c_{1i}, c_{2i})$ denotes the geographical coordinates for the observations, longitude and latitude.

We fit the model with different numbers of basis dimensions. A useful general purpose approach to choose the basis dimensions is fitting the model and extracting the deviance residuals. Afterwards, for each smooth term in the model, we fit an equivalent, single smooth to the residuals, using a substantially increased K to see if there is a pattern in the residuals which could potentially be explained by increasing K. AIC values also are a guiding factor for which model (which we use different numbers of dimension basis for the smooths) to choose. Hence, we use K=100 for the thin plate splines and K=20 for the cubic splines where the correlated logCO3 and logOH values are used, and K=10 for the averaged values over small areas of logCO3 and logOH. The knots of the cubic splines are placed evenly throughout the covariate values to which the term refers.

Fixed effects	**t**	**p-value**
Intercept	25.958	0.0000
Elevation	-2.675	0.0077
Sig. of smooths	**F**	**p-value**
δ (LONG,LAT)	1.939	< 0.001
γ_1 (logCO3)	215.418	< 0.001
γ_2 (logOH)	114.540	< 0.001
η_1 (mlogCO3)	14.774	< 0.001
η_2 (mlogOH)	13.603	< 0.001
	StdDev	
Intercept	48.561	
Residual	66.531	

Table 3.4: Estimates of coefficients (with p-values), intercept and the random effects standard deviation and the approximate significance of smooth trems.

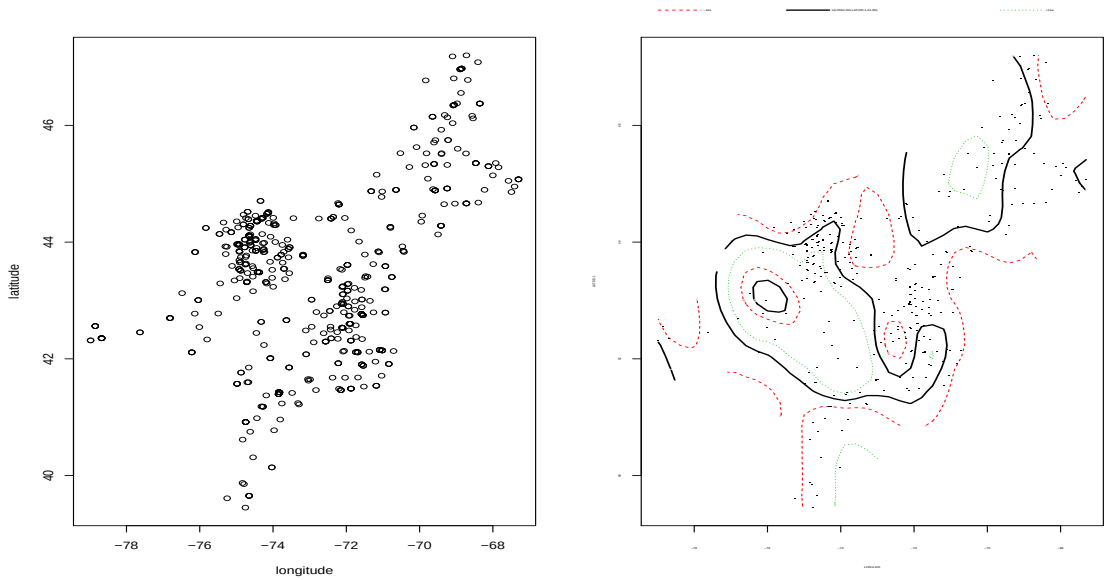

Figure 3.3: Left; plot of longitude and latitude of lakes. Right; smooth of location, K=100.

The data was analyzed using **R** (R Development Core Team, 2009) and since the linear mixed model can be viewed as a special case of the generalized linear mixed model with the identity link, the **R** package mgcv (Wood, 2011) was used. We analyzed the data by using mixed effects models. We used measures of carbon trioxide and hydroxyl levels in the lakes' water as our covariates, Hydrologic Unit Codes as random effects and our choice was the elevation of the lakes as fixed effects. In our model, we had several smooth functions and they were estimated by thin plate splines and cubic splines where we could rewrite the cubic splines in additional form. To assess the validity of the small area estimation analysis, we performed a simulation study where we estimated bias and variance of the standard deviations of the errors and the random effects. Because the first three rules of statistics are "draw a picture, draw a picture, draw a picture" (Michael Starbird), we provided figures for the thin plate splines simulations and the smooth terms estimates from the data analysis.

In the Figures 3.3 to 3.5, the solid lines and curves are the estimated effects. Gray areas indicate 95% confidence bands (Bayesian credible intervals). The bottom of each plot, the rug plot, shows the values of the covariates of each smooth. The number we see at

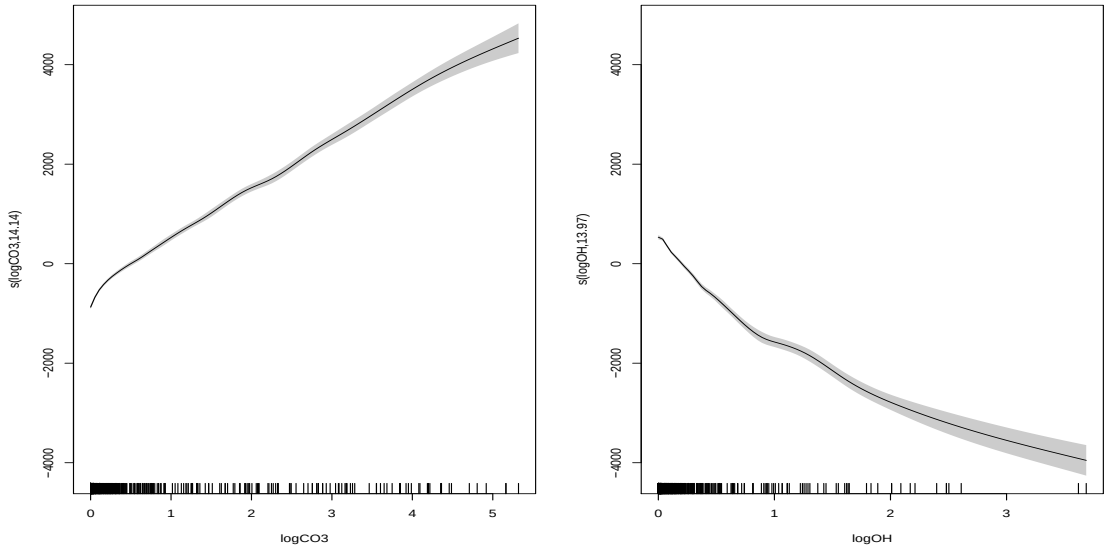

Figure 3.4: Left; smooth of logCO3, K=20. Right; smooth of logOH, K=20.

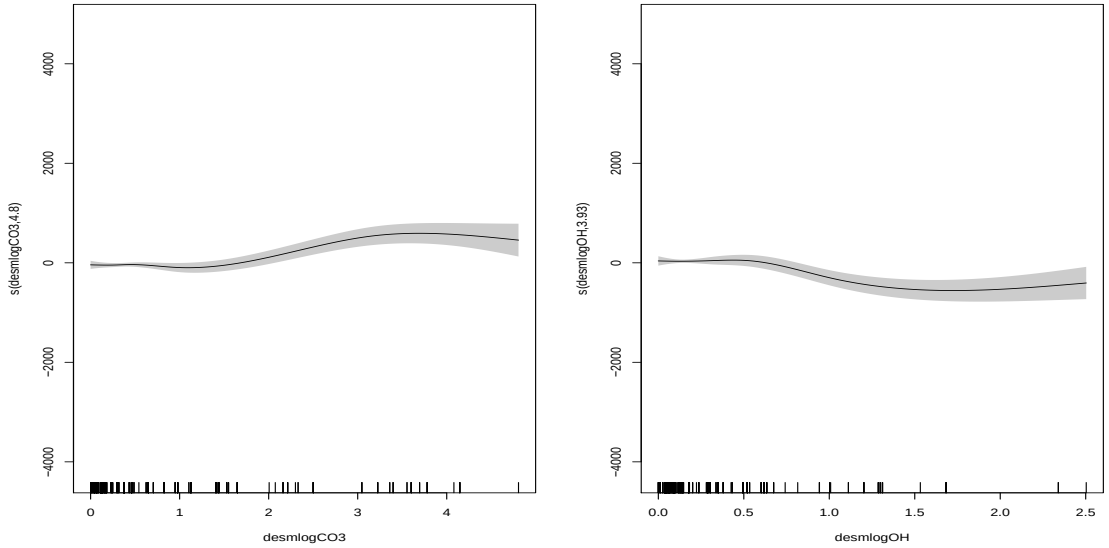

Figure 3.5: Left; smooth of mean logCO3, K=10. Right; smooth of mean logOH, K=10.

the y-axis caption represents the effective degrees of freedom of the term being plotted. Certainly, for the function estimate of δ we have no confidence bands but can see its

significance from the R print out.

What we find is firstly, the control functions seem to be all significantly different from zero and, secondly, all functions are clearly non-linear except maybe γ_1. We can therefore see that: (a) location matters even after having controlled for the other variables like elevation, and its impact is not easy to be captured; (b) as Breidt, Opsomer, Johnson and Ranalli (2007) pointed out, simple linear mixed effects models are often not flexible enough to reflect correctly complex relationships such as those in our environmental problem; (c) the crucial and always applied independence assumption is typically problematic and in our case clearly violated. Control functions are necessary to filter the possible dependence between covariates and area effects as otherwise all small area inference would be invalid.

4 From the log of gravity toward a semi-mixed effects gravity model for intra-trade in domestic markets

4.1 Abstract

The common procedure with which to estimate Gravity equations was based on the OLS of the transformed log-linear specification. For panel data, several fixed effects were included to control for country unobserved heterogeneity. In recent years, both the classic gravity model and the estimation method have been criticized, and different alternatives have been suggested. While a controversial discussion has been going on about the classic and newly proposed approaches, the empirical research on trade has mainly ignored this, still applying the popular log-linear model combined with OLS estimation. In this work, first some of the main criticism and alternatives considered so far are revisited. Then, the ongoing discussion is extended to panel data analysis before an original new model and estimation method is proposed. This new proposal tries to reconcile the existing ones and can easily be applied with several of the standard software packages. This model is especially designed for studying intra-trade of domestic markets and integration areas. It is used to study the intra-trade of the European Union before the recent economic crisis of 2007.

This chapter is a joint work with Proença, I. and Sperlich, S..

4.2 The classic gravity model and some criticisms in brief

To analyze trade flows, the gravity model has been experiencing a revival, in particular thanks to the new economic geography. It was mainly due to the work of Tinbergen (1962) that this modeling approach influenced strongly the research on trade (especially the empirical one) though it has a much broader field of applications in economics. Niedercorn and Bechdolt (1969) were probably the first in trying to give an economic derivation of this modeling idea, followed by Anderson (1979) who concentrated on trade and started out from consumer expenditure system theory. More recent theoretical contributions in this direction are provided for example by Deardorff (1998) who proved the consistency of the gravity model with the Heckscher-Ohlin trade theory, and Anderson and van Wincoop (2003), who basically revisited Anderson's model but extending it slightly to deal with the border puzzle. For more contributions see also references herein.

These articles by nature look at the economic derivations to clarify the basic economic model specification and consequences for comparative statics analysis. At the same time there was a lot of empirical and (applied) econometric research going on trying to find model specifications, relevant covariates and estimation procedures which fit the data well to a seemingly reasonable model. The work of Bergstrand (1985) marked an important contribution to bridge economic theory and common empirical practices, again based on consumer expenditure and general equilibrium theory. But still, this discussion of correct econometric specification and estimation can differ a lot from the aforementioned; to see this, compare his work with Mátyás (1997).

Quite recently, Santos Silva and Tenreyro (2006) have pointed out a problem that has been ignored so far by both of the above mentioned communities. They make the point that for basically all trade models logarithmic transformations (i.e. typically log-log-linear models) are used. However, if the error terms are not independent from the regressors, ordinary least square estimation (henceforth OLS) allows for consistent estimation of the parameters of that log-model but not of the model of interest. They argue that these trade-models typically exhibit heteroscedasticity, i.e. the lack of independence. It is clear that neither the use of generalized least square estimation (GLS) nor the introduction of some fixed effects, etc., can resolve this problem. Even though this is true for any log-transformed model, they argue that the consequences are particularly severe in the context of gravity modeling for trade analysis. As an alternative they proposed for example to

apply the Poisson pseudo maximum likelihood (henceforth PPML) to estimate the original model without the use of a log-transformation. Other authors like Martinez-Zarzoso (2011) criticized that approach and argued, based on intensive simulation studies, in favor of OLS and feasible GLS combined with the log-transformation. Santos Silva and Tenreyro (2008) replied (to an early discussion paper version, therefore the earlier date) they wanted to rise mainly the consistency problem, not the question of efficiency performance. It is clear that the PPML has not the same problems with zero responses as a log-transformation has, but the problems are similar if we face a zero-inflation. One of the further problems is the inefficiency due to possible overdispersion which makes the feasible GLS outperforming PPML. Consequently, one could argue that just like economists have long been aware of the Jensen's inequality (to pick up the argument of Santos Silva and Tenreyro), statisticians have always been aware of the fact that the mean squared error of an estimator is more important than unbiasedness, we are not sure if the econometricians are aware of this. We will therefore propose an extension that on the one hand will respect the interest of a correct identification of the parameters of interest and on the other hand the legitimate request for a more efficient estimation performance.

Panel data have become widely used as means to control for unobserved cross-section heterogeneity. Applications with the gravity model to explain panel bilateral country trade flows are an example. In that context, Mátyás (1997) was one of the first authors to call attention to control for unobserved heterogeneity specific to importer country, exporter country and time, that may lead to endogeneity, with fixed effects. Another approach, which is more in use nowadays, instead considers fixed country-pair effects in order to prevent inconsistency due to omitted time invariant determinants specific to the bilateral relation such as common language, common border or a common relevant ethnic group capturing immigrants' links to their own country. Anderson and van Wincoop (2003) give different reasons for including either country specific or country-pair specific effects. Moreover, Baltag, Egger and Pfaffermayr (2003) show the importance of controlling for all interactions based on importer, exporter and time effects. Their model is more general, encompassing both the specific importer or exporter and the country-pair fixed effects models.

One of the drawbacks of these fixed effects approaches is that they either tend to overfit or they require the application of differencing estimators which might be appropriate for linear models (i.e. models which are linear with respect to the parameters) but less for nonlinear

ones, neither they allow for the identification of the impact of time invariant variables on trade. This problem is addressed by Serlenga and Shin (2007) in the estimation of a gravity model to explain bilateral trade among 15 EU countries for a panel data large in time. They control for unobserved heterogeneity using a general multifactor error structure following Pesaran (2006). To account for dependency between unobserved heterogeneity and some explanatories they combine the estimation procedure of Pesaran (2006) with Hausman and Taylor instrumental variables in a way that enables them to estimate the coefficients of time invariant variables. This last procedure needs a minimum of exogenous variables varying in time and a minimum of exogenous variables constant in time depending on the number of endogenous variables in each category. It is only applicable for panels with a large number of observations in time. Moreover, it assumes that one is provided with adequate instruments which is not at all evident in practice – but both weak instruments or the lack of that particular exogeneity lead to a failure of identification. Different to the above mentioned contribution, Westerlund and Wilhelmsson (2009) proposed a Poisson fixed effects estimator (with robust variances) for a panel data analysis to study the trade effects of the 1995 European Union enlargement. Their argumentation follows partly Santos Silva and Tenereyro (2005) but they give a stronger emphasis on the problem of zero responses. Therefore, it is surprising that they do not consider the problems of zero-inflation and overdispersion. They certainly do not identify other time invariant impacts than the fixed effects.

Now it is well known that a natural extension of a Poisson modeling is the introduction of subject specific random effects which automatically will capture the overdispersion. If the mean function is exponential with multiplicative random effects which are supposed to follow a gamma distribution, then the resulting likelihood is a (pseudo) negative binomial one. This is probably the most popular extension of the Poisson (pseudo) likelihood modeling. In mixed effects literature and also in different software packages, it is preferred to consider normal distributed additive random effects inside the exponential mean function, for details see next sections. To add then a zero inflation to a Poisson, both conditioned on the same explanatories, is quite straight and standard in applied statistics but less often needed if random effects take care of possible overdispersion. Notice that we have a rather flexible likelihood with subject specific heterogeneity but allowing for the identification of all parameters of interest, including those that are time invariant. It helps us further to overcome possible zero inflation and / or serious efficiency losses.

The obvious problem that occurs now is to prevent misspecification due to the independence assumption for the random effects. But in the context of small area statistics, Lombardía and Sperlich (2011) introduced a new class of semi-mixed effects models. In terms of panel econometrics, one could say that they extended the Mundlak device for random effects models in several ways. They actually included a nonparametric component in the equation that captures all country unobserved heterogeneity correlated with the explanatories without compromising the estimation of the effect of time invariant variables nor the estimation of the untransformed nonlinear gravity equation. The remaining heterogeneity is still modeled by independent random effects. For our purpose this is important for the allowance of overdispersion.

To summarize, in this paper we introduce a semiparametric gravity model for panel data. One of the main ideas is to add a nonparametric term in the exponential conditional mean function which depends on observable proxy variables in order to filter possible dependency between some explanatories and the unobserved individual heterogeneity (constant over time). Time fixed effects and interaction terms can be included without difficulty. The proposed model can be estimated with different programs provided in standard software packages like e.g. **R** and `Stata`. As far as we know, there does not exist a similar semiparametric specification for panel count data models in the literature. Most of the semiparametric approaches to panel count data are based on random effects, that is, assuming that unobserved individual heterogeneity is distribution free and not correlated with explanatories. Examples are, among others, Gurmu, Rilstoneb and Sternc (1999) and Zheng (2008) who recurs to Bayesian techniques. A distinct approach is given by Wellner and Zhang (2007) who consider a semiparametric count panel data model by introducing a multiplicative term in the exponential conditional mean and equal to a monotonously increasing (but unknown) function that depends on the time period. Note that does by no means fit our problem. Racine and Li (2004) introduce a nonparametric single index model, estimated with kernels, and apply it to panel data reporting the number of successful patent applications.

In what concerns nonlinear specifications of the gravity model for panel data, Henderson and Millimet (2008) compare the usual parametric exponential specification of the gravity model with a nonparametric alternative. The parametric model is estimated using PPML with fixed importer country effects and fixed exporter country effects. They hypothesize that the nonparametric approach might be more valuable for depicting im-

portant heterogeneity on the effect of some explanatories. However, the authors found that their specification combined with PPML was superior on predicting trade flows (both in- and out-of-sample) compared to their nonparametric estimates and predictions. For this reason one might want to limit the use of nonparametrics to filter possible dependence between exploratories and subject specific random effects. There is nevertheless an issue in the model specification that makes additive nonparametric modeling still attractive, at least for explorative purposes. We will see in the next section that all the gravity model specifications and estimation methods work with logarithmic inputs. Unfortunately, for a logarithmic transformation the chosen units can easily make an important difference for the coefficient estimates.

The rest of the paper is organized as follows. In the next section the model is introduced and the problems of the classic estimation methods, pointed out by Santos Silva and Tenereyro (2005), are briefly revised. We then develop a new model and estimator in order to reconcile their suggestions with criticism e.g. of Martinez-Zarzoso (2011), and to give an extension to panel data analysis with time invariant regressors. Section 4.4 is dedicated to the study of the trade flows among the 25 members of the European Union after the big expansion in 2004 toward the East, when the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia, Slovenia, and Cyprus became members. To not distort the results by the financial crises, we concentrate on the period of 2004 to 2007, i.e. a four years panel. Section 4.5 concludes.

4.3 A semi-mixed effects gravity model for panel data

The specification of our gravity model follows the derivations of the above discussed literature. The bilateral trade T_{ijt}, say export from country i to country j at time (year) t, is assumed to be determined by the GDPs y_{it}, y_{jt} of countries i and j, multilateral trade barriers and trade cost factors which might be represented by consumer price indexes (like in Anderson and van Wincoop (2003), pair specific information like contiguity and distance, common language and /or ethnic groups as well as country specific information like landlockedness and population size. Alternatively to GDP and population one can find model specifications via GDP per capita. In panel specifications one might also include time fixed effects.

Let us for a moment classify the independent variables into the following groups: the GDP, binary information D_{ijt}, non-binary time invariant information z_{ij}, and the remaining set x_{ijt}. Then, typically for $i, j = 1, ..., N,\ t = 1, ..., T$ the following panel gravity model is considered

$$T_{ijt} = \exp[\ln(x_{ijt})\beta + \ln(z_{ij})\gamma + D_{ijt}\delta + \alpha_t + \eta_{ij}]\ y_{it}^{\beta_{yi}}\ y_{jt}^{\beta_{yj}} + \epsilon_{ijt} \quad (4.1)$$

$$= \exp[\ln y_{it}\beta_{yi} + \ln y_{jt}\beta_{yj} + \ln(x_{ijt})\beta + \ln(z_{ij})\gamma + D_{ijt}\delta + \alpha_t + \eta_{ij}]v_{ijt}, \quad (4.2)$$

where we included time fixed effects α_t, unexplained heterogeneity η_{ij}, and error terms v_{ijt} and ϵ_{ijt} respectively. Here ϵ_{ijt} is a zero-mean random variable, and $v_{ijt} = 1 + \epsilon_{ijt}/exp[...]$ consequently a heteroscedastic though multiplicative disturbance term with a conditional mean equal to one. Further, x_{ijt}, z_{ij} and D_{ijt} are row vectors, β, γ and δ column vectors of unknown coefficients of corresponding size, and β_{yi}, β_{yj} unknown scalar coefficients. When panels are large in time, then there are alternative specifications for the time effect that may be preferable to the fixed effects and can be easily accommodate in the framework that will be introduced in the remainder of this paper, see for instance Pesaran (2006).

While Santos Silva and Tenreyro (2006) criticized the inconsistency of OLS or GLS estimators of

$$\ln T_{ijt} = \ln y_{it}\beta_{yi} + \ln y_{jt}\beta_{yj} + \ln(x_{ijt})\beta + \ln(z_{ij})\gamma + D_{ijt}\delta + \alpha_t + \eta_{ij} + \ln v_{ijt}, \quad (4.3)$$

due to the fact that the conditional expectation of $\ln v_{ijt}$ is a function of the independent variables and η_{ij} (if a random effect), Westerlund and Wilhelmsson (2009) emphasized the distortion caused by zero trade observations. Both proposed PPML to overcome these problems. In contrast, Martinez-Zarzoso (2011) argued that nevertheless OLS and GLS showed better performance in practice, and several authors discussed ways to incorporate zero-responses differently. Henderson and Millimet (2008) found that the nonparametric alternatives they studied could not outperform the above parametric models, neither in in-sample nor in out-of-sample prediction. Recall further the above discussed problems that occur if the η_{ij} are modeled as fixed effects.

Based on these observations, we propose to estimate equation (4.1) with the aid of a mixed effects PPML, where the η_{ij} are random effects. The well known problem is that that if this unknown heterogeneity is related with the included explanatory variables, then also this estimator is inconsistent. The best known possible remedy is probably the Mundlak (1978) device. He proposed to include the temporal means of the explanatories linearly in model

(4.1) wherever possible (i.e. in our case basically of the $\ln x_{ijt}$ as most of the dummies are time invariant, too). Although this proposal can often be found in the literature – see also the much less practical but better motivated one of Chamberlain (1984) – it has never been accepted in applied econometrics as a real way out of the dependency dilemma of random (respectively mixed) effects models. Note further that for panels short in time, the inclusion of temporal means runs the estimation numerically less stable and inherits complex covariance structures with large variances for the coefficient estimates.

In the context of small area statistics, Lombardía and Sperlich (2011) introduce a semi-parametric filter to get rid of the possible dependency between this random heterogeneity and included explanatory variables. Following their idea, we claim that for a set of time-invariant but else continuous variables w_{ij} there exists an unknown function $\psi(.)$ such that $E(\eta_{ij}|\psi(w_{ij}), x_{ijt}, z_{ij}, D_{ijt}) = 0$ or in other words that $\eta_{ij} = \psi(w_{ij}) + u_{ij}$ with u_{ij} being an unobserved random effect, uncorrelated with x_{ijt}, z_{ij}, D_{ijt} and w_{ij}. If we additionally assume that u_{ij} is independent of ϵ_{ijt}, then model (4.1) becomes

$$T_{ijt} \quad = \quad \exp[\ln(x_{ijt})\beta + \ln(z_{ij})\gamma + D_{ijt}\delta + \alpha_t + \psi(w_{ij}) + u_{ij} + \epsilon_{ijt}]. \tag{4.4}$$

The existence of such a filter or function $\psi(\cdot)$ is not a mystery at all. As w_{ij} is continuous we can imagine the extreme case where ψ simply does a kind of interpolation such that it corresponds in fact to fixed effects and we get $\eta_{ij} \equiv 0$. Another extreme case it that there is actually no dependency between the included explanatory variables and η_{ij}; then we can set $\psi \equiv 0$. In most cases, however, ψ will be something in between and be estimated accordingly.

The variables w_{ij} can be considered as proxies for the relation of x_{ijt} and z_{ij} with η_{ij}. Certainly, their availability will depend on the particular problem. For some applications one may have a clear idea of the causes of the dependency between explanatory variables and the individual unobserved heterogeneity term. For instance, if we modeled wages the last is due in good part to individual ability, and we would look for corresponding proxies like the IQ. If we are not sure and lacking additional information, one can follow Mundlak's suggestion of taking the temporal means of the time varying explanatories for w_{ij} but then has to be very careful with the coefficients' correct interpretation, see discussion above. Note, however, that our proposal is much more flexible and general, given the fact that we allow these variables to enter nonparametrically. If one is not particularly interested in estimating right the coefficient γ then the respective time invariant z_{ijl} can be part of

w_{ij}.

To avoid smoothing problems, the curse of dimensionality, and to provide the chance to interpret ψ (the impact of w_{ij}), we will consider $\psi(w_{ij})$ as an additively separable function. Furthermore, we would like to get confidence intervals around the estimated additive functionals. Finally, recall that we said that ψ has to be estimated accordingly to really act as a filter. When using smoothers for ψ, this problem basically boils down to the question of smoothing parameter selection along a well defined objective function. Note that Lombardía and Sperlich (2011) considered only a multidimensional kernel estimate combined with cross validation or a modified Hausman test for the exogeneity of explanatories. All the points we call for have recently been solved in the P-spline literature on mixed effects model estimation; see, for example, Wood (2006). Programs which can handle this estimation of equation (4.4) with PPML are provided in **R** and `Stata`. In the moment of estimation, we face at least one remaining problem which is due to the nonparametric nature of ψ on the one hand, and the wanted correlation of regressors and w_{ij} on the other hand. This can easily lead to multifunctionality (the analogue to multicollinearity). There are basically two ways to handle this: variable selection as is usually done in complex high dimensional regression problems, and the restriction of flexibility of ψ, e.g. by limiting the impact of the Mundlak device variables to linearly.

It should be added that an extension of our (else parametric) model with nonparametric filter to the more complex semi- and nonparametric world is straight forward, and can even be performed with the same software. Henderson and Millimet (2008) argue that the added value of such an extension is quite poor if not questionable.

In the next section, we use our semi-random effects gravity model (4.4) to analyze the trade flows among the EU25 countries from 2004 till 2007. Recall that this was the period following the big extension to Eastern Europe until the financial crises in the Western world. The dependent variable of interest will be the import flows, given that countries often tend to monitor their imports more carefully than their exports.

4.4 Trade flows insider the European Union after the big Eastward Enlargement

4.4.1 Data and model

As said above, we will consider the import flows as our dependent variable. With regard to the independent time-variant variables, we use as proxies for the overall economic mass of the countries in the bilateral trade flow the GDPs (*MGDP* and *XGDP* for importer and exporter countries respectively), and the populations (*MPOP* and *XPOP* for importer and exporter countries respectively). We expect a positive effect for *MGDP* and *XGDP* while for *MPOP* and *XPOP* the literature documents an ambiguous effect. For multilateral trade barriers we follow Anderson and van Wincoop (2003) including the consumer price indexes *MCPI* and *XCPI* of the importing and exporting countries. We consider also a measure of the difference in terms of factor endowment of both trading countries, *RFE*, given by the absolute value of the difference between their *per capita* GDPs, and a measure of the degree of similarity in bilateral size, *SIM*, an index that is bounded between 0 (absolute divergence in size) and 0.5 (equal country size). The higher is *RFE*, the larger is the difference between each country factor endowment resulting in a lower intensity of intra-industry trade but a higher volume of inter-industry trade. Consequently, the total effect has an ambiguous sign. Serlenga and Shin (2007) found positive signs for the effect of these variables in explaining total trade while Baltagi, Egger and Pfaffermayr (2003) found a negative sign for RFE and a positive one for SIM when modeling exports. All these regressors enter our model in logarithmic terms.

In what concerns time-invariant variables, we use the distance (logarithm of kms between the capitals) of the two trading partners, *DIST*, as a proxy for the effect of transportation costs. Therefore, we expect this variable to have a negative effect on trade. We consider also some time-invariant dummy variables like *NEIGH* which is equal to one if both trading countries share a land or sea border, *COMLANG* which is equal to one if both trading partners share the same official language, *ETHNIC* which is equal to one if in the importer country there is an ethnic minority of the exporter country, *EU15* which is equal to one if both countries belong to the European 15, *MLOCK / XLOCK* which is equal to one if the importer / exporter country has no direct connection to the sea (i.e. are landlocked), and

the dummy *GERMAN* which is equal to one if one of the countries involved in the trade flow is Germany. We admit that this is not really an 'explaining' variable but as it has been shown in many different empirical works, skipping this dummy from the regression increases significantly the remaining unobserved heterogeneity in an asymmetric way. We expect that all these dummy variables have a positive effect on trade except for *MLOCK* and *XLOCK*. We also introduced time dummies as well, denoted by *D05*, *D06* and *D07* for the years 2005, 2006 and 2007.

For the proxies, say W_{ij}, that shall filter possible dependence between the remaining heterogeneity η_{ij} and the regressors enumerated above, we consider the area of importer and exporter country, respectively *MAREA* and *XAREA* for size. To control for the technological development (in absolute and relative terms) we take the yearly average of the number of patents from 2003 to 2005 and its per capita counterpart, denoted by *NPAT* and *NPATpc*, respectively. All these four variables were included in logarithmic terms.

Table 4.1 reports some descriptive statistics about all of the used variables. Further details on the way the variables were obtained can be found in the Appendix. It should be mentioned that GDP and population are correlated by more than 80% for importing and exporting countries, the consumer price indexes MCPI and XCPI exhibit a correlation of about 70% among themselves and one of 6 to above 9% (in absolute values) with population and GDP, RFE and SIM a negative correlation of about -27%. Moreover, while RFE shows a correlation of 10 to 20% with 5 regressors, SIM has mostly a pretty low correlation with other regressors (varying from zero to five percent). Finally, being landlocked is to 30% correlated with population size and to 23% with GDP. The AREA variables are to almost 80% correlated with the corresponding population sizes and to 60% with the corresponding GDPs.

The estimation of all our models have been performed by the use of the `gamm` procedure of the software package **R**. For all models for which alternative estimation procedures existed, no matter whether in **R** or `Stata`, we double checked our results. Where numerical problems occurred due to multicollinearity or multifunctionality, the **R** routines turned out to be numerically more robust. In general, the **R** procedures were mostly faster so that all results presented in the following refer to the output obtained in **R**. Where offered by the routines, we tried different specifications of the distribution of our random effects

Variable	Obs	Mean	Std. Dev.	Min	Max
M	2400	37.5504	95.4791	0.0013	920.7711
MGDP	2400	12.0023	1.5235	9.1749	14.7005
XGDP	2400	12.0031	1.5242	9.1749	14.7005
MPOP	2400	16.0941	1.2717	13.0280	18.2287
XPOP	2400	16.0884	1.2706	13.0280	18.2287
MCPI	2400	4.6125	0.0408	4.5183	4.7649
XCPI	2400	4.6125	0.0408	4.5183	4.7649
RFE	2400	0.9767	0.7696	0.0045	7.9311
SIM	2400	-1.5861	0.9152	-4.7496	-0.6931
DIST	2400	7.0065	0.6413	4.0431	8.1194
NEIGH	2400	0.1333	0.3400	0	1
COMLANG	2400	0.0300	0.1706	0	1
ETHNIC	2400	0.0167	0.1280	0	1
EU15	2400	0.3500	0.4771	0	1
GERMAN	2400	0.0808	0.2726	0	1
MLOCK	2400	0.2000	0.4001	0	1
XLOCK	2400	0.2000	0.4001	0	1
MAREA	2400	11.5227	1.1928	7.8579	13.2123
XAREA	2400	11.5227	1.1928	7.8579	13.2123
NPAT	2400	5.6722	2.2519	2.1511	10.0200
NPATpc	2400	3.3400	1.7927	-0.2066	5.5994

Table 4.1: Descriptive statistics

u_{ij}. Typically the outcome did not vary significantly along the different specifications (namely gamma for the multiplicative version, and normal for the additive one). All results presented in this article refer to the normal specification.

4.4.2 Summary of main estimation results

We start with the estimation of parametric panel data models following the PPML approach. A first trial showed that variables RFE should be skipped but SIM be kept, compare also with discussion above. For the pooled regression (model 1) the estimates of

the elasticities of importer GDP, exporter GDP and distance have size smaller (in absolute value) than the usual estimates obtained with the loglinear model, though this is a tendency depicted also in Santos Silva and Tenreyro (2006) (for different bilateral partners than those considered here) and Proença et al (2008) (for the same bilateral partners addressed here), both using the PPML for cross section data. The impact of importer population is not relevant while we find a positive significant effect for the population size of the exporting country. As expected, the CPI of the importer is positive whereas the one of the exporter is negative. If countries share borders, if Germany is involved, or if both trading partners belong to the EU15, we have a significant positive effect. Landlockedness of the importing country, and distance between partners have a significant negative impact. All included time fixed effects are significant and increasingly positive (over time).

In a second step we try the parametric Mundlak device. We denote the temporal means of our regressors by simply putting an "m" in front of the original name of the variable. If we compare the two models of which the estimates are given in Table 4.2, we see quite important changes from model 1 to model 2: the impact of XGDP decreases by almost 60% whereas the impact of MGP goes up by more than 10%. XPOP changes its sign, the consumer price index variation becomes insignificant like EU15 and almost GERMAN. The impact of NEIGH cuts almost to the half, SIM changes the sign and becomes clearly independent. Note also that the estimated σ_u is almost ten percent smaller in model 2 than in model 1. Certainly, the problem is that the interpretation of β does indeed change importantly from model 1 to model 2. However, what we have in model 2 is mainly a projection of the time invariant heterogeneity on the temporal means.

It would therefore be preferable to filter possible dependence between the pair effects η_{ij} (see equation (4.3)) and our regressors $\ln x_{ijt}$ by means of our proxies W_{ij} and function ψ, compare model (4.4). This has first been done by only including the area and patent variables but without (further) temporal means and is denoted as model 3.1. further , the alternatives are to include the temporal means of the population sizes, let it be nonparametrically (model 3.2) or parametrically (model 3.3), and the temporal means of the GDPs (model 3.4) respectively. The estimation results for model 3.1 and 3.2 are given in Table 4.2, and those for models 3.3 and 3.4 in Table 4.4 of the Appendix. In general we can say that these models seem to be located in-between models 1 and 2 (except model 3.4), which is in accordance with our intuition. The remaining pair effect u_{ij}, which is supposed to be random, shows a quite similar standard deviation for all models of this

class. Models 3.2 and 3.3 are basically the same what simply means that the impact of the temporal means of the population size is log-linear.

What is evident is that each time we include the temporal mean of one of our regressors of interest, its β coefficient changes drastically. Compare especially the coefficients of XPOP of model 3.1 versus models 3.2, 3.3 and the coefficients of the GDPs of model 3.1 versus model 3.4. Also the coefficients of the CPIs change a lot due to its high absolute correlations of more than 8% with mXPOP and of more than 15% with mXGDP. What is not shown here is that the model changes a lot if the temporal means of the GDPs enter nonparametrically. This indicates that we may face rather a problem of functional misspecification than of omitted variables. The functional forms of our additive ψ are given in Figure 4.1 together with 95% confidence intervals. The shown graphs refer to model 3.1. The corresponding estimates of $\psi(W_{ij})$ for the other models are given in the Appendix. The nonparametric impacts of the temporal means of population sizes, see model 3.2, are given in Figure 4.2.

One might want to be on the safe side (concerning possible exogeneity) and include always all temporal means, but Chamberlain (1982) showed already for simpler models that this can easily fail to work - without discussing the details of proper interpretation. We can include temporal means only for our $\ln x_{ijt}$ but not for the other explanatories. Furthermore, for our data, larger models than those presented in this articles turned out to be overidentified facing problems of multicollinearity and multifunctionality. Finally, a high significance of the temporal means may simply reflect the inappropriateness of the log-linear specification or the choice of units before doing the log-transform, recall also our above made comments.

All our studies made so far indicate that the impact of GDP is likely to be more complex than log-linear of needs a different choice of units before the log-transform is done. The easiest way to relax our model (4.4) is to replace $\beta_1 \ln MGDP$ and $\beta_2 \ln XGPD$ by $\beta_1(\ln MGDP)$ and $\beta_2(\ln XGPD)$, where now β_1 and β_2 are nonparametric functions. Note that we alternatively tried to include the squared and cubic terms of log-GDP but unfortunately without similar success. We denote our new model 4.1 in reference to model 3.1 because only the functional shape of the impact of GDP has changed. The similarities to model 3.1 and differences to model 1 or model 2 are as expected. The standard deviation

Model:	**1**	**2**	**3.1**	**3.2**	**4.1**
Const	-5.3178 .0001	-70.2684 .0691	-7.1251 .0000	-0.61577 .6640	9.7555 .0000
MGDP	0.7183 .0000	0.7961 .0000	0.7900 .0000	0.77226 .0000	n.p.e
XGDP	0.5008 .0000	0.2139 .0000	0.4192 .0000	0.29200 .0000	n.p.e
MPOP	0.0041 .2917	0.0003 .9307	0.0023 .5485	0.00022 .9537	0.0028 .4300
XPOP	0.2344 .0000	-0.0890 .0499	0.2810 .0000	-0.09098 .0455	0.2627 .0000
MCPI	0.3438 .0514	-0.0126 .9509	0.1510 .4623	0.14191 .4789	0.2336 .2871
XCPI	-0.9498 .0000	-0.2398 .2118	-0.6499 .0007	-0.33542 .0763	-1.1525 .0000
SIM	-0.0509 .0771	0.0651 .2580	-0.0631 .0260	-0.01382 .6478	-0.1352 .0000
DIST	-1.2632 .0000	-1.3165 .0000	-1.1971 .0000	-1.19854 .0000	-1.2422 .0000
NEIGH	0.4054 .0013	0.2459 .0351	0.2773 .0076	0.25257 .0112	0.3279 .0017
COMLANG	-0.2609 .1817	-0.0443 .8046	0.1004 .5296	0.09505 .5353	0.0255 .8736
ETHNIC	0.1206 .6185	0.2227 .3157	0.0158 .9359	0.03128 .8683	-0.0679 .7304
EU15	0.4317 .0000	0.1351 .2371	0.8220 .0000	0.70199 .0000	0.7954 .0000
MLOCK	-0.3581 .0000	-0.3313 .0000	0.0673 .5352	-0.01673 .8679	0.0133 .9006
XLOCK	0.0243 .7659	0.0271 .7211	0.1136 .2618	0.08796 .4337	0.1676 .0872
GERMAN	0.1805 .0213	0.1350 .0679	0.0715 .4009	-0.02853 .7566	0.1587 .0444
2005	0.0256 .0004	0.0234 .0010	0.0229 .0015	0.02183 .0019	0.0485 .0000
2006	0.0818 .0000	0.0847 .0000	0.0775 .0000	0.07985 .0000	0.1368 .0000
2007	0.0998 .0000	0.1066 .0000	0.0938 .0000	0.09870 .0000	0.1868 .0000
mMGDP		-0.1368 .1379			
mXGDP		0.6149 .0000			
mMPOP		0.2329 .0005		n.p.e	
mXPOP		0.0978 .2157		n.p.e	
mMCPI		13.0632 .0169			
mXCPI		0.0719 .9897			
mSIM		0.0090 .8951			
W_{add}			n.p.e	n.p.e	n.p.e
σ_u	0.708668	0.640852	0.559236	0.531391	0.562986

Table 4.2: Estimates of coefficients (with p-values), intercept and the random effects standard deviation for different model specifications. *n.p.* stands for nonparametric estimates. The latter are shown in the corresponding figures: for model 3.1. the estimated impact of the additional W-instruments MAREA, XAREA, NPAT, and NPATpc are plotted in Figure 4.1; for models 3.3 and 4.1 see Appendix.

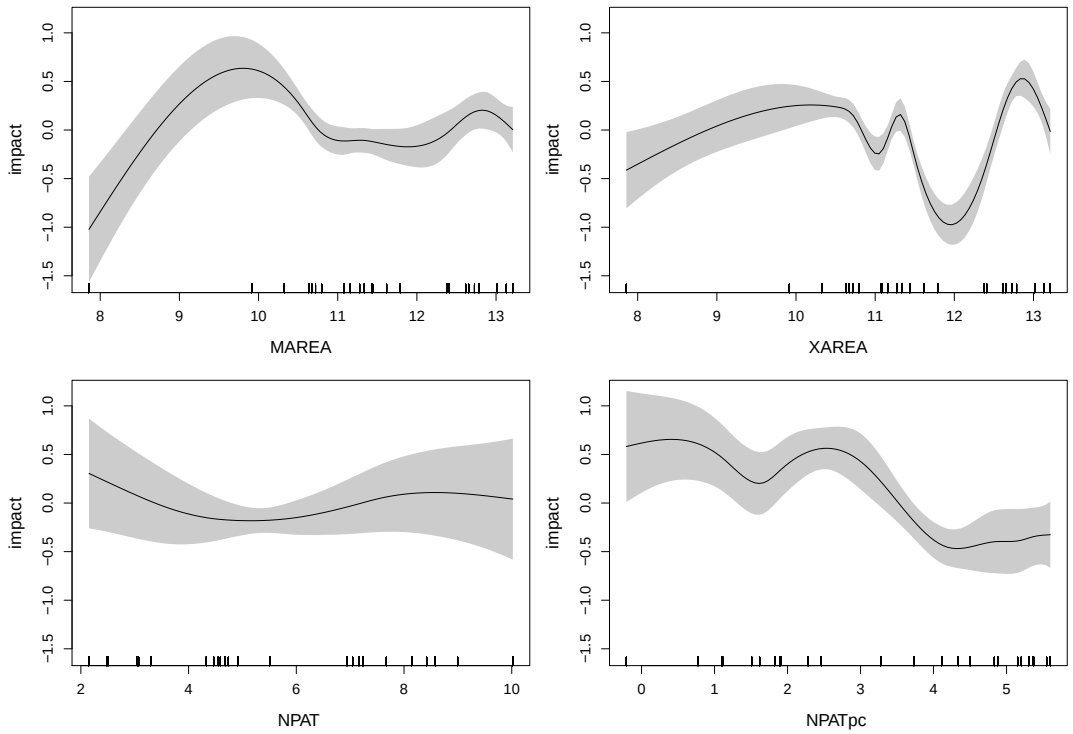

Figure 4.1: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 3.1 with 95% confidence intervals.

of the random effect has hardly changed compared to the class of our models 3.1 to 3.4. The next steps would be again to see what changes if we include additionally temporal means (see model 4.2 and 4.3 in the Appendix) or finally to try an almost purely non-parametric, say generalized additive model, compare model 5 in the Appendix. But before we start to further compare coefficients and discuss reasons for significant changes, it is probably more reasonable to look for a model selection criteria and see whether we really gain something with all these extensions. Table 4.3 provides such a comparison based on the Akaike, the Bayesian Information Criterion, and the Log-likelihood. We see that along these criteria model 4.3 seems to be the best but in our opinion is little helpful for reasonable economic interpretation. We should also be careful with looking at the absolute numbers. In fact, compared to model 4.1, only model 4.3 can improve more than 10% in all criteria. However, between model 3.1 and 4.1 we have again a loss / gain of another 10%. Respective the question of endogeneity, a Hausman type test is not available for our

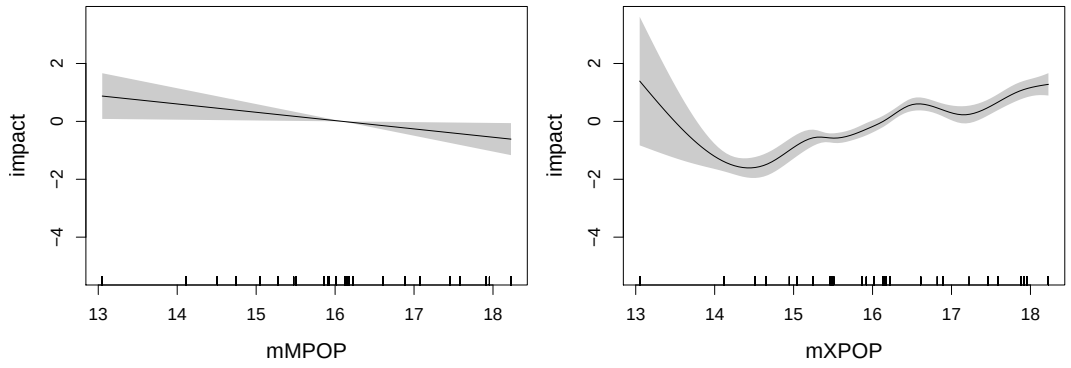

Figure 4.2: Estimates of $\psi_5(mMPOP)$ and $\psi_6(mXPOP)$ for model 3.2 with 95% confidence intervals.

models in any software. We programmed one based on (wild) bootstrap but numerically it is not very stable and therefore just not reliable. The typical but not surprising message is, however, that a simple random (or mixed) effects model suffers from endogeneity whereas nonparametric filters turn the test statistic insignificant. It does – and maybe should – not further serve or help for model selection.

Model:	**1**	**2**	**3.1**	**3.2**	**4.1**
AIC	1603.479	1388.531	1421.918	1306.513	1144.407
BIC	1724.927	1550.461	1589.631	1497.359	1323.687
LL	-780.740	-666.265	-681.959	-620.257	-541.204
Model:	**3.3**	**3.4**	**4.2**	**5**	**4.3**
AIC	1311.954	1301.396	1028.780	1079.962	975.020
BIC	1491.234	1480.676	1219.626	1293.941	1194.782
LL	-624.977	-619.698	-481.390	-502.981	-449.510

Table 4.3: Different model selection criteria for all estimated model specifications.

4.5 Concluding remarks

In this paper we have introduced a methodology to estimate a semiparametric specification of a gravity model for panel data with mixed effects and applied it to explain the trade flows among the EU25 countries from 2004 to 2007. We use proxy variables to filter nonparametrically the country pair effects that are supposed to be correlated with the explanatory variables. Given that these proxies' impact is modeled nonparametrically, our new model class includes the two extreme cases of fixed and random effects models. The resulting model is a semi-mixed effects model in the sense that it still has a residual random effects component. Thanks to this modeling tool we are now able to extend the suggestion of Santos Silva and Tenreyro (2006) to use PPML estimation for gravity models. Our extension is aimed to address different pitfalls which have been the source of a quite controversial discussion about which is the best model or method to analyze gravity models and / or trade. We introduced our method directly for the case of panel data analysis but should emphasize that this extension works equally well for other types of data (unbalanced panels, cohorts, cross sectional, etc.). It is well known that the introduction of an additional randomness in the (pseudo-)poisson models can substantially increase the efficiency of estimation as it allows for overdispersion, zero trade, and can easily be extended to zero inflation.

We discussed several reasons why we are convinced that, especially for short-time panels, fixed effects Poisson regression is not appropriate to estimate trade flows. On the other hand, random effects and pooled Poisson regression ignore that country unobserved effects may be correlated with explanatories which may lead to inconsistency. Our approach is a good compromise and has shown to give sensible results. As a side effect, we can also explore and handle the problem of model specification concerning possible model specification problems caused by, let's say a bad choice of units before taking the log-transform of GDP.

It should finally be emphasized that, though the idea and the model class are completely new, our approach is already now applicable as **R** and `stata` provide commands which help the practitioner to estimate these kind of models.

4.6 Appendix

4.6.1 Countries included in the data set

Austria (AU), Belgium (BE), Bulgaria (BU), the Czech Republic (CZ), Denmark (DK), Estonia (EE), Finland (FI), France (FR), Germany (DE), Greece (GR), Hungary (HU), Ireland (IR), Italy (IT), Latvia (LV), Lithuania (LH), Luxembourg (LU), the Netherlands (NE), Poland (PL), Portugal (PT), Romania (RO), Slovakia (SK), Slovenia (SV), Spain (SP), Sweden (SW) and the United Kingdom (UK).

4.6.2 Further details about the used variables

Dependent Variable:

- IMPORTS: Nominal Import (cif) flows in 10^{11} euros.

Independent Variables:

- GDPM/GDPX: Importer/Exporter country's Nominal Gross Domestic Product at Market Prices, expressed in millions of euro. Yearly data obtained from the Eurostat's New Cronos Database.
- IMPOP/EXPOP: Importer/Exporter country's Population, expressed in thousands of people at the end of the period. Data obtained from the Eurostat's New Cronos Database.
- CPIM/CPIX: Consumer Price Indexes of importer/exporter country.
- $RFE_{ijt} = \left| \log\left(\frac{GDP_{it}}{POP_{it}}\right) - \log\left(\frac{GDP_{jt}}{POP_{jt}}\right) \right|$
- $SIM_{jt} = \log\left[1 - \left(\frac{GDP_{it}}{GDP_{it}+GDP_{jt}}\right)^2 - \left(\frac{GDP_{jt}}{GDP_{it}+GDP_{jt}}\right)^2\right]$
- DISTANCE: Absolute Distance, expressed in kilometers, is the geodesic distance between capitals (in the case of the Netherlands, Amsterdam substitutes Den Haag), measured as the surface distance between two points of latitude and longitude (great circle distance). Values obtained from www.wcrl.ars.usda.gov/cec/java/lat-long.htm.

- NEIGH: Neighboring Dummy Variable is equal to one if two trading partners share a land or sea border, zero otherwise. From CIA's World Factbook 2003 on www.cia.gov/cia/publications/factbook/index.html.
- COMLANG: Common Language Dummy Variable is equal to one if two trading partners share the same official language, zero otherwise. From CIA's Factbook 2003 on www.cia.gov/cia/publications/factbook/index.html.
- ETHNIC: Ethnic Dummy Variable is equal to one if there is in the importer country an ethnic minority of the exporter country that represents more than 5% of total population of the latter, zero otherwise. From the CIA's The World Factbook 2003.
- EU15: Dummy equal to 1 if the exporting country belongs to the European 15.
- GERMAN: Dummy equal to 1 if the two involved countries is Germany.
- MLOCK/XLOCK: Landlockedness Dummy Variable for the Importer/Exporter country, is equal to one if the importing country has no direct connection to sea, zero otherwise.

Additional PROXY Variables to filter possible dependency:

- IMPLANDAREA/EXPLANDAREA: Importer/Exporter country's area. (constant in time)
- MEANPAT: Average of the number of patents of importer country recorded as EPO (European patent office) patent applications (Direct EPO filings + EURO-PCT in regional phase) in the period 2003-2005; source OECD. (constant in time)
- MEANPATPC: Average number of patents per million inhabitant of importer country recorded as EPO (European patent office) patent applications per million inhabitant in the period 2003-2005; source EUROSTAT. (constant in time)
- not to forget, some of the averages of the time varying variables from above (recall Mundlak device).

4.6.3 Further estimation results

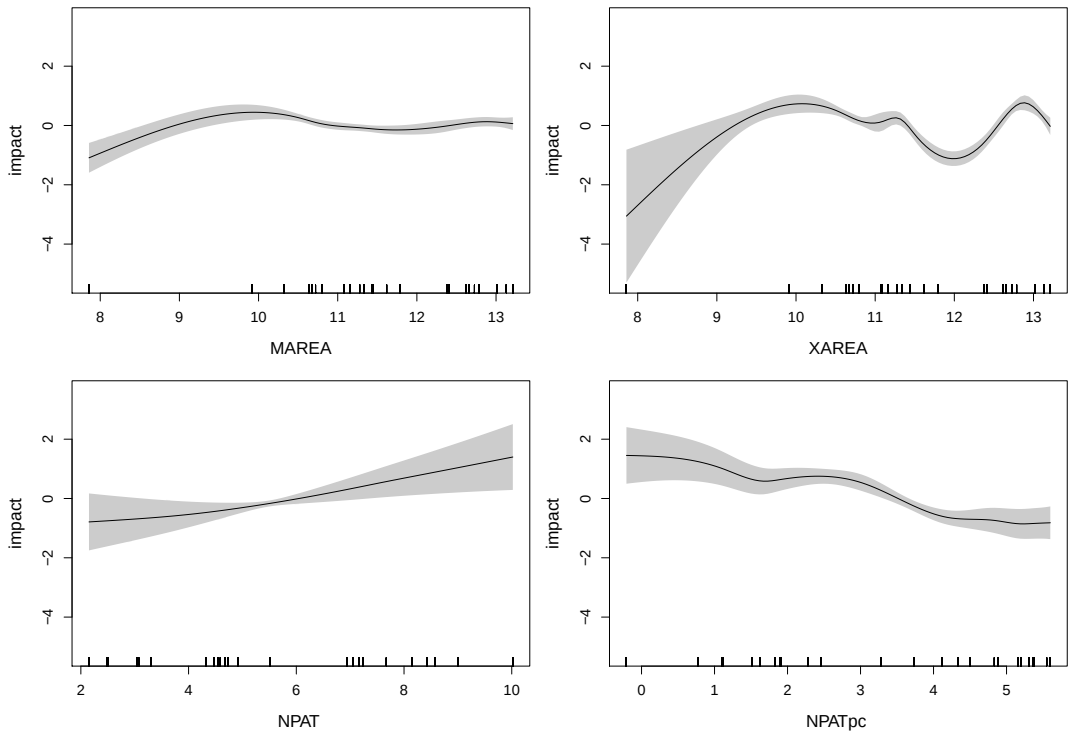

Figure 4.3: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 3.2 with 95% confidence intervals.

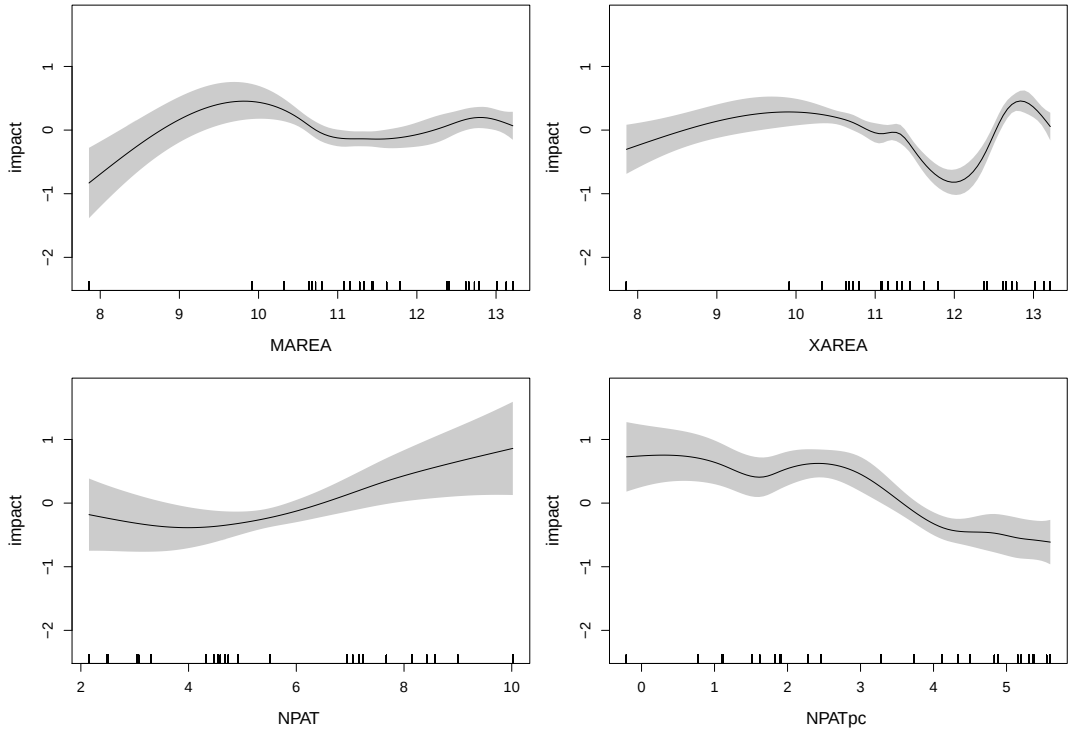

Figure 4.4: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 3.2 with 95% confidence intervals.

Model:	**3.3**	**3.4**	**4.2**	**5**	**4.3**
Const	-7.7228 .0030	-1.4173 .4553	9.9938 .0000	1.2897 .0000	-74.743 .0662
MGDP	0.7727 .0000	0.8084 .0000	n.p.e	n.p.e	n.p.e
XGDP	0.3112 .0000	0.2181 .0000	n.p.e	n.p.e	n.p.e
MPOP	0.0002 .9502	0.0012 .7500	0.0019 .5901	n.p.e	0.001 .7434
XPOP	-0.0911 .0452	0.0449 .2633	0.0641 .0931	n.p.e	-0.046 .2740
MCPI	0.1449 .4704	-0.0339 .8688	0.0642 .7690	n.p.e	0.086 .6914
XCPI	-0.3940 .0365	-0.2011 .2954	-0.5103 .0211	n.p.e	-0.880 .0001
SIM	-0.0178 .5245	0.0499 .0910	-0.0076 .8042	n.p.e	0.028 .6015
DIST	-1.1925 .0000	-1.2100 .0000	-1.2489 .0000	n.p.e	-1.231 .0000
NEIGH	0.2270 .0256	0.2774 .0073	0.3005 .0034	0.4371 .0001	0.266 .0071
COMLANG	0.1613 .3033	0.1378 .3852	0.0971 .5378	0.0607 .7005	0.127 .4030
ETHNIC	0.0560 .7708	0.2200 .2643	0.1468 .4522	-0.0551 .7760	0.095 .6094
EU15	0.7910 .0000	0.2664 .0054	0.3062 .0012	0.8044 .0000	0.520 .0000
MLOCK	0.0261 .7995	-0.3506 .0000	-0.3239 .0001	-0.1421 .2527	-0.298 .0002
XLOCK	-0.0080 .9364	0.0886 .3522	0.0894 .3485	0.1283 .1927	0.104 .2821
GERMAN	-0.0680 .4168	0.0721 .3793	0.1207 .1177	0.1269 .1010	0.036 .6394
2005	0.0223 .0015	0.0225 .0013	0.0449 .0000	0.0775 .0000	0.052 .0000
2006	0.0805 .0000	0.0817 .0000	0.1342 .0000	0.1753 .0000	0.150 .0000
2007	0.0995 .0000	0.1020 .0000	0.1854 .0000	0.2132 .0000	0.211 .0000
mMGDP		-0.5860 .0000	-0.4320 .0004		-0.614 .0000
mXGDP		0.5547 .0000	0.5398 .0000		0.504 .0000
mMPOP	-0.2905 .0302				-0.093 .4635
mXPOP	0.7297 .0000				0.304 .0005
mMCPI					-10.415 .0836
mXCPI					29.270 .0000
mSIM					-0.042 .5081
W_{add}	n.p.e	n.p.e	n.p.e	n.p.e	n.p.e
σ_u	0.547692	0.559103	0.555421	0.550515	0.531086

Table 4.4: Estimates of coefficients (with p-values), intercept and the random effects standard deviation for different model specifications. *n.p.* stands for nonparametric estimates. The latter are shown in the corresponding figures.

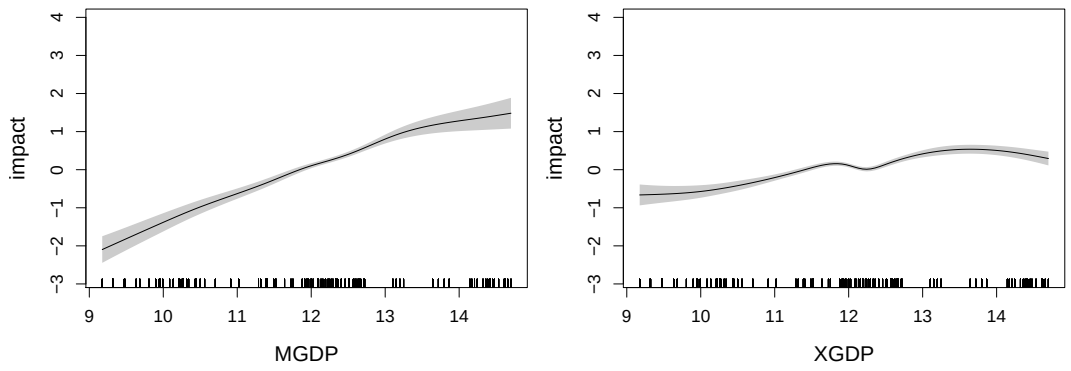

Figure 4.5: Estimates of $\beta_1(mMGDP)$ and $\beta_2(mXGDP)$ for model 4.2 with 95% confidence intervals.

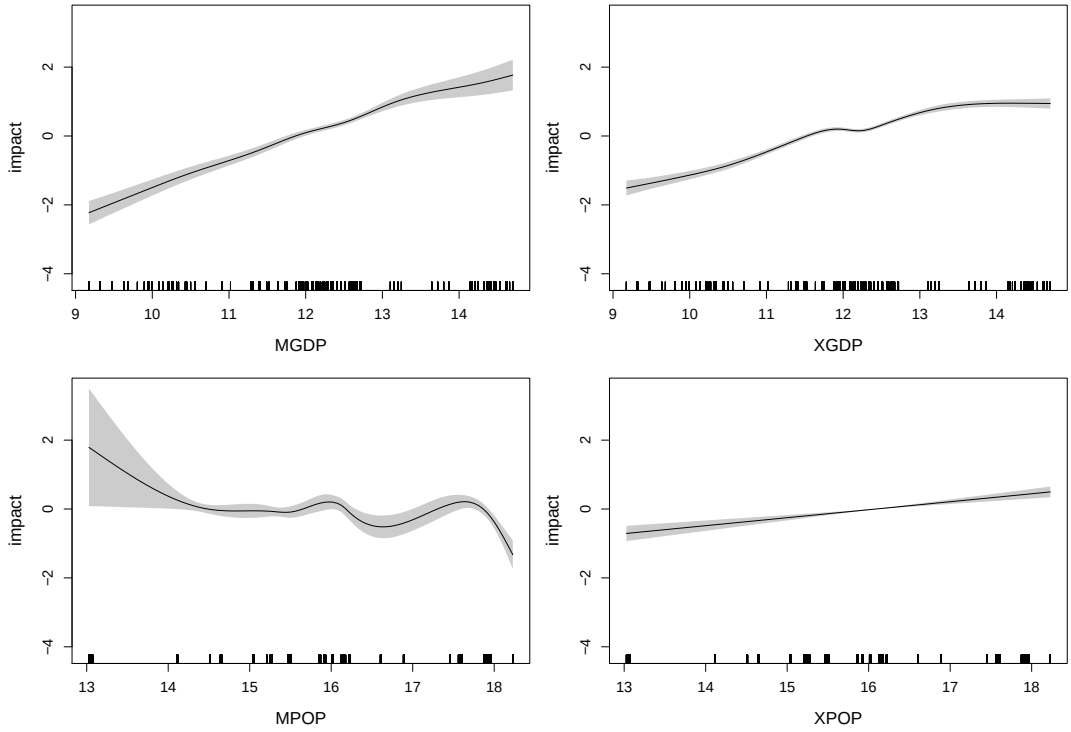

Figure 4.6: Estimates of $\beta_1(MGDP)$, $\beta_2(XGDP)$ $\beta_3(MPOP)$, $\beta_4(XPOP)$ for model 5 with 95% confidence intervals.

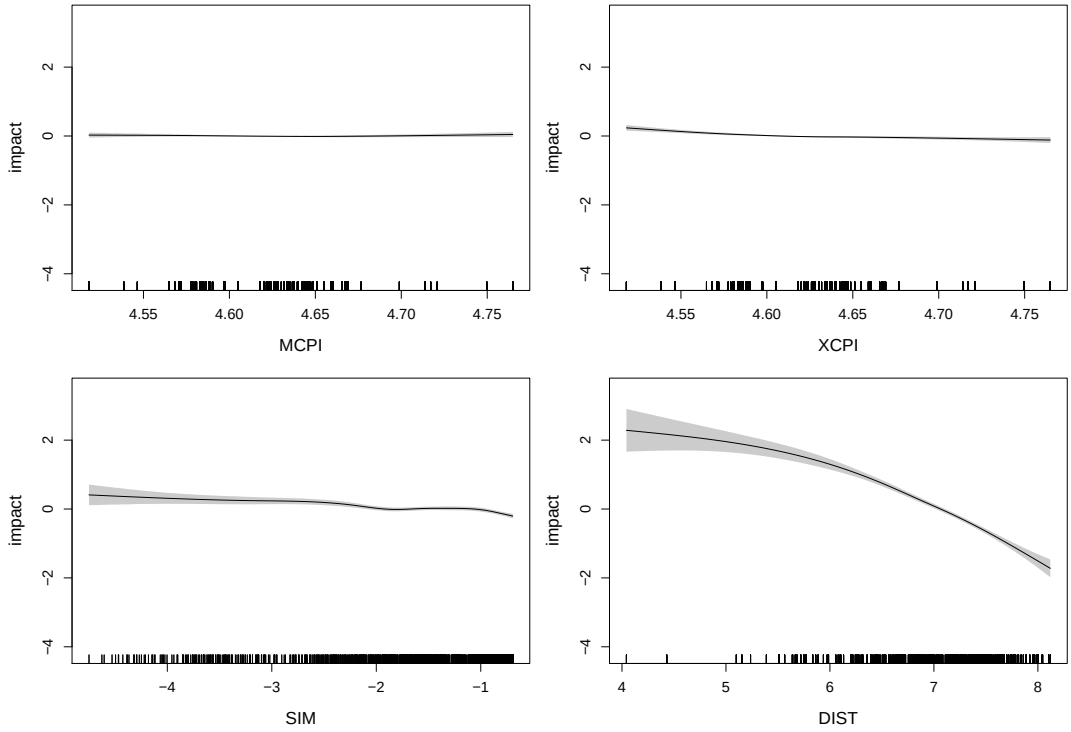

Figure 4.7: Estimates of $\beta_5(MCPI)$, $\beta_6(XCPI)$ $\beta_7(SIM)$, $\beta_8(DIST)$ for model 5 with 95% confidence intervals.

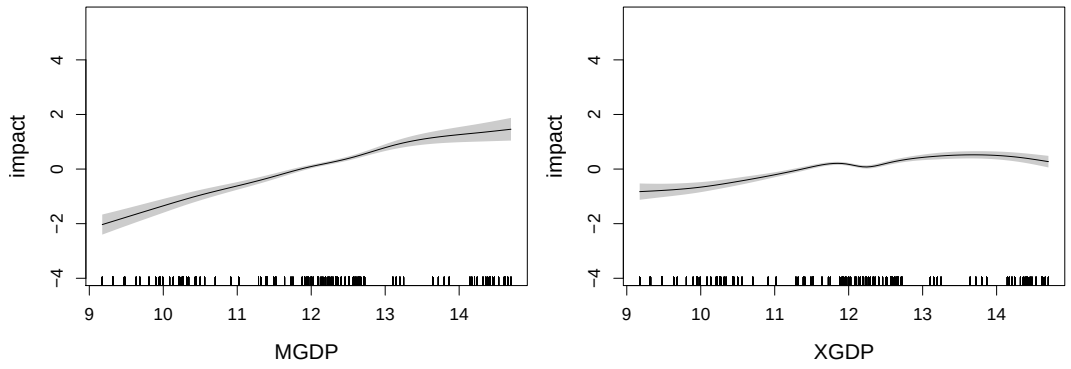

Figure 4.8: Estimates of $\beta_1(mMGDP)$ and $\beta_2(mXGDP)$ for model 4.3 with 95% confidence intervals.

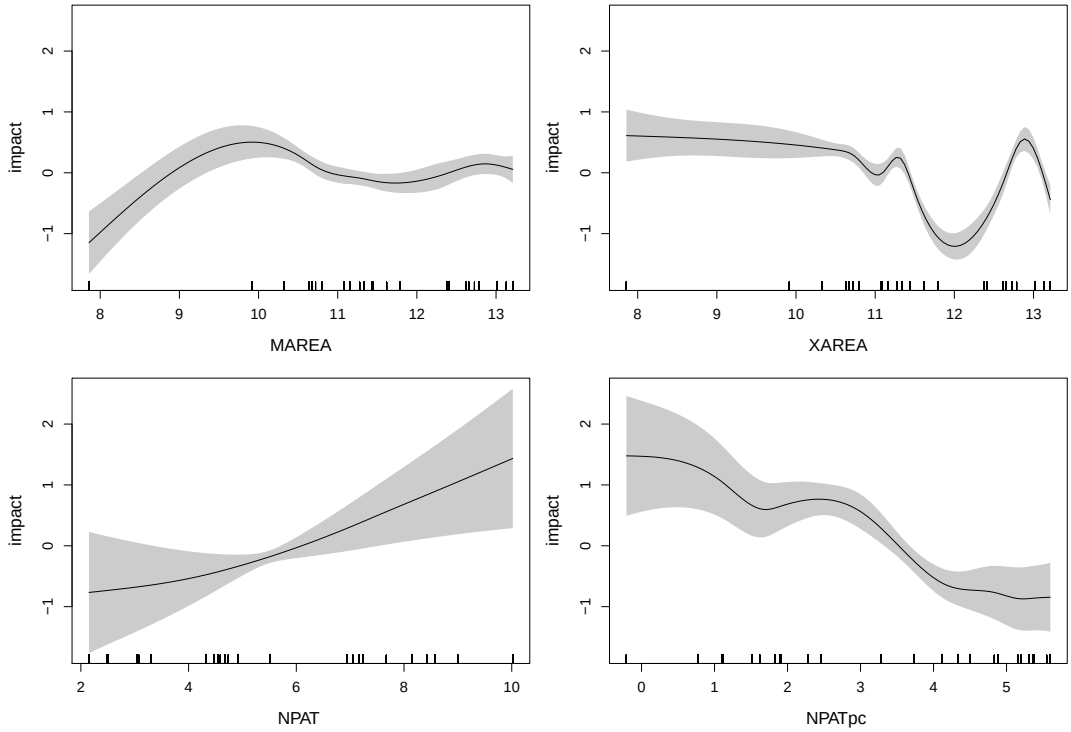

Figure 4.9: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 3.3 with 95% confidence intervals.

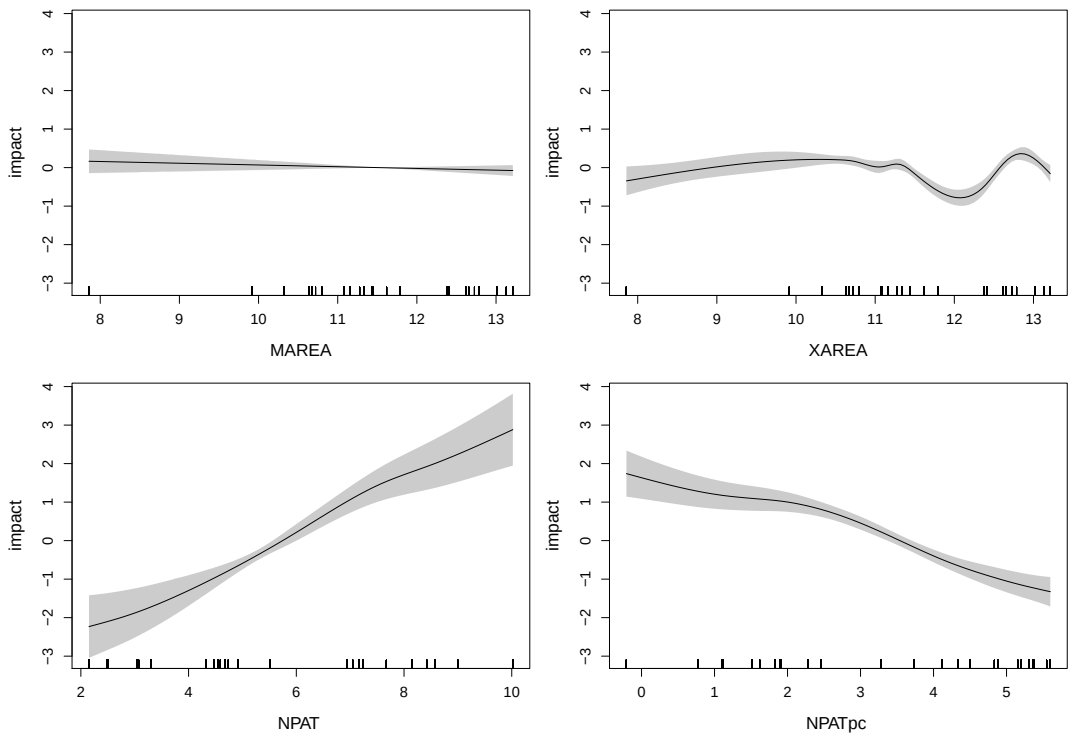

Figure 4.10: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 3.4 with 95% confidence intervals.

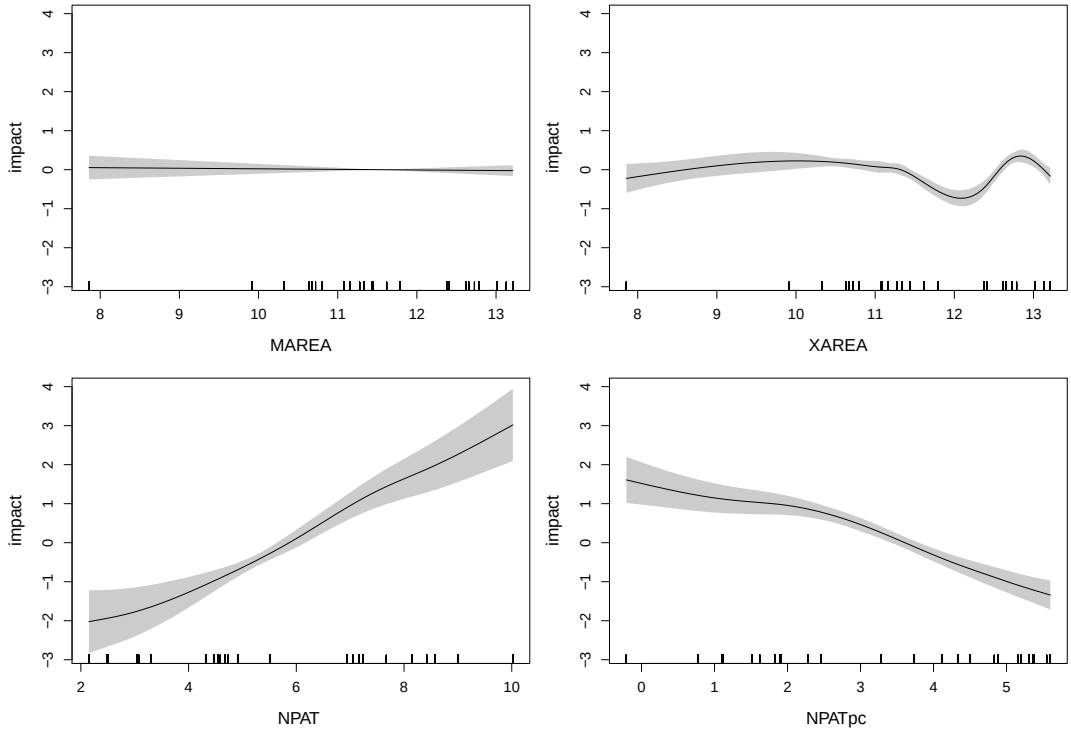

Figure 4.11: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 4.2 with 95% confidence intervals.

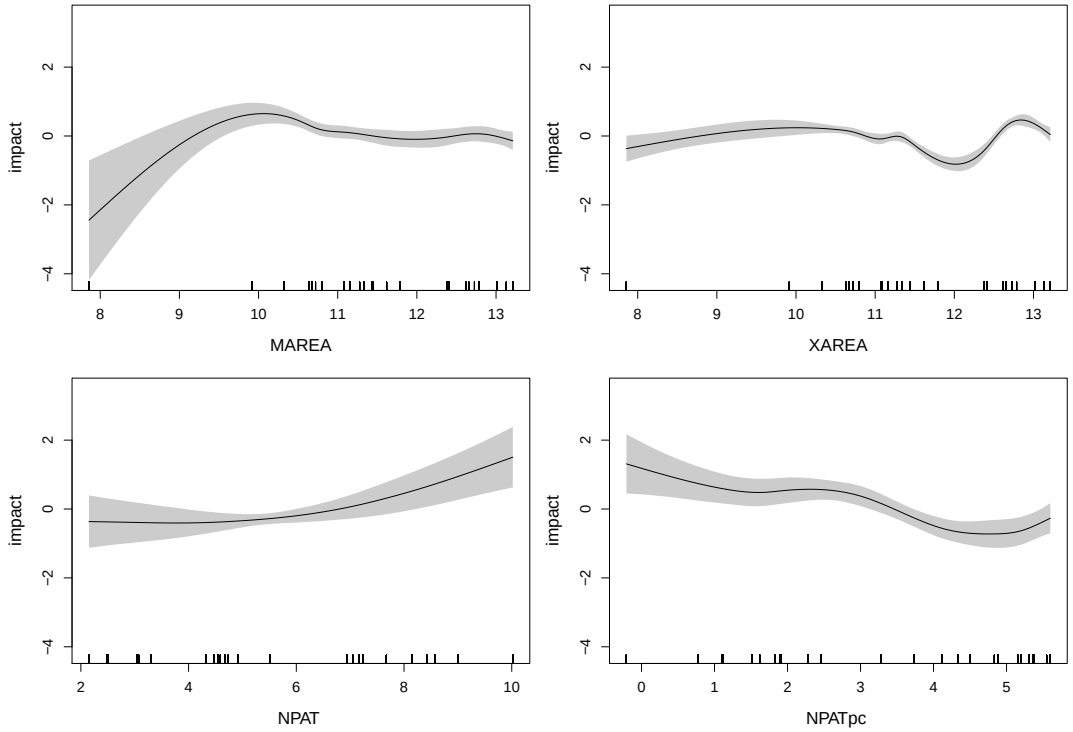

Figure 4.12: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 5 with 95% confidence intervals.

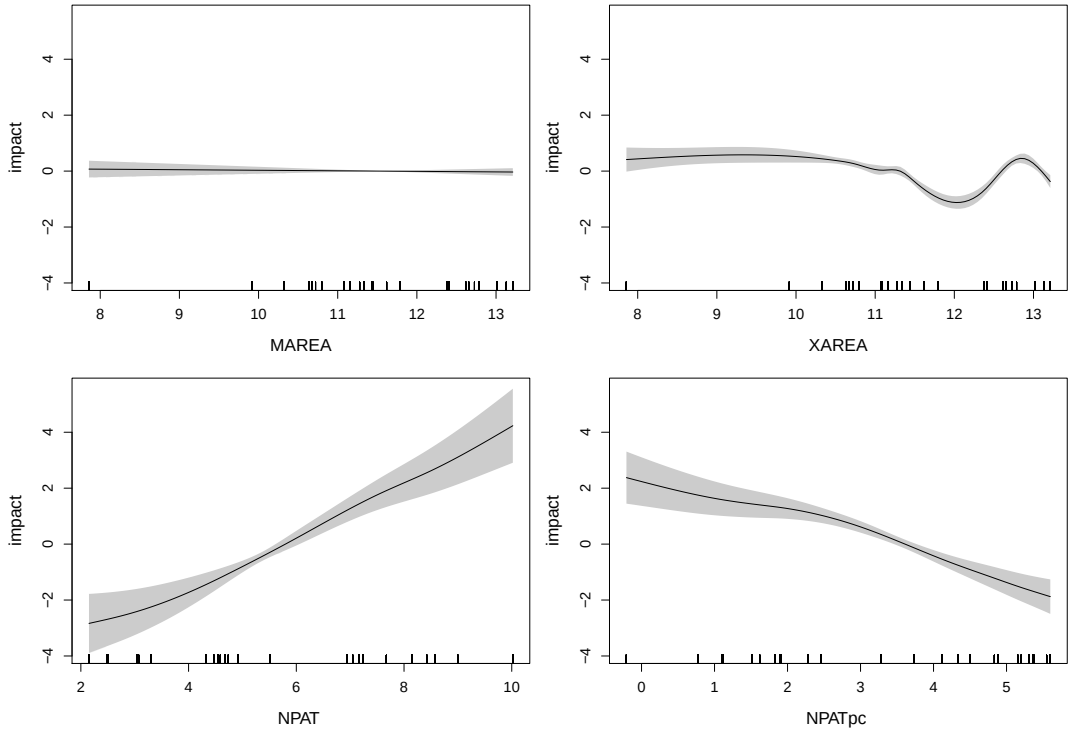

Figure 4.13: Estimates of $\psi(W) = \psi_1(MAREA) + \psi_2(XAREA) + \psi_3(NPAT) + \psi_4(NPATpc)$ for model 4.3 with 95% confidence intervals.

Summary

In this dissertation, we focused on semiparemetric estimation using mixed effects models in panel data and small area estimation where it is intended to relax the independence assumption between random effects and the covariates which can presently be considered as a main challenge in the practical use of mixed effects models.

In Chapter 2, we presented the mixed effects models and their typical applications along with small area estimation and panel data analysis. We introduced the idea along with the semi-mixed effects model referring mainly to Lombardía and Sperlich (2011). Our main focus was on the extension to P-splines. After a detailed introduction we performed and studied different implementations. We allowed for partly choosing and partly fixing the smoothness of the nonparametric components in the model. We tried different alternative estimation methods for the variance of the random effects and the pseudo-variances of the splines. We estimated the variance components simultaneously by ML method with a correction. For notational convenience, we re-wrote estimation of the variance components in a way that distinguishes the random part from the splines' pseudo random parts. Further alternatives have been implemented to account for the possibility of only estimating the variances of the truly random parts but fix the smoothness of function or pre-determine the slider for our dependence filter function. It turned out that to distinguish between random and pseudo-random parts only makes sense if we want to assume different distributions, and even then it is not evident what numerically happens and if one wants to fix the smoothness, it is better from an implementation point of view to fix one of the variances of the pseudo-random effects. We conclude that one of these implementations is sufficient. Other extensions were implemented to account for possible heteroskedasticity of either the random effects, the pseudo-random effects, or the residuals. Our implementation for the heteroskedasticity of the residuals did work but not very well. We basically followed White's well known approach in econometrics by simply using the squares of residuals on

the variance matrices. A much more successful extension to incorporate simultaneously heteroskedasticity for the error term (recall that this effects the smoothness parameter and, in our case certainly also the random effects prediction) can be found in Crainiceanu, Ruppert, Carroll, Joshi, and Goodner (2007), Krivobokova, Crainiceanu, and Kauermann (2008) or Wiesenfarth, Krivobokova, Klasen, and Sperlich (2011).

In Chapter 3, we concentrated mainly on small area estimation and we discussed the main statistical challenges when using mixed effects model estimation and prediction. The aim of the chapter was to avoid possible dependence between the random effects and the covariates with a semiparametric modeling approach using splines. After reconsidering the P-splines, we introduced briefly the basic ideas of thin plate spline regression to better account for spatial smoothing. We reviewed briefly how the prediction mean squared error can be calculated for our spline estimation approach. We carried out a small Monte Carlo simulation study to illustrate the estimation performance of the proposed model. We considered models somewhat more complex than those considered in the earlier section of Chapter 2 that were closer to the model we used in our small area environmental study. The data we used were collected by the Environmental Monitoring and Assessment Program of the US Environmental Protection Agency where they surveyed 334 lakes out of a population of 21026 in the north-eastern states of the U.S.A. between the years 1991 and 1996. While analyzing the data, we used measures of carbon trioxide and hydroxyl levels in the lakes' water as our covariates, Hydrologic Unit Codes as random effects and the elevation of the lakes as fixed effects. In our model, we had several smooth functions and they were estimated by thin plate splines and cubic splines where we could rewrite the cubic splines in additional form. We concluded that the location did matter even after having controlled for the other variables like elevation. Also, as Breidt, Opsomer, Johnson and Ranalli (2007) pointed out, simple linear mixed effects models are often not flexible enough to reflect correctly complex relationships such as those in our environmental problem. On the other hand, the crucial and always applied independence assumption is typically problematic and in our case it was clearly violated. What we also concluded was that the control functions were necessary to filter the possible dependence between covariates and area effects as otherwise all small area inference would be invalid.

In Chapter 4, we presented an application with the gravity model to explain panel bilateral country trade flows. We applied our semiparametric approach to panel gravity model via adding a nonparametric term in the transformed conditional mean in order

to capture the dependency between the explanatory variables and the unobserved individual heterogeneity term. For this application, we used the generalized additive mixed effects model, which is an additive extension of generalized mixed effects model. Based on the observations, we proposed to estimate our model with the help of a mixed effects PPML, where the unexplained heterogeneity term were random effects. The well known problem is that if this unknown heterogeneity is related with the included explanatory variables, then also this estimator is inconsistent. The best known possible remedy to this problem is the Mundlak (1978) device which is including the temporal means of the explanatories linearly in the model. In the context of small area statistics, Lombardía and Sperlich (2011) introduce a semiparametric filter to get rid of the possible dependency between this random heterogeneity and included explanatory variables. Following their idea, we claimed that for a set of time-invariant variables, where we applied the Mundlak device, there exists an unknown function and we considered this function to be an additively separable function. We allowed these time-invariant variables to enter our model nonparametrically. We used our semi-random effects gravity model to analyze the trade flows among the EU25 countries from 2004 till 2007. Given that the proxies' impact was modeled nonparametrically, our new model class included the two extreme cases of fixed and random effects models and the resulting model was a semi-mixed effects model in the sense that it still had a residual random effects component. With the help of this modeling tool, we were able to extend the suggestion of Santos Silva and Tenreyro (2006) to use PPML estimation for gravity models. We introduced our method directly for the case of panel data analysis but should emphasize that this extension works equally well for other types of data (unbalanced panels, cohorts, cross sectional, etc.).

Bibliography

Aerts, M., Claeskens, G. and Wand, M. (2002). Some Theory for Penalized Spline Additive Models. *Journal of Statistical Planning and Inference*, **103**: 455-470.

Anderson, James E. (1979). A Theoretical Foundation for the Gravity Equation. *American Economic Review*, **69** (1): 106116.

Anderson, James E. and Van Wincoop, E. (2003). Gravity with Gravitas: A Solution to the Border Puzzle. *American Economic Review*, **93**: 170-192.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Baltagi, B. H. (2005). *Econometric analysis of panel data*. John Wiley & Sons.

Baltagi, B. H., Egger, P. and Pfaffermayr, M. (2003). A Generalized Design for Bilateral Trade Flow Models, *Economics Letters*, **80** (3): 391-397.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988).An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, **83**: 28-36.

Bellman, R. E. (1957) *Dynamic Programming*. Princeton University Press, Princeton.

Bergstrand, J. H. (1985). The gravity equation in international trade: some micro-economic foundations and empirical evidence. *Review of Economics and Statistics*, 67, 474-481.

Breidt, F. J., Opsomer, J. D., Johnson, A. A. and Ranalli, M. G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, **33**: 35-44.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **88** (421): 9-25.

Brumback, B. and Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**: 9611006.

Chamberlain, G. (1982). Multivariate Regression Models for Panel Data. *Journal of Econometrics*, **18**: 5-46.

Chamberlain, G. (1984) Panel Data. *Handbook of Econometrics*, Vol II. Edit.: Griliches, Z. and Intriligator, M.D., Elsevier Sience Publisher. Chapter 22; 1247-1318.

Claeskens, G., Krivobokova, T., Opsomer, J. D. (2009). Asymptotic Properties of Penalized Spline Estimators. *Biometrika*, **96**: 529-544.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A., and Goodner, B. (2007). Spatially adaptive bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, **16** (2): 265-288.

Das, K., Jiang, J. and Rao, J. N. K. (2004). Mean squared error of empirical predictor. *Annals of Statistics*, **32**: 818-840.

Davidian, M., and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. New York: Chapman and Hall.

Davis, B. (2003). *Choosing a method for poverty mapping*. Food and Agriculture Organization of the UN, Rome.

Deardorff, A. V. (1998). Determinants of Bilateral Trade: Does Gravity Work in a Neoclassical World?. *In The Regionalization of the World Economy*, edited by Jeffrey A. Frankel. Chicago: University of Chicago Press, 722.

Diggle, P. J., Heagerty, P., Liang, K. L., and Zeger, S. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford University Press.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Solobev spaces. In *Construction Theory of Functions of Several Variables*. Berlin: Springer.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11** (2): 89-121.

Deaton, A. and Muellbauer, J. (1980). *Economics and Consumer Behavior*. Cambridge University Press, Cambridge.

Elbers, C., Lanjouw, J. O., and Lanjouw, P. (2003). Micro-level Estimation of

Poverty and Inequality. *Econometrica*, **71**, 355-364.

Fahrmeir. L., Lang, S. (2001). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Journal of the Royal Statistical Society*, C, **50** (2): 201220.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer Series in Statistics, Springer-Verlag: New York.

Fay, R. E. and Herriot, R. A. (1979). Estimates of Income for Small Places. An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**: 269-277.

Foulley J. L. and Quaas R. L. (1995). Heterogeneous variances in Gaussian linear mixed models. *Genet. Sel. Evol.* **27**: 211-228.

Friedman, J. H. and Stuetzle, W. (1981. Projection Pursuit Regression. *Journal of the American Statistical Association*, **76**: 817823.

Ghosh, M., and Rao J. N. K. (1994). Small area estimation: An appraisal. *Statistical Science*, **9** (1): 55-76.

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, **93**: 273-282.

Greene, W. (2003). *Econometric Analysis*, Fifth Edition, Prentice Hall.

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

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

Gurmu, S., Rilstoneb, P. and Sternc, S. (1999). Semiparametric estimation of count regression models . *Journal of Econometrics*, **88** (1): 123-150.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A. (2004). *Nonparametric and Semiparametric Models*. New York: Springer Series in Statistics, Springer.

Hastie, T. J., Tibshirani, R. J. (1986). Generalized additive models. *Stat. Sci.*, **1**: 297-318.

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

Chapman and Hall.

Henderson, C. R., Kempthorne, O., Searle S. R. and von Krosigk, C. M. (1959). The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, **15** (2): 192218.

Henderson, D. J., and Millimet, D. L. (2008). Is Gravity Linear?. *Journal of Applied Econometrics*, **23**: 137-172.

Horowitz, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. Springer Series in Statistics.

Hsiao, C. (2003). *Analysis of Panel Data*, Second Edition, Cambridge University Press.

Ichimura, H. (1993). Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models. *Journal of Econometrics*, **58**: 71120.

Jiang, J. and Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. *Test*, **15** (1), 1-96.

Kauermann, G. and Opsomer, J. D. (2011). Data-driven Selection of the Spline Dimension in Penalized Spline Regression. *Biometrika*, **98**: 225-230.

Kneib, T., Fahrmeir, L. (2007). A Mixed Approach for Geoadditive Hazard Regression. *Board of the Foundation of the Scandinavian Journal of Statistics*, **34**: 207-228.

Krivobokova, T., Crainiceanu, C., and Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, **17** (1): 1-20.

Lahiri, P. and Rao, J. N. K. (1995). Robust estimation of mean squared error of small area estimators. *J. Amer. Statist. Assoc.*, **90**: 758-766.

Lahiri, P. (2003). On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science*, **18**: 199-210.

Larsen, D. P., Thornton, K. W., Urquhart, N. S. and Paulsen, S. G. (1994). The Role of Sample Surveys for Monitoring the Condition of the Nation's Lakes. *Environmental Monitoring and Assessment*, **26**: 2428-36.

Lin, X. H. and Carroll, R. J. (2001). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.*, **96**: 10451056.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed model using smoothing splines. *Journal of the Royal Statistical Society*, Series B, **61**: 381400.

Lombardía, M. J. and Sperlich, S. (2006). Notes on Semiparametric Inference in Small Areas. *Working Paper*, Georg-August-Universität Göttingen.

Lombardía, M. J. , Sperlich, S. (2008). Semiparametric Inference in Generalized Mixed Effects Models. *Journal of the Royal Statistical Society*, B, **70** (5): 913-930.

Lombardía, M. J. and Sperlich, S. (2008). Multi-level Regression Between Fixed Effects and Mixed Effects Models, *manuscript*.

Lombardía, M. J. and Sperlich, S. (2011). A new class of Semi-Mixed Models and its Application in Small Area Estimation. Forthcoming.

Martinez-Zarzoso, I. (2011). The Log of Gravity Revisited. *Discussion paper at the University of Göttingen, under revision*.

Mátyás, L. (1997). Proper econometric specification of the gravity model. *The World Economy*, **20** (3): 363-368.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models*, Second Edition. Chapman and Hall, London.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear and Mixed Models*. Wiley-Interscience, New York.

Militino, A. F., Ugarte, M. D. and Goicoa T. (2006). Combining sampling and model weights in agriculture small area estimation. *Environmetrics*, **18**: 87-99.

Mundlak, Y. (1978). On the Pooling of Time Series and Cross Section Data. *Econometrica,* **46**: 69-85.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, **370** (3): 370-384.

Niedercorn, J. H., Bechdolt, B. V. (1969). An Economic Derivation of the "Gravity Law" of Spatial Interaction. *Journal of Regional Science*, **9** (2): 273-282.

Opsomer, J., Claeskens, G., Ranalli, M. G., Kauermann, G., and Breidt, F. J. (2008). Nonparametric Small Area Estimation Using Penalized Spline Regression. *Journal of the Royal Statistical Society*, Series B, **70**: 265-286.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58** (3): 545554.

Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, **74**: 967-1012.

Polachek , S. W. and Kim M. (1994). Panel estimates of the gender earnings gap: individual-specific intercept and individual specific slope models. *Journal of Econometrics*, **61**: 23-42.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, **85**: 163-171.

Pratesi, M., Ranalli, M. G. and Salvati, N. (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the North-eastern US. *Environmetrics*, **19**: 687-701.

Proença, I., Fontoura, M. P. and Martnez-Galn, E. (2008). Trade in the Enlarged European Union: a new approach on trade potential. *Portuguese Economic Journal*, **7**: 205224.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, **119** (1): 99-130.

Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley and Sons, Inc., New York.

Robert-Granié C., Heude B. and Foulley J.L. (2002). Modelling the growth curve of Maine-Anjou beef cattle using heteroskedastic random coefficients models. *Genet. Sel. Evol.*, **34**: 423-445.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6** (1): 1532.

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

Santos Silva, J. and Tenreyro, S. (2005). The Log of Gravity. *C.E.P.R. Discussion Papers*, **5311**.

Santos Silva, J. and Tenreyro, S. (2006) The log of gravity. *The Review of Economics and Statistics*, **88** (4): 641-58.

Santos Silva, J. and Tenreyro, S. (2008). Comments on "The log of gravity revisited". *University of Essex, Department of Economics*

Searle, S. R. (1971). *Linear Models*. New York: Wiley.

Searle, S. R., Casella, G., and McCulloch, C. E. (1982). *Variance Components*. New York. Wiley.

Serlenga, L. and Shin, Y. (2007). Gravity models of intra-EU trade: application of the CCEP-HT estimation in heterogeneous panels with unobserved common time-specific factors. *Journal of Applied Econometrics*, **22** (2): 361-381.

Severini, T. A., Staniswalis, J. G. (1994). Quasi-likelihood Estimation in Semiparametric Models. *Journal of the American Statistical Association*, **89** (426): 501-511.

Sperlich, S. (1998). *Additive Modelling and Testing Model Specification*. Shaker Verlag, Aachen, Germany.

Sperlich, S., Härdle, W., Aydinli, G. (2006). *The Art of Semiparametrics*. Physica-Verlag.

Sperlich, S. and Lombardía, M. J. (2010). Local Polynomial Inference for Small Area Statistics: Estimation, Validation and Prediction. *Journal of Nonparametric Statistics*, **22** (5): 633-648.

Tinbergen, J. (1962). *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: The Twentieth Century Fund.

Ugarte, M. D., Goicoa, T., Militino A. F. and Durbán, M. (2009). Spline Smoothing in small area trend estimation and smoothing. *Computational Statistics and data Analysis*, **53**: 3616-3629.

Verbeke, G. and Molenberghs, G. (2009). *Linear mixed models for longitudinal data*, 2nd ed. Springer Verlag.

Wahba, G. (1980). Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data. In *Approximation Theory III* (ed.W. Cheney), 905912. New York: Academic Press.

Wahba, G. (1990). Spline models for observational data. *CBMS-NSF Regl Regional Conference Series in Applied Mathematics*, **59**.

Wand, M. P. (2003). Smoothing and Mixed Models. *Computational Statistics*, **18**:

223-249.

Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics*, **35** (5): 2106-2142.

Westerland, J. and Wilhelmsson, F. (2009). Estimating the gravity model without gravity using panel data. Forthcoming in *Applied Economics*.

White, H. (1980). A heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity. *Econometrica*, **48** (4): 817-838.

Whittier, T. R., Paulsen, S. G., Larsen, D. P., Peterson, S. A., Herlihy, A. T., and Kaufmann, P. H. (2002). Indicators of Ecological Stress and Their Extent in the Population of Northeastern Lakes: A Regional-Scale Assessment *BioScience*, **52** (3): 235-247.

Wiesenfarth, M., Krivobokova, T., Klasen, K., and Sperlich, S. (2011). Direct Inference for Local Adaptive Additive Models and its Application to Model Undernutrition, Working Paper University of Göttingen.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

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

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

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.

Zheng, X. (2008). Semiparametric Bayesian estimation of mixed count regression models. *Economics Letters*, **100** (3): 435-438.