PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS

Functional regression models

Supervisor: **doc. RNDr. Eva Fišerová, Ph.D.**
Author: **Mgr. Veronika Římalová**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2023

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Mgr. Veronika Římalová

Název práce: Funkcionální regresní modely

Typ práce: Disertační práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Eva Fišerová, Ph.D.

Rok obhajoby práce: 2023

Abstrakt: Cílem této disertační práce je představit nový neparametrický přístup k testování významnosti parametrů ve funkcionálním lineárním regresním modelu pro prostorová data pro případ heteroskedasticity nebo prostorové korelace. Pro testování významnosti regresních parametrů v prostorovém funkcionálním regresním modelu navrhujeme využít metodiku založenou na permutačních testech. Tyto metody jsou navrhnuty tak, aby bylo možno s jejich pomocí pracovat s daty s heterogenní prostorovou strukturou a to na základě Freedmanova a Laneova permutačního schématu. Přirozeně, vlivem prostorové závislosti mezi daty nejsou residua regresního modelu permutovatelná, čímž je porušen základní předpoklad Freedmanova a Laneova permutačního schématu. Abychom tento problém vyřešili, navrhujeme v případě heteroskedasticity modelovat data pomocí vážené metody nejmenších čtverců a poté vydělit odhadnutá residua jejich směrodatnou odchylkou, čímž získáme asymptoticky permutovatelná residua. V případě prostorově závislých dat navrhujeme odhadnout varianční strukturu pomocí variogramu a poté dekorelovat odhadnutá residua a založit permutační test na těchto přibližně permutovatelných residuích. Za účelem ohodnocení navrhovaných testů z hlediska empirické velikosti a síly testu byly provedeny simulační studie zkoumající chování testů pro různé varianční struktury dat. Ukážeme, že zanedbání varianční struktury residuí (tedy permutování přímo heteroskedastických či prostorově korelovaných dat) vede v permutačním testování k buďto velmi liberálním, či velmi konzervativním výsledkům testů, zatímco empirická velikost námi navrhovaných testů je v případě heteroskedasticity či prostorové korelace blízká té nominální. Navrhované metody jsou též prezentovány na reálných datech z oblastí geochemie a turismu. Mimo to je potenciál analýzy funkcionálních dat ukázán na datech z oblasti dopravního výzkumu s důrazem na regresní model s funkcionálním regresorem a funkcionální závisle proměnnou.

Klíčová slova: Analýza funkcionálních dat, funkcionální regresní modely, permutační testy, neparametrická inference, heteroskedasticita, prostorová korelace, funkcionální geostatistika, prostorově filtrovaná residua, permutovatelnost

Počet stran: 110

Počet příloh: 0

Jazyk: anglický

BIBLIOGRAPHICAL IDENTIFICATION

Author: Mgr. Veronika Římalová

Title: Functional regression models

Type of thesis: Dissertation thesis

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Eva Fišerová, Ph.D.

The year of presentation: 2023

Abstract: This thesis proposes a novel nonparametric approach to the significance testing of the null hypothesis in a functional linear model for data with a heterogeneous spatial structure. A permutation approach is introduced to test for the effect of covariates in a spatial functional regression model with heteroscedastic or spatially correlated residuals. In this context, the proposed methods account for the heterogeneous spatial structure of the data by grounding on the Freedman and Lane permutation scheme for the estimated residuals of the functional regression model. Indeed, due to the spatial dependence among the data, the residuals of the regression model are not exchangeable. Therefore, the basic assumption of the Freedman and Lane permutation scheme is violated. To overcome this issue in the case of heteroscedasticity, we propose to fit a weighted least squares model to the observations, and then to divide the estimated residuals by their corresponding standard deviation, leading to asymptotically exchangeable, and thus, permutable residuals. In the case of spatially correlated observations it is proposed here to estimate the variance-covariance structure of the residuals by variography, remove this correlation by spatial filtering of the residuals and base the permutation test on these approximately exchangeable residuals. To evaluate the performance of the proposed methods in terms of the empirical size and power, simulation studies are conducted, examining the behaviour of the tests under different covariance settings. We show that neglecting the spatial structure of the residuals in the permutation scheme, i.e., permuting the heteroscedastic or spatially correlated residuals directly, yields very liberal or conservative results, whereas the proposed procedures are close to the nominal size of the test. The results of modelling and testing on the case studies are shown and discussed on the data from geochemistry and tourism. Moreover, the potential of FDA methodology is shown in the field of transportation research, focusing on the permutation-based inference in a regression model with functional covariates and functional response.

Key words: Functional data analysis, functional regression models, permutation tests, nonparametric inference, heteroscedasticity, spatial correlation, functional geostatistics, spatially filtered residuals, exchangeability

Number of pages: 110

Number of appendices: 0

Language: English

Statement of originality

I hereby declare that this dissertation thesis has been completed independently, under the supervision of doc. RNDr. Eva Fišerová, Ph.D. All the materials and resources are cited concerning scientific ethics, copyrights and the laws protecting intellectual property. This thesis or its parts were not submitted to obtain any other or the same academic title.

In Olomouc ....

Contents

List of Figures

Acknowledgement

I would like to express my gratitude to my supervisor Eva Fišerová for her support, motivation, valuable advice, patience and guidance through my PhD studies. I am also grateful for the guidance, inspiration and help from Alessandra Menafoglio and Alessia Pini during my research stay in Milan. My parents, family and friends are gratefully appreciated for being there for me. I would also like to thank my colleagues from studies for a great and inspiring working environment. Finally, I am particularly grateful to Pavel for his understanding, encouragement and endless support during my studies, research and writing this thesis.

Introduction

In many practical tasks involving high-dimensional measurements, such as curves or surfaces, it is convenient to treat the data not as a sequence of single measurements taken one after another, but as the whole entities. Such approach has developed into functional data analysis (FDA), where the basic elements of the statistical analysis are the functional observations as a whole [28, 62]. In these days, functional data occur in many scientific fields, such as an analysis of human gait by means of linear models [1], analysis of variance [69], modelling the time trends and forecasting the future rates of age-specific breast cancer mortality using time series models [25, 26], or predicting the risk of drought by principal component logistic regression [27]. The overview of the recent trends in FDA can be found in summary surveys [8, 9, 35].

Most FDA methods strongly rely on the assumption of independence among the observations. Under the presence of spatial dependence, applying these methods is inappropriate, and the analysis could fail because of the consistency problems [39]. Spatial dependence needs to be properly treated; for this purpose, the classical geostatistical methods can be extended to the functional framework [15, 51, 52].

In the geostatistical framework, the functional observations can occur, e.g., as soil or weather measurements taken from the same site with possibly daily, weekly or monthly frequency (see, e.g., [41, 62, 67]). In general, the extension of the geostatistical methods to the functional setting is well-developed these days, including ordinary or universal kriging [34, 50, 51, 52], or other aspects related to spatial statistics, such as the change point analysis [44], clustering [68], tests

for detecting the spatial autocorrelation [33], or the regression with differential regularization to model dependent functional data in space [11].

In functional data analysis, regression models play an important part of the methods used. A thorough overview of the methodology with application can be found in vast amount of the literature, see, e.g., [61, 62], focusing on situations where at least the covariates, or the response, or both is functional. The cases of the concurrent model, as well as the functional-on-functional model, where the functional domains of the covariate and the response differ, i.e., $\beta(t_1, t_2)$ is a function of these two domains, are detailed here. In the work of [39], the methodology for models for independent, dependent, and spatially correlated data is introduced. In the work of [28], the focus is on the analysis of nonparametric functional data, including the models for such data. In the work of [43], functional regression models, including the generalized models and models for spatially correlated data, are concerned. An overview of recent trends in FDA can be found in summary surveys [8, 9, 35], including also topics related to functional regression models, see, e.g., the work on variables selection, such as [17], or [13], where the variables selection is based on smoothed centred ridge regression approach. In the work of [23], a nonparametric approach to the construction of prediction bands in the multi-functional regression framework is concerned.

Focusing on the inference in the context of functional data, testing for the significance can be performed by means of either parametric or nonparametric methods. The nonparametric approach appears particularly promising, as it allows for minimal assumptions on the data-generative model. In this context, permutation tests have been recently successfully developed, leading to flexible approaches for testing in FDA [1, 60]. Indeed, the permutation tests are only based on the assumption of exchangeability between the units under the null hypothesis [42, 49], meaning that, under H_0, the data distribution does not change if the units are randomly resampled [36, 58]. In practice, this is evaluated by comparing the test statistic evaluated on the original data to the distribution of the test statistics obtained from the permuted data. If the observed statistic lies

in the tail of this distribution, the null hypothesis can be rejected [12, 61].

In the context of a two-population test, the idea behind the permutation scheme is as follows (see, e.g., [58]). Under the null hypothesis of equality in distribution of two populations, the distribution of the two samples is the same as the distribution of any two samples obtained upon permuting the data. A test statistic computed for the original sample (e.g., the difference of the means) should thus have the same distribution as that calculated from any permutation of the data. Hence, the evidence against H_0 is provided by values of the test statistic under permutations that are statistically different from the one obtained on the original data [49, 58]. Due to a large number of all possible permutations (a factorial of the sample size) it would be hardly feasible to consider all possible rearrangements of the data. Instead, the permutation distribution of the test statistic can be approximated with Monte Carlo methods by randomly generating a subset of permutations [30, 47, 58].

In the framework of linear models, it is not possible to directly observe the exchangeable quantities under the null hypothesis unless the model only contains a single covariate. The permutation scheme can be adapted in different ways to obtain the approximate exchangeability. In the work of [31], permuting the estimated residuals from the model under the null hypothesis (also referred to as the reduced model) was proposed; in the work of [71], permuting the residuals of the full model was proposed; other authors, see [48, 54], proposed to permute other quantities, such as the covariates' values, or the responses. In the work of [6], Freedman and Lane's methodology (see [31]) and other approaches were compared by a simulation study, concluding that the permutation scheme proposed by Freedman and Lane [31], relying on the permutation of the estimated residuals from the reduced model, gives the best empirical results in terms of the power and the size of the test. From a theoretical point of view, since the estimated residuals of the null model are asymptotically exchangeable, the obtained test is asymptotically exact. In the work of [75], this approach was applied to the multivariate case of testing for the differences between two sets of images; in the

work of [1], Freedman and Lane permutation scheme was adapted to testing in a functional regression model for the knee movement.

The aim of this thesis is to introduce a novel approach to the inference in functional regression models, aiming also at more complex data structures. In particular, a testing procedure for a regression model with functional covariate and functional response, based on a Freedman and Lane permutation scheme is outlined in Chapter 3, being a novel approach in the field of transportation research, grounding on a real-world task from this research area.

Extending this methodology from homoscedastic [1, 64] to the space-time setting, both the observed data and the residuals are likely to be heterogeneous by virtue of their geographical closeness and thus treating them as exchangeable may bias the results. This methodology is developed in Chapter 4. Firstly, a case of heteroscedasticity among the spatial observations is concerned [66]. Under the presence of more groups in the data, an interesting research question is whether these groups have different properties. Assuming the same variance within the groups and the different variance between the groups, a permutation testing procedure based on a spatial regression model is here proposed. To deal with the nature of spatial observations, which depend on their spatial coordinates, we propose to permute the residuals of the spatial regression model and not the observations themselves. For this purpose, the permutation scheme by Freedman and Lane (see [31]) is adapted here. An emphasis shall be given to the exchangeability of the residuals, since this assumption can be easily violated whenever the data are heteroscedastic.

Secondly, a case of spatially correlated observations is concerned [65]. To ensure the approximate exchangeability of the units being permuted, a spatial filtering of the residuals of the functional linear model is proposed. More precisely, the spatial covariance of the residuals is estimated through the functional variography [39] and then the residuals are decorrelated by using the inverse-square root of the estimated variance-covariance matrix, eventually obtaining approximately exchangeable residuals, suitable for using the Freedman and Lane permutation

scheme. A similar approach can be found in [10], where the spatial covariance is estimated through the trace variogram and used to fit the functional analysis of variance model with GLS. Then, the inference is based on an approximate chi-squared distribution of the test statistic, that is derived based on the assumption of normality of the residuals, whereas a nonparametric permutational approach is proposed in this work.

To summarize the previous paragraphs, the aims of this thesis are to

1. model functional data and to set up the methodology for spatiotemporal observations, the estimation of the spatial relationship among the observations through variography (Chapter 4),

2. introduce a novel approach to the inference in spatial functional regression models based on a permutation scheme by Freedman and Lane [31], aiming at the inference in case of (i) heteroscedastic and (ii) spatially correlated functional observations (Chapter 4),

3. apply the proposed methodology to real-world spatiotemporal data (Chapter 4),

4. show a potential of FDA framework in the field of transportation research, focusing on the inference in functional-on-functional regression models (Chapter 3).

In Chapter 1, the real-world data which are further analysed in this thesis using the methods concerned here are introduced to provide examples of the functional observations. The basic aspects of functional data analysis and data processing are summarized in Chapter 2. Functional-on-functional regression model, together with the inferential framework suitable for this setting, is concerned in Chapter 3, followed by an application to the driving speed data. Chapter 4 introduces the inferential framework to regression models for spatiotemporal functional data, focusing on heteroscedastic and spatially correlated observations. The performance of the proposed methodology is examined through extensive

simulation studies. Finally, the methods for spatiotemporal data are applied to real functional observations from the fields of geochemistry and tourism.

To avoid confusion, the following notation is used in this thesis. Spatiotemporal data, i.e., the observations recorded in both time and space, are denoted as $\mathcal{X}_s(t)$, where s refers to a position of the observation in space D and $t \in T$ refers to a (functional) temporal domain. On the other hand, temporal data, i.e., the observations recorded in time, are denoted as $y(t)$.

Moreover, it should be noted here that the *random errors* in a regression model are referred to as *residuals* in Chapter 4, and their estimates, *residuals*, are referred to as the *estimated residuals*, consistently with the commonly used notation in the geostatistical literature. In Chapter 3, the standard notation, i.e., *random errors* and *residuals*, is used.

This dissertation thesis is based on the following papers that were published during my PhD. study:

- **V. Římalová**, A. Menafoglio, A. Pini, V. Pechanec, and E. Fišerová. A permutation approach to the analysis of spatiotemporal geochemical data in the presence of heteroscedasticity. Environmetrics, 31(4):e2611, 2020 (Chapter 4).

- **V. Římalová**, E. Fišerová, A. Menafoglio, and A. Pini. Inference for spatial regression models with functional response using a permutational approach. Journal of Multivariate Analysis, 189:104893, 2022 (Chapter 4).

- **V. Římalová**, J. Elgner, J. Ambros, and E. Fišerová. Modelling the driving speed on expressway ramps based on floating car data. Measurement, 195:110995, 2022 (Chapter 3).

Chapter 1

Motivational examples

1.1. The driving speed on expressway ramps based on floating car data

The first case study is based on the data set collected on the interchange number 2432A012 (road I/42 × II/384, Žabovřeská × Kníničská street) in Brno, Czech Republic, during the years 2014 and 2015. This interchange can be further divided into six unique ramps with several auxiliary lanes (i.e., merging and exit lanes), see Figure 1.1. There are merging lanes into ramps 1 and 2 (see Figure 1.2a), and exit lanes from ramps 3 and 5, (see Figure 1.2b).

These data were retrieved from the floating car data (FCD), collected by a fleet of company vehicles, which is maintained by a third party. The raw data include time, position and speed recorded at the 4 Hz frequency (4 times per second) and are stored by the Vetronics on-board units. According to the data provider, the position accuracy in terms of CEP (Circular Error Probable, i.e., the median error radius) was 2.5 m; the speed accuracy was 2 km/h. The data provider collects this data for the purpose of the fleet management, monitoring of the fuel consumption, vehicle tracking, recording of the travel diaries, etc. Nevertheless, FCD may also be used for secondary tasks, such as the analysis of the kinematic characteristics (speed, acceleration, etc.), as demonstrated, e.g., by [3, 4, 5].

It is upon interest to (i) investigate the smoothness of the speed when passing a ramp, (ii) point out the sections with a rapid change (decrease or increase)

of the speed, (iii) describe and evaluate the effect of the ramp curvature and of the auxiliary lanes on the vehicle's speed, and (iv) identify the parts where the observations from the central and the auxiliary lanes are indistinguishable in terms of their speed. The theoretical background and the case study are concerned in Sections 3.1 - 3.5, and the case study is detailed in Section 3.6.

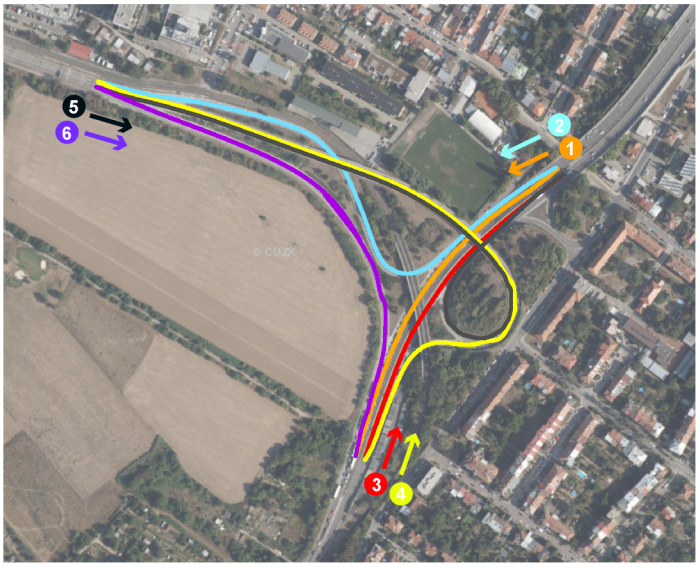


Figure 1.1: Scheme of the interchange number 2432A012 (road I/42 × II/384, Žabovřeská × Kníničská street).

(a) Merging into ramps 1 and 2.

(b) Exit from ramps 3 and 5.

Figure 1.2: A view of the selected parts of the ramps, Google street view, June 2014.

1.2. The soil samples collected on the border between field and forest

The second case study is based on the data resulting from the analysis of the soil samples collected in the growing seasons (March-October) of the years 2015 and 2016 at the site located near Křtiny, Czech Republic.

The site contains 11 sampling points equally distributed on a straight line, perpendicular to the border between the field and the forest. The sampling points are located every 3 metres, and the total length of the studied site is 30 metres. The site is naturally divided into two parts by the central sampling point, the ecotone. The soil samples were taken 5 cm beneath the surface.

The main goal here is to compare the chemical properties (KCl pH, H_2O pH, and the percentage of organic carbon) of the field (A) and the forest (B). Using the state-of-the-art permutation tests, the problem would be dealt with by (i) randomly reassigning the data to the two groups and (ii) comparing the distributions of the randomly reassigned samples to the distribution of the original ones. Under the null hypothesis that (A) and (B) do not differ the distributions of the two groups should not differ when the data are permuted. However, such permutation procedure grounds on the key assumption that the data are exchangeable under the null hypothesis. This assumption may clearly be violated in the case of spatial data, as they are likely to depend on their coordinates. Instead of permuting the data directly, a spatial trend is fitted to the data, and the estimated residuals from this model, assumed to be approximately exchangeable, are permuted instead. The model and the procedure are formally introduced in Sections 4.1 - 4.5 and the case study is detailed in Section 4.8.

1.3. The production of municipal waste in the Venice province

The third case study is based on a data from the field of tourism, collected in 49 cities in the Venice province during the years 1997 and 2011. The data include

the yearly records of the per capita amount of the municipal waste, which shall be modelled as the functional observations of time. This data set was used by [15], where the focus was on fitting a spatiotemporal regression model with differential regularization [14, 46]. We here focus on testing the significance of the covariates within a spatial functional regression model inspired by that of [15], but fitted by the generalized least-squares. To model the production of the waste, the spatial coordinates of the centre of each city is accounted for, as well as the proportion of the number of the beds in the accommodation facilities to the number of residents, to account for the tourism in each city.

Chapter 2

Overview of functional data analysis

2.1. Basic steps in functional data analysis

This section aims at introducing the basic concepts of functional data analysis. In the FDA framework, we assume that the observed discrete data are generated by some unknown function, i.e., [62]

$$x_g = y(t_g) + \xi_g, g \in \{1, \ldots, G\}, t \in T \subset [0, \infty), \tag{2.1}$$

where G is the number of discrete observations of the function $y(t)$ and ξ_g is the random error. Moreover, it is assumed that the underlying function is smooth, therefore the two adjacent discrete data points are likely to be similar to each other.

Assume now that $i \in \{1, \ldots, n\}$ functions were observed. Then, the i-th function $y_i(t)$ can be represented by a linear combination of the known basis functions and the unknown basis coefficients as [62]

$$y_i(t) = \sum_{k=1}^{K_{y,i}} c_{ki}\phi_{ki}(t) = \boldsymbol{\phi}_i'(t)\mathbf{c}_i, i \in \{1, \ldots, n\}, \tag{2.2}$$

where $K_{y,i}$ is the number of the basis coefficients of the i-th functional observation, $\phi_{ki}(t)$ are the known basis functions and c_{ki} are the unknown basis coefficients, which can be estimated, e.g., by the least-squares method. The symbol $'$ denotes the transposition.

In the following, only the B-spline basis system is detailed, as it allows for a great flexibility for fitting the nonperiodical functional data [62], being the most convenient choice for the real-world data introduced in Chapter 1 of this thesis.

To construct a B-spline basis system, the functional domain T is partitioned into τ subintervals by $\tau - 1$ knots. At each subinterval, a B-spline is a polynomial of order 4 (note that, a cubic spline is the most popular choice implying continuous second derivatives) with a non-zero property over 4 adjacent subintervals. An example of a B-spline basis with 17 basis functions is provided in Figure 2.1.

In general, $K_{y,i}$, the number of basis functions, helps to control the smoothness of the functional observation $y_i(t)$. This process is also referred to as the least-squares smoothing. When estimating curves from the raw observations, the goal is to reach a trade-off between the bias and the variance [62]. Under this setting, it can be achieved by choosing the number of the basis functions. The greater $K_{y,i}$ is, the better $y_i(t)$ fits the observed data, but at the same moment it increases the variance and can lead to overfitting. On contrary, the smaller $K_{y,i}$ is, the smoother and less noisy $y_i(t)$ becomes, but if the number of the basis functions becomes too low, important data features can be lost due to oversmoothing, and $y_i(t)$ may not give sufficient information about the observed data. An appropriate number of the basis functions can be found, e.g., by means of the generalized cross-validation [62].

A better control over the smoothness can be provided by the roughness penalties, where the roughness of a function can be quantified as its curvature, i.e., the square of the second derivative: $[D^2 y_i(t)]^2$. Then, the roughness can be measured by an integrated squared second derivative [62]

$$\mathrm{PEN}_2(y_i) = \int_T [D^2 y_i(t)]^2 dt. \tag{2.3}$$

Note that, in general, one can assume the ν-th power of the ν-th derivative, that is, $[D^\nu y_i(t)]^\nu, \nu \in \{1, 2, 3, \ldots\}$. Following Equation (2.2), Equation (2.3) can be rewritten as $\mathbf{c}_i' \mathbf{R}_{y,i} \mathbf{c}_i$, where $\mathbf{R}_{y,i}$ represents the roughness penalty matrix for the i-th function $y_i(t)$. The (j,k)-th entry of $\mathbf{R}_{y,i}$ is given as

$\mathbf{R}_{y,i_{(j,k)}} = \int_T D^2\phi_{ji}(t)D^2\phi_{ki}(t)dt$. An optimal fit to the observed discrete data is then found by minimization of the penalized residual sum of squares criterion [62]

$$\text{PENSSE}_{\lambda_i}(y_i|\mathbf{x_i}) = [\mathbf{x_i} - y_i(\mathbf{t})]'[\mathbf{x_i} - y_i(\mathbf{t})] + \lambda_{y,i}\mathbf{c}_i'\mathbf{R}_{y,i}\mathbf{c}_i, \tag{2.4}$$

where $\mathbf{x_i}$ is a set of discrete observations of $y_i(t)$ and $\lambda_{y,i} \geq 0$ is a smoothing parameter controlling the smoothness of $y_i(t)$, which can be selected, e.g., by the generalized cross-validation. Increasing $\lambda_{y,i}$ puts more emphasis on the smoothness of $y_i(t)$ and less on the closeness of its fit to the observed data points. Conversely, decreasing $\lambda_{y,i}$ is less and less penalizing the roughness, leading to a high variability of $y_i(t)$. Note that, in general, $\lambda_{y,i}$ can vary from function to function.

An important theorem can be found in the work of [20] stating that the function $y_i(t)$ which minimizes Equation (2.4) is a cubic spline with knots placed at the data points [62]. However, in the case of a very large number of the recorded data, placing a knot at each data point would become unfeasible. This is the case of, e.g., the analysis of floating car data (see the introduction in Section 1.1 and the case study in Section 3.6). Recall that, the driving speed was recorded with the frequency of four observations per second, obtaining on average lower thousands of raw data points per ride (the exact number depends on the length of each ramp). By that, the sequence of the knots for defining a B-spline basis may be initially chosen by an expert assessment to capture enough variation and data features. Then, a roughness penalty (2.4) can be used to obtain the smooth functional observations. Figure 2.1 provides an example of a cubic B-spline basis as described in this paragraph, as well as an example of the real-world observations and the smoothed functional data. To compare the different levels of smoothness of the functional data, two different sets of curves are shown based on different values of the smoothing parameter λ.

To take into account the whole data set, define the diagonal basis functions matrix $\mathbf{\Phi} = \text{diag}(\boldsymbol{\phi}_1'(t), \ldots, \boldsymbol{\phi}_n'(t))$, the set of all discrete observations $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$, and a symmetric block-diagonal matrix

$\mathbf{R}_y = \text{diag}(\lambda_{y,1}\mathbf{R}_{y,1}, \ldots, \lambda_{y,n}\mathbf{R}_{y,n})$. Then, the vector of the basis coefficients $\mathbf{c} = (\mathbf{c}_1', \ldots, \mathbf{c}_n')'$ can be estimated as [61]

$$\widehat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{\Phi} + \mathbf{R}_y)^{-1}\mathbf{\Phi}'\mathbf{x}. \tag{2.5}$$

Define further the total number of the basis coefficients $K_y = \sum_{i=1}^{n} K_{y,i}$, the K_y-dimensional column vector $\boldsymbol{\phi} = (\boldsymbol{\phi}_1'(t), \ldots, \boldsymbol{\phi}_n'(t))'$ and the $n \times K_y$ matrix of the basis coefficients $\widehat{\mathbf{C}} = \text{diag}(\widehat{\mathbf{c}}_1', \ldots, \widehat{\mathbf{c}}_n')$. Then, one can write $\widehat{\mathbf{y}} = \widehat{\mathbf{C}}\boldsymbol{\phi}$.

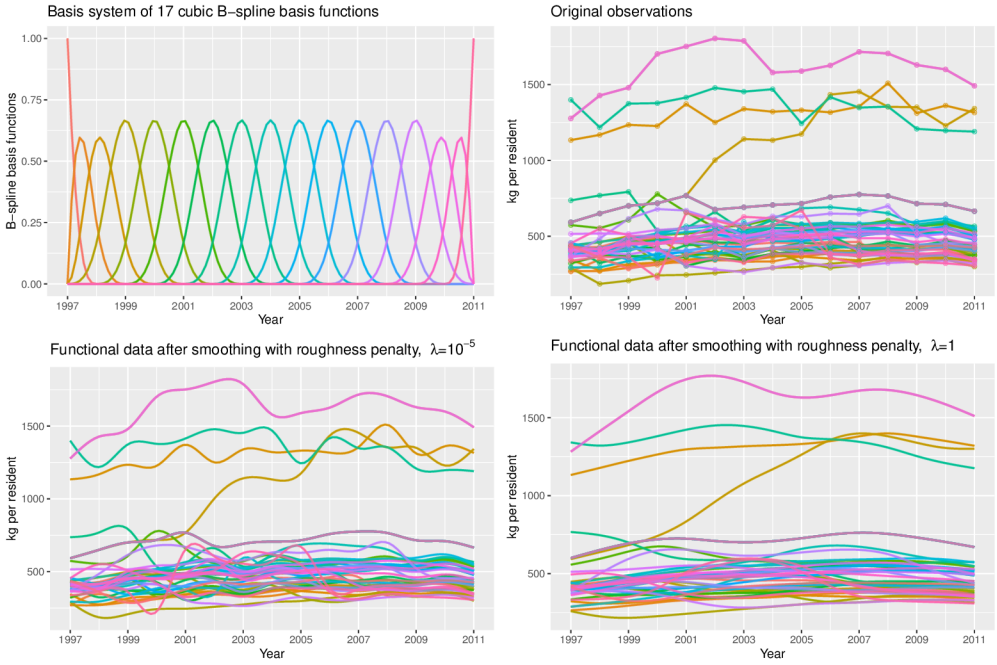


Figure 2.1: Top left: A view of the cubic B-spline basis system with 17 basis functions. Top right: Raw observations, for clarity connected by lines. Bottom: Functional data after smoothing using the roughness penalty approach with smoothing parameter $\lambda = 10^{-5}$ (bottom left) and $\lambda = 1$ (bottom right). For further details regarding this data, see Section 4.9.

2.2. Basic definitions of functional data

Suppose now that a function $y_i(t)$ is a realization of a random function $Y_i(t)$. To describe and quantify the attributes of this function, we consider the following [43]. The random function $Y = \{Y(\omega, t), t \in T\}$, where T is the functional domain, is defined on a probability space Ω, that means, $\forall \omega \in \Omega, Y(\omega)$ is a deterministic function. The realizations $Y(\omega), \omega \in \Omega$, are assumed to be elements of $L^2(T)$ (or L^2 for short), a space of the square-integrable functions, i.e., $\forall \omega \in \Omega$, it holds [43]

$$\|Y(\omega)\|^2 = \int_T [Y(\omega, t)]^2 dt < \infty, \tag{2.6}$$

where $\|\cdot\|$ represents the L^2-norm.

Define now the mean function as [43]

$$\mu(t) = E[Y(t)], t \in T, \tag{2.7}$$

the variance function as [43]

$$\text{var}(t) = E[Y(t) - \mu(t)]^2, t \in T, \tag{2.8}$$

and the covariance function as [43]

$$\text{cov}(t_1, t_2) = E\{[Y(t_1) - \mu(t_1)][Y(t_2) - \mu(t_2)]\}, t_1, t_2 \in T. \tag{2.9}$$

Assume now that $y_1(t), \ldots, y_n(t)$ is a sample of n functions. Define the point-wise estimators where

$$\bar{y}(t) = \frac{1}{n} \sum_{i=1}^{n} y_i(t), t \in T, \tag{2.10}$$

is the point-wise sample mean function [43, 62] and

$$\widehat{\text{var}}(t) = \frac{1}{n-1} \sum_{i=1}^{n} [y_i(t) - \bar{y}(t)]^2, t \in T, \tag{2.11}$$

is the point-wise sample variance function [62]. One also can define the point-wise sample standard deviation function [62] as $\sqrt{\widehat{\text{var}}[y(t)]}$.

To describe the dependence of the observations across different values of the functional domain T, let

$$\widehat{\text{cov}}(t_1, t_2) = \frac{1}{n-1} \sum_{i=1}^{n} [y_i(t_1) - \bar{y}(t_1)][y_i(t_2) - \bar{y}(t_2)], t_1, t_2 \in T, \tag{2.12}$$

be the point-wise sample covariance function [43, 62].

Chapter 3

Regression models for functional covariates and functional response under homoscedasticity

The driving speed observations introduced in Section 1.1 can be viewed as the functions of distance travelled from the starting point of the ramp, and the methods specified in Chapter 2 can be applied to represent the observed floating car data as the functional observations of the driving speed. Then, these can be analysed, and a relationship between the functional observations of the driving speed and the explanatory variables can be examined by the functional-on-functional regression model. Under this setting, the estimates of the regression functions $\beta(t)$ will also be functions over the same functional domain T as the functional observations $y_i(t)$. Throughout this chapter, the functional domain T will represent the distance from the starting point of the ramp. This allows, e.g., to see where and how is the driving speed changing over T. Also, as $\beta(t)$ are functions of T, the regression model can describe the relationship between the dependent and the explanatory variables in a way which allows to assess the effect of the explanatory variable at a specific part of T, not being evaluated just as an overall information.

In the first part of this chapter, the methods for the functional-on-functional regression model are summarized in Sections 3.1 - 3.3, together with a permutation-based testing procedure suitable for this setting (Section 3.4). Then, the

interval-wise testing procedure for the comparison of the means of the two functional populations, allowing to detect the dissimilarity between the curves of speed in the merging, the exit, and the central lanes, is recapitulated as well (Section 3.5). Lastly, a real-world example from the field of transportation research is presented in Section 3.6.

3.1. Model specification

Assume that a sample of functions $y_i(t), i \in \{1, \ldots, n\}$, has been observed. In the following, $y_i(\cdot)$ represents an element of the Hilbert space of the square-integrable functions $L^2(T)$. We further assume that the functional observations can be represented through the linear model [1]

$$y_i(t) = \beta_0(t) + \sum_{\ell=1}^{L} \beta_\ell(t) f_{i\ell}(t) + \varepsilon_i(t), t \in T, i \in \{1, \ldots, n\}, \tag{3.1}$$

where $f_{i\ell}(t), \ell \in \{1, \ldots, L\}, t \in T$, are the known covariates, possibly scalar, categorical, or functional, and $\beta_\ell(t), \ell \in \{0, 1, \ldots, L\}, t \in T$, are the unknown regression functions. The random errors $\varepsilon_i(t), i \in \{1, \ldots, n\}, t \in T$, are independent, identically distributed random functions with zero mean and constant variance, that is, the variance-covariance matrix is $\mathbf{\Sigma} = \sigma^2 \mathbf{I}$, where σ^2 is an unknown parameter and $\mathbf{I}$ is an $n \times n$ identity matrix.

Using a matrix notation, model (3.1) can be rewritten into the form [62]

$$\mathbf{y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $\mathbf{y} = (y_1(t), \ldots, y_n(t))', t \in T$, are the functional observations, $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_L(t))', t \in T$, are the unknown regression functions, and $\boldsymbol{\varepsilon} = (\varepsilon_1(t), \ldots, \varepsilon_n(t))', t \in T$, are the random error functions. A known $n \times (L+1)$ functional design matrix $\mathbf{F}$ can be symbolically expressed through a common notation with the i-th row $(1, f_{i1}(t), \ldots, f_{iL}(t)), t \in T$, where $f_{i\ell}(t), \ell \in \{1, \ldots, L\}$, are the known covariates related to the i-th observation. As the covariates can in general be functions, the element $f_{i\ell}(t)$ of the design matrix would in that case represent the functional covariate.

3.2. Estimation of the regression functions

Express now the unknown regression functions through the basis expansion [61, 62]

$$\beta_\ell(t) = \sum_{k=1}^{K_{\beta,\ell}} b_{k\ell}\theta_{k\ell}(t) = \boldsymbol{\theta}'_\ell(t)\mathbf{b}_\ell, \ell \in \{0, 1, \ldots, L\}, \tag{3.3}$$

where $K_{\beta,\ell}$ is the number of the basis coefficients of the ℓ-th regression function, $\theta_{k\ell}(t)$ are the known basis functions and $b_{k\ell}$ are the unknown coefficients. In this way, the problem of estimating the functions $\beta_\ell(t)$ reduces to estimating the coefficients $b_{k\ell}$ by, e.g., the least-squares method.

In order to express the model (3.2) with respect to the basis expansion (3.3), define the total number of the basis coefficients of the regression functions, i.e., $K_\beta = \sum_{\ell=0}^{L} K_{\beta,\ell}$, and stack the vectors $\mathbf{b}_\ell$ vertically to obtain a K_β-dimensional column vector $\mathbf{b} = (\mathbf{b}'_0, \mathbf{b}'_1, \ldots, \mathbf{b}'_L)'$. By defining an $(L+1) \times K_\beta$ matrix $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{\theta}'_1(t), \ldots, \boldsymbol{\theta}'_L(t))$, model (3.2) can be rewritten as [62]

$$\mathbf{y} = \mathbf{F}\boldsymbol{\Theta}\mathbf{b} + \boldsymbol{\varepsilon}, \tag{3.4}$$

and then can be formally transformed to a *constant coefficient linear model* [62] by defining an $n \times K_\beta$ matrix $\tilde{\mathbf{F}} = \mathbf{F}\boldsymbol{\Theta}$ as

$$\mathbf{y} = \tilde{\mathbf{F}}\mathbf{b} + \boldsymbol{\varepsilon}. \tag{3.5}$$

Note that, by performing this step, we moved from the estimation of $L+1$ regression functions in model (3.2) to the estimation of K_β scalar coefficients in model (3.4).

Following the general assumption that the roughness penalty term $\text{PEN}_2(\beta_\ell) = \int_T [D^2\beta_\ell(t)]^2 dt$ can be different for each regression function $\beta_\ell, \ell \in \{0, 1, \ldots, L\}$, one can, following Equation (3.3), write $\text{PEN}_2(\beta_\ell) = \mathbf{b}'_\ell \mathbf{R}_{\beta,\ell}\mathbf{b}_\ell$, where $\mathbf{R}_{\beta,\ell} = \int_T D^2\boldsymbol{\theta}_\ell(t) D^2\boldsymbol{\theta}'_\ell(t) dt$ represents the roughness penalty matrix. Then, a symmetric block-diagonal $K_\beta \times K_\beta$ matrix $\mathbf{R}_\beta = \text{diag}(\lambda_{\beta,0}\mathbf{R}_0, \ldots, \lambda_{\beta,L}\mathbf{R}_L)$, where $\lambda_{\beta,\ell}, \ell \in \{0, 1, \ldots, L\}$, are the smoothing parameters, can be created. The penalized least-squares estimator of the

regression coefficients $\mathbf{b}$ is then defined as [62]

$$\widehat{\mathbf{b}} = \left[\int_T \tilde{\mathbf{F}}'(t)\tilde{\mathbf{F}}(t)dt + \mathbf{R}_\beta \right]^{-1} \int_T \tilde{\mathbf{F}}(t)'\mathbf{C}\boldsymbol{\phi}(t)dt. \tag{3.6}$$

Then, $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_L(t))', t \in T$, can be estimated as $\widehat{\boldsymbol{\beta}} = \boldsymbol{\Theta}\widehat{\mathbf{b}}$.

To characterize the precision of the estimator $\widehat{\boldsymbol{\beta}}$ and to obtain the point-wise confidence intervals for the regression functions $\beta_\ell(t), \ell \in \{0, 1, \ldots, L\}$, consider the following substitution [62]

$$\mathbf{A} = \left[\int_T \tilde{\mathbf{F}}(t)'\tilde{\mathbf{F}}(t)dt + \mathbf{R}_\beta \right]^{-1} \int_T \boldsymbol{\phi}'(t) \otimes \tilde{\mathbf{F}}'(t)dt, \tag{3.7}$$

where the symbol $\otimes$ denotes the Kronecker product. Then, the estimator of the regression coefficients can be written as $\widehat{\mathbf{b}} = \mathbf{A}\mathbf{c}$, where $\mathbf{c}$ is a vector of the basis coefficients (see Equation (2.5)). Then, the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be estimated as [61]

$$\widehat{\mathrm{var}}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2\boldsymbol{\Theta}\mathbf{A}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\mathbf{A}'\boldsymbol{\Theta}', \tag{3.8}$$

where [66]

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|\kappa_i(t)\|^2. \tag{3.9}$$

Here, the symbol $\|\cdot\|$ represents the L^2-norm and $\boldsymbol{\kappa} = \widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}$ is the vector of functional residuals. The approximate 95% point-wise confidence intervals for $\beta_\ell(t), \ell \in \{0, 1, \ldots, L\}$, are then the intervals with the limits $\widehat{\beta}_\ell(t) \pm 2\widehat{\mathrm{se}}[\widehat{\beta}_\ell(t)]$, where $\widehat{\mathrm{se}}[\widehat{\beta}_\ell(t)] = \widehat{\sigma}\sqrt{\{\boldsymbol{\Theta}\mathbf{A}(\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\mathbf{A}'\boldsymbol{\Theta}'\}_{\ell,\ell}}$ is an estimated standard error of the estimator $\widehat{\beta}_\ell(t)$. Note that, in general, the coverage probability of such confidence intervals may differ from 95%.

3.3. Model validation

The accuracy of regression models can be assessed by the (normalized) root mean squared error (N)RMSE

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|y_i(t) - \widehat{y}_i(t)\|^2}, \ \mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\|\bar{y}(t)\|}, \tag{3.10}$$

where $\widehat{y}_i(t)$ are the fitted values and $\bar{y}(t) = \frac{1}{n}\sum_1^n y_i(t)$ is the functional mean. The R^2 (the coefficient of determination)

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^n \|y_i(t) - \widehat{y}_i(t)\|^2}{\sum_{i=1}^n \|y_i(t) - \bar{y}(t)\|^2} \tag{3.11}$$

characterizes how accurately a model explains a phenomenon. Note that, the model selection can be performed, e.g., by the backward elimination, or by the generalized cross-validation [62].

3.4. The global permutation tests for the effect of the covariates

This section focuses on testing for the significance of the regression functions in model (3.2). Most linear hypotheses about the effect of the covariates can be expressed in a compact matrix form as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}, \text{ against } H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}, \tag{3.12}$$

where $\mathbf{C}$ is an $h \times (L+1)$ matrix of constants and $1 \leq h \leq L+1$ is a number of linearly independent hypotheses. An important special case, helping to assess the model, is the omnibus test that at least one covariate has a significant effect on the response variable. Symbolically, in model (3.2), we aim at testing the hypothesis

$$H_0 : \beta_1(t) = \beta_2(t) = \ldots = \beta_L(t) = 0 \ \forall t \in T, \text{against} \tag{3.13}$$

$$H_1 : \beta_\ell(t) \neq 0 \text{ for some } \ell \in \{1, \ldots, L\} \text{ and some } t \in T. \tag{3.14}$$

Using the notation introduced in Section 3.2, the omnibus hypothesis can be equivalently written as

$$H_0 : \mathbf{b}_1 = \mathbf{b}_2 = \ldots = \mathbf{b}_L = 0, \text{ against } H_1 : \mathbf{b}_\ell \neq 0 \text{ for some } \ell \in \{1, \ldots, L\}, \tag{3.15}$$

or in a matrix form as

$$H_0 : \boldsymbol{\mathcal{C}}\mathbf{b} = \mathbf{0}, \text{ against } H_1 : \boldsymbol{\mathcal{C}}\mathbf{b} \neq \mathbf{0}, \tag{3.16}$$

where $\boldsymbol{\mathcal{C}}$ is a $(\sum_{\ell=1}^{L} K_{\beta,\ell}) \times K_\beta$ block matrix of the form $\boldsymbol{\mathcal{C}} = (\mathbf{0}|\mathbf{I})$, where $\mathbf{0}$ is a $\sum_{\ell=1}^{L} K_{\beta,\ell} \times K_0$-dimensional matrix of zeros and $\mathbf{I}$ is a $(\sum_{\ell=1}^{L} K_{\beta,\ell} \times \sum_{\ell=1}^{L} K_{\beta,\ell})$ identity matrix. To verify the omnibus hypothesis, one can use the test statistic

$$\mathcal{T} = \frac{1}{\widehat{\sigma}^2} \widehat{\mathbf{b}}' \boldsymbol{\mathcal{C}}' \left[\boldsymbol{\mathcal{C}} (\tilde{\mathbf{F}}' \tilde{\mathbf{F}} + \mathbf{R}_\beta)^{-1} \boldsymbol{\mathcal{C}}' \right]^{-1} \boldsymbol{\mathcal{C}} \widehat{\mathbf{b}}. \tag{3.17}$$

The test for the significance of a single regression function $\beta_\ell(t), \ell \in \{1, \ldots, L\}$, through the hypothesis

$$H_0^\ell : \mathbf{b}_\ell = 0, \text{ against } H_1^\ell : \mathbf{b}_\ell \neq 0, \tag{3.18}$$

can be run using the test statistic

$$\mathcal{T}_\ell = \frac{1}{\widehat{\sigma}^2} \widehat{\mathbf{b}}_\ell' (\tilde{\mathbf{F}}_\ell' \tilde{\mathbf{F}}_\ell + \mathbf{R}_l) \widehat{\mathbf{b}}_\ell, \tag{3.19}$$

where $\tilde{\mathbf{F}}_\ell$ denotes the block-diagonal part of $\tilde{\mathbf{F}}$ corresponding to $\mathbf{b}_\ell$.

In this section, only the omnibus and the single-hypothesis tests are concerned. The case of the more general linear hypotheses is detailed in Section 4.4.

A decision about the hypotheses (3.16) and (3.18) can be made by the Freedman and Lane permutation scheme (see [31]), adapted to the functional data setting [1, 66]. Algorithm 1 provides a detailed overview of the scheme for testing the hypothesis (3.16). Similarly, one can proceed in the case of a single-covariate hypothesis (3.18). This procedure allows one to estimate the distribution of the test statistics $\mathcal{T}$ (or $\mathcal{T}_\ell$) under the random permutations by Monte Carlo techniques. In particular, the residuals $\widehat{\boldsymbol{\varepsilon}}$ of model (3.2) are estimated from the

reduced model (i.e., the model fitted under the null hypothesis) and they are permuted at each iteration, each time creating a permuted data set over which the test statistic is evaluated. Under H_0, the distribution of $\mathcal{T}$ (or $\mathcal{T}_\ell$) is asymptotically invariant to permutations. The asymptotical invariance is a consequence of performing the test on the estimated and therefore asymptotically exchangeable residuals. Finally, one can compute the global p-value of the $\mathcal{T}$ (or $\mathcal{T}_\ell$) test as the proportion of permutations leading to a value of the $\mathcal{T}$ (or $\mathcal{T}_\ell$) statistic that is higher than or equal to the statistic from the observed data.

Algorithm 1 Freedman and Lane permutation scheme for the hypothesis (3.16).

1. Compute the test statistic $\mathcal{T}_{\text{data}}$ from the original data.
2. During the e-th permutation, $e \in \{1, \ldots, \mathcal{E}\}$,
   (a) Determine the residuals $\widehat{\varepsilon}_{i,r}$ of the reduced model $y_{i,r}(t) = \boldsymbol{\theta}_0'(t)\mathbf{b}_0 + \varepsilon_{i,r}(t)$. Obtain $\widehat{\mathbf{b}}_0, \widehat{\varepsilon}_{i,r}(t) = y_{i,r}(t) - \boldsymbol{\theta}_0'(t)\widehat{\mathbf{b}}_0$, $i \in \{1, \ldots, n\}$.
   (b) Permute the residuals $\widehat{\varepsilon}_{1,r}(t), \ldots, \widehat{\varepsilon}_{n,r}(t)$, obtaining $\widehat{\varepsilon}^*_{i,r}(t)$, $i \in \{1, \ldots, n\}$.
   (c) Compute the permuted responses $y^*_{i,r}(t)$ through the fitted reduced model and the permuted residuals $\widehat{\varepsilon}^*_{i,r}(t)$ as $y^*_{i,r}(t) = \boldsymbol{\theta}_0'(t)\widehat{\mathbf{b}}_0 + \widehat{\varepsilon}^*_{i,r}(t), i \in \{1, \ldots, n\}$.
   (d) Estimate the regression coefficients $\mathbf{b}$ of the full model from the permuted responses $\mathbf{y}^*_r$ by the OLS (Equation (3.6)). Obtain $\widehat{\mathbf{b}}^*$, $\widehat{\varepsilon}^*_{i,p}(t), i \in \{1, \ldots, n\}$.
   (e) Estimate the parameter σ^2_p from the residuals $\widehat{\varepsilon}^*_{i,p}(t)$, $i \in \{1, \ldots, n\}$, from step (d), obtaining $\widehat{\sigma}^2_p$
   (f) Compute the test statistic $\mathcal{T}^*_e$ as $\mathcal{T}^*_e = \frac{1}{\widehat{\sigma}^2_p}\widehat{\mathbf{b}}^{*\prime}\boldsymbol{\mathcal{C}}'\left[\boldsymbol{\mathcal{C}}(\tilde{\mathbf{F}}'\tilde{\mathbf{F}} + \mathbf{R}_\beta)^{-1}\boldsymbol{\mathcal{C}}'\right]^{-1}\boldsymbol{\mathcal{C}}\widehat{\mathbf{b}}^*$.
3. Compute the global p-value of the $\mathcal{T}$ test as the proportion of all permutations for which $\mathcal{T}^*_e \geq \mathcal{T}_{\text{data}}, e \in \{1, \ldots, \mathcal{E}\}$.

3.5. Detecting the differences in the means of two functional populations

The presence of the auxiliary lanes in the interchange introduced in Section 1.1 has motivated the task of identifying where exactly does the mean speed of the vehicles driving in the main lane and in the auxiliary lane differ, and where it cannot be distinguished. For this purpose, we apply the interval-wise permutation test proposed in the work of [60], allowing to identify the parts of the functional domain T where the two groups of the functional observations significantly differ.

Let $y_{1i}(t), i \in \{1, \ldots, n_1\}$, and $y_{2i}(t), i \in \{1, \ldots, n_2\}$, denote the functional observations from the two groups. Let $\mu_1(t)$ and $\mu_2(t)$ be their functional means and let $\mathcal{I} \subseteq T$ be an arbitrary interval $(t_1, t_2) \subseteq T$ or its complement $T \backslash (t_1, t_2)$. The aim is, for each $\mathcal{I}$, to test the hypothesis [60]

$$H_0 : \mu_1^{\mathcal{I}}(t) = \mu_2^{\mathcal{I}}(t) \; \forall t \in \mathcal{I}, \text{ against } H_1 : \mu_1^{\mathcal{I}}(t) \neq \mu_2^{\mathcal{I}}(t) \text{ for some } t \in \mathcal{I}, \tag{3.20}$$

using the test statistic [60]

$$\mathrm{T}^{\mathcal{I}} = \frac{1}{|\mathcal{I}|} \int_{|\mathcal{I}|} [\bar{y}_1(t) - \bar{y}_2(t)]^2 dt, \tag{3.21}$$

where $\mu_m^{\mathcal{I}}(t)$ denotes $\mu_m(t)$ restricted on the sub-interval $\mathcal{I}$ of T, $|\mathcal{I}|$ is the length of $\mathcal{I}$ and $\bar{y}_m(t)$ is the sample functional mean of the m-th group, $m \in \{1, 2\}$.

The permutation test described in the work of [38] and a test statistic (3.21) can be used to compute the p-value $p^{\mathcal{I}}$ of the $\mathrm{T}^{\mathcal{I}}$ test (3.21). For some $\mathcal{I}$, given two random independent samples $\mathcal{Y}_1^{\mathcal{I}} = y_{1i}^{\mathcal{I}}(t)$ and $\mathcal{Y}_2^{\mathcal{I}} = y_{2i}^{\mathcal{I}}(t)$, denote by $\mathcal{P}^{\mathcal{I}}$ the pooled sample $\mathcal{Y}_1^{\mathcal{I}} \cup \mathcal{Y}_2^{\mathcal{I}}$. For $e \in \{1, \ldots, \mathcal{E}\}$, where $\mathcal{E}$ is the total number of permutations, the test statistics $\mathrm{T}_1^{\mathcal{I}*}, \ldots, \mathrm{T}_{\mathcal{E}}^{\mathcal{I}*}$ are computed by randomly resampling n_1 values from $\mathcal{P}^{\mathcal{I}}$, assigning them to $\mathcal{Y}_1^{\mathcal{I}}$, while the remaining n_2 values are assigned to $\mathcal{Y}_2^{\mathcal{I}}$. Then, the p-value $p^{\mathcal{I}}$ is calculated as the proportion of all permutations for which $\mathrm{T}_e^{\mathcal{I}*} \geq \mathrm{T}_{\text{data}}^{\mathcal{I}}, e \in \{1, \ldots, \mathcal{E}\}$. Finally, define for each $t \in T$ the interval-wise adjusted p-value $\widetilde{p}(t)$ of the test (3.21) as [60]

$$\widetilde{p}(t) = \sup_{t \in \mathcal{I}} p^{\mathcal{I}}, t \in T. \tag{3.22}$$

The details of the computation of the interval-wise adjusted p-value (Equation (3.22)) can be found in [60].

3.6. Application: Analysis of the driving speed on expressway ramps based on floating car data

In this section, the methodology introduced in Chapter 2 and Sections 3.1 - 3.5 is applied to floating car data introduced in Section 1.1. All the steps which are necessary to perform the analysis, starting with preprocessing the raw observations, proceeding to testing for the significance in a functional-on-functional regression model or to detecting differences in the means of two groups of functional data, are detailed here.

3.6.1. Data preprocessing

Prior to proceeding to the analysis of the functional data, the raw observations obtained from the on-board GPS units were preprocessed as follows. Based on FCD, the area of interest was determined by a polygon function. Given our focus on the specific sections of the ramp, the data were trimmed in its immediate vicinity and the directions were deducted based on the GPS coordinates and the recorded time of each ride. The detailed characteristics of the specific sections of the ramp were detected as well (e.g., GPS coordinates of the road centres and their cumulative distance). The trimmed data respect the same start and end of each ride. Moreover, the rides distinguish the specific driving directions. Due to different recording frequencies that affected the final cumulative distance, additional cleaning of the data was considered. Atypical rides, e.g., clear errors or unrealistic accelerations in the source data, were omitted.

To achieve a higher accuracy of the recorded data, it was necessary to transform the WGS84 [degrees, EPSG:4326] coordinates to the planar S-JTSK [m, EPSG:5514] coordinates. The latter are standardly used coordinates for the data collected in the Czech Republic. This transformation can be conducted using the `sf_project` function (`R` package `sf` [56]).

Finally, the identification of centres and breakpoints of the curvature of the monitored ramp was needed. The least-square-circle fit method [18] was applied to determine the curvature, i.e., the reciprocal value of the circular radius for each ride with respect to the central lane, the merging lane, and the exit lane. For this purpose, the `lsfit.circle` function can be used (`R` package `circular` [2]). According to [7], a section with a circular radius greater than 2000 m is considered a tangent (straight) section. Lower values of the circular radius indicate the circular curve. For ramp 2, the process is illustrated in Figure 3.1, having identified six breakpoints of the curvature.

The potential presence of the spiral (transition) curves between a circular curve and a tangent section was not considered here due to the complexity of the determination of the parameters of the spiral curve from the point data, compared to the relatively simple determination of the parameters of the circular curve. In addition, several authors (see, e.g., [32, 55, 57]) found out that the transition curves do not significantly influence the speed, and had disregarded them in their studies. Following these findings, the spiral curves were not included into this analysis either.

3.6.2. Floating car data as the functional observations

In the first step, the data were cleaned from the rides with nearly zero speed, where there presumably was a congestion, as only the standard rides are of interest. Then, using the `R` package `FDA` [63] the data were turned into functional observations using a projection over a cubic B-spline basis with 100 basis functions and equispaced knots, preserving enough variation of the original discrete data, yet not being too noisy. The smoothing was performed using the penalized residual sum of squares criterion, where, for each ramp, an appropriate value of the smoothing parameter λ was selected via the generalized cross-validation [62]. The summary of the final data set available for the analysis is provided in Table 3.1. To identify the sections with a rapid change of the speed, the observations themselves and their first derivatives are studied, allowing to analyse the

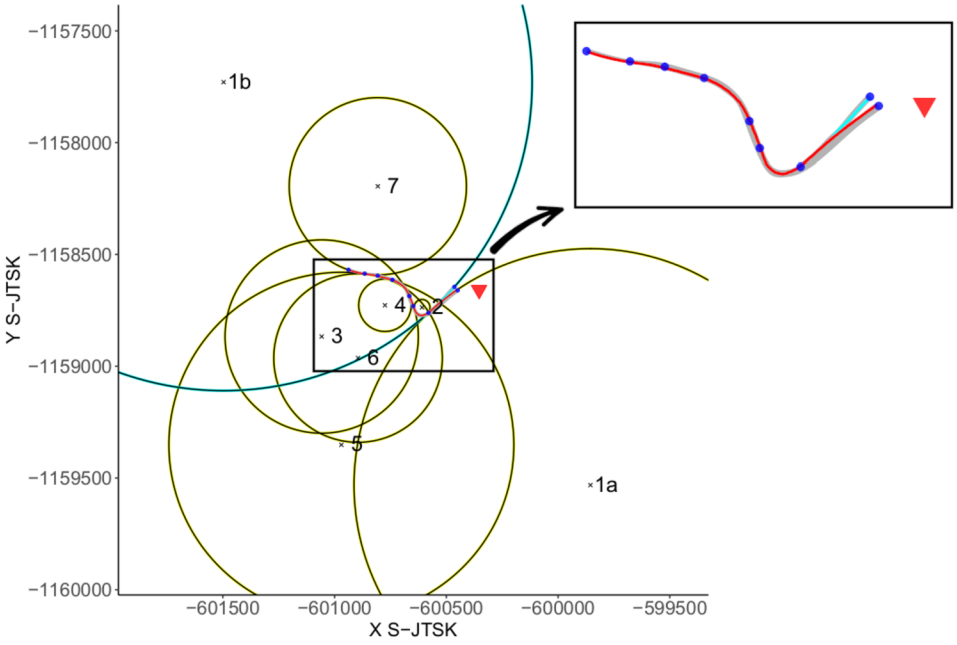


Figure 3.1: A circular radii for ramp 2 (the central lane by red, the merging lane by cyan), obtained from the least-square-circle fit method. Blue points denote the breakpoints of the ramp curvatures. The fitted yellow circles correspond to the central lane, the blue circle corresponds to the merging lane. The red triangle indicates the beginning of ramp 2. Upper-right corner: detailed view of ramp 2.

driving speed and the acceleration, respectively. To model the driving speed on each ramp, a set of potential covariates is selected, including the auxiliary lane, i.e., merging/exit lane (indicating from which direction the vehicle joined/left the ramp, 1 for the auxiliary lane, 0 otherwise) and the radius of the circle fitted to the particular part of the ramp (in metres), measuring its curvature.

3.6.3. Basic characteristics of the driving speed

In this section, to focus on the smoothness of the functional observations of the driving speed on each ramp, we study the observations of the driving speed and their first derivatives. The behaviour of the functions of the driving speed on the less complex (i.e., relatively straight) ramps 1 and 3 has only minimal fluctuations (Figures 3.2a and 3.2c). In contrast, the driving speed and the acceleration on

Table 3.1: ramp ID, ramp length (m), number of the observations from the central, merging and exit lane and the total number of the observations per ramp. Symbol - denotes the absence of the merging/exit lane in the ramp.

ramp ID	ramp length	$n_{\text{central lane}}$	$n_{\text{merging lane}}$	$n_{\text{exit lane}}$	n_{total}
1	395	473	9	-	482
2	653	365	40	-	405
3	386	684	-	23	707
4	839	168	-	-	168
5	888	309	-	18	327
6	604	141	-	-	141

the more complex loop ramps 2, 4, 5, and 6 show higher variability (Figures 3.2b, 3.3a, 3.3b, and 3.3c).

The behaviour of the vehicles driving in the auxiliary lanes is also of interest. The local conditions, including the speed limits, need to be considered in order to provide the correct interpretation.

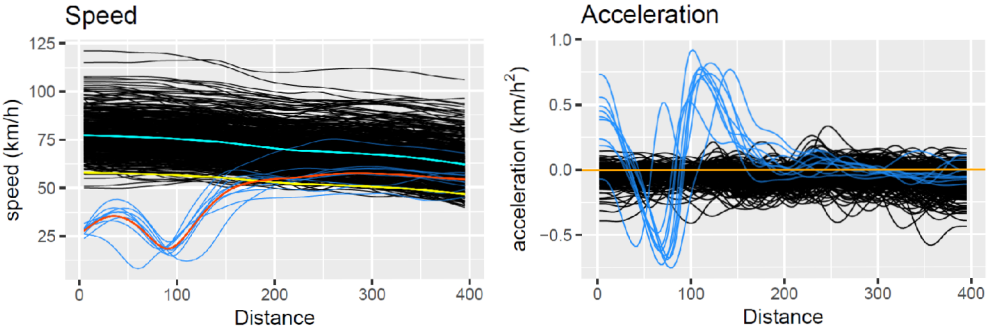
Firstly, consider merging into ramps 1 and 2, posted by a STOP sign located at 64.6th meter. The merging lane ends at 94.3th meter and the vehicles must stop before reaching this point. It can be concluded that the driving speed in the merging lane is the lowest at approximately 90th meter (Figure 3.2). Note that, although it is obliged to stop there, the minimum speed is above zero. This is in fact a consequence of smoothing the data with the purpose of reducing its variability and ensuring the overall smoothness of functions of the driving speed. After the STOP sign, a weaving section follows, where the exiting drivers are obliged to drive not faster than 40 km/h. Therefore, the merging drivers need to accelerate from the zero speed, while at the same time they need to decelerate in case of the necessary weaving manoeuvres. This creates turbulent patterns in both the speed and the acceleration (Figure 3.2), as evidenced also by the previous studies of the behaviour in the weaving sections [74]. The overall speed limit in the central lanes at ramps 1 and 2 is 60 km/h. Note that, the mean speed in the central lane at ramp 1 is around 80 km/h, exceeding the maximum allowed speed (Figure 3.2a). The driving patterns in the merging lanes at ramps 1 and 2 are further analysed in Section 3.6.4.

Secondly, consider the exit lanes from ramps 3 and 5. Although the exit is not posted by any speed limit sign, the speed should not exceed 50 km/h. From the driver's point of view, the radius of the curve is relatively small, which is probably the reason why the observed speed is reduced even below 40 km/h. The acceleration also drops significantly, especially in the case of ramp 3 (see Figure 3.3).

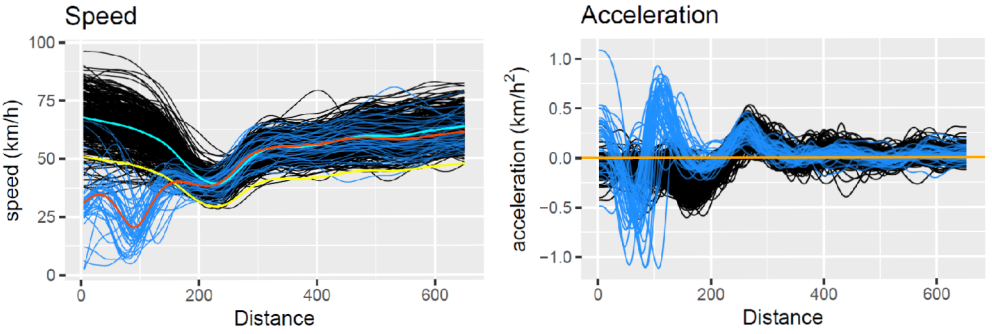
3.6.4. Examining the properties of the merging lanes

In order to maintain the safety and the smoothness of the traffic, it is recommended for the merging lane to be long enough to reach the speed of 3/4 of the speed limit for the central lane [73]. However, this assumption is in practice often fulfilled only in case of the roads which comply with the technical standards. In reality, these standards are often violated, e.g., by the presence of a STOP sign. For this reason, we suggest to approximate this recommendation by 3/4 of the mean speed of the vehicles driving in the central lane for the sake of examining the driving speed at merging lanes 1 and 2. Concluding from a visual inspection of the functions of speed, the vehicles from the merging lane at ramp 1 reach the 3/4 of the mean speed of the vehicles driving in the central lane at approximately 180th meter and, compared to the central lane, their mean speed is lower $\forall t \in T$ (Figure 3.2a).

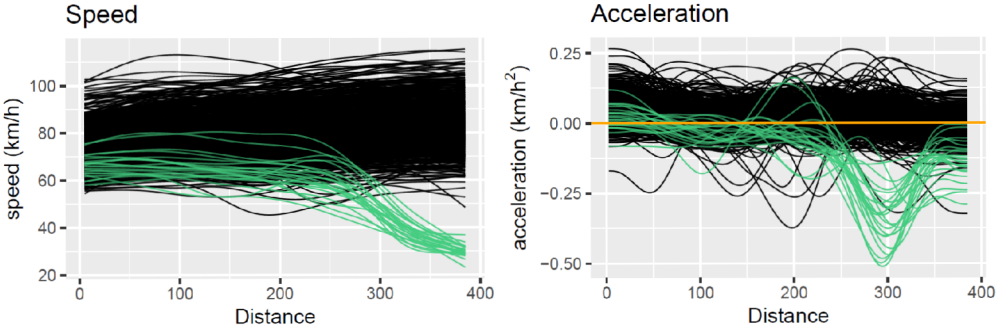
The vehicles driving in the merging lane at ramp 2 reach 3/4 of the mean speed of the vehicles driving in the central lane at approximately 150th meter and they quickly reach similar speed as the vehicles driving in the central lane (Figure 3.2b). Note that, these findings are based on descriptive statistics and visual inspection of the data. A more concise analysis regarding the (dis)similarity of the observations from the central and the merging lane is given in Section 3.6.5.



(a) Ramp 1

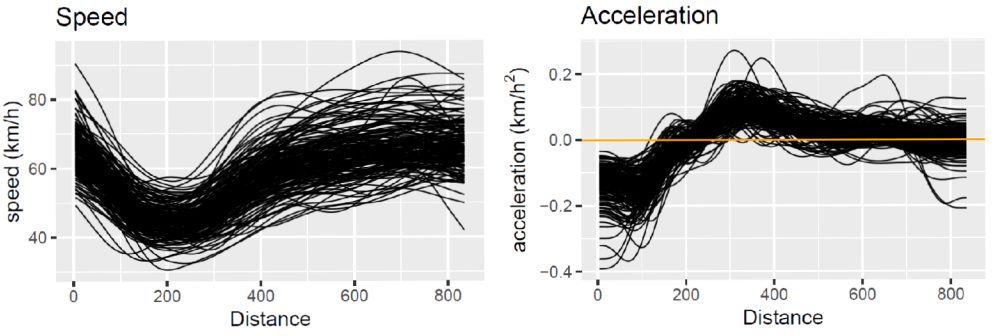


(b) Ramp 2

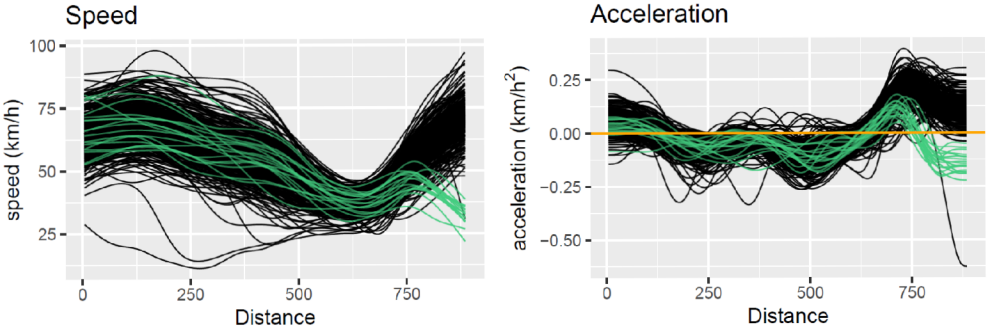


(c) Ramp 3

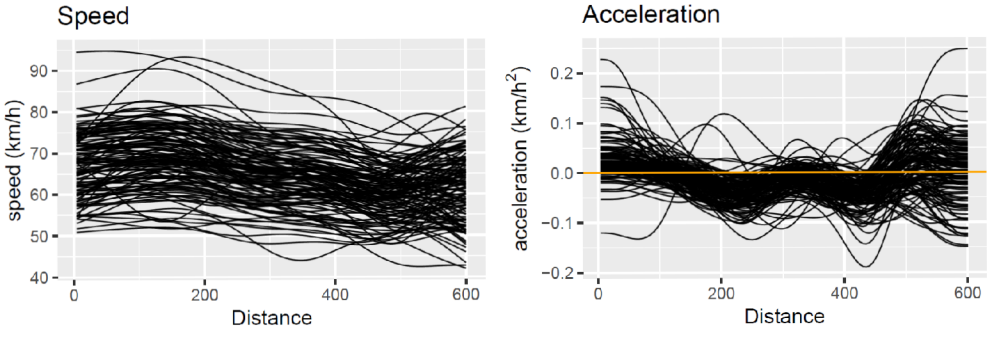
Figure 3.2: Left: observations of the speed in kilometres per hour after smoothing; right: acceleration in kilometres per hour squared. The black curves represent the observations from the central lane, the blue curves represent the observations from the merging lane, the green curves represent the observations from the exit lane. The cyan and yellow curves represent the mean of the speed and the 3/4 of the mean of the speed of the vehicles driving in the central lane, respectively. The red line represents the mean of the speed of the vehicles driving in the merging lane.



(a) Ramp 4



(b) Ramp 5



(c) Ramp 6

Figure 3.3: Left: observations of the speed in kilometres per hour after smoothing; right: acceleration in kilometres per hour squared. The black curves represent the observations from the central lane, the green curves represent the observations from the exit lane.

3.6.5. Differences in the means of the driving speed from two distinct lanes

In case of the ramps with both the central and the auxiliary lanes, it is of interest to focus on the specific parts of the ramp from/to which the observations from the central lane and from the auxiliary lane become indistinctible in terms of the driving speed. This can be analysed by testing the hypothesis of the equality of the means of the driving speed in the central and the auxiliary lane, applying the interval-wise testing procedure [60] (the methodology is discussed in Section 3.5), using the `R` package `fdatest` [59].

The test (3.21) for equality of the means of the driving speed for the central and the merging lane at ramp 2, based on 1000 permutations, leads to rejecting the null hypothesis (Figure 3.4b). To conclude, it takes approximately 180 metres for the vehicles driving in the merging lane to reach the same mean speed as the vehicles driving in the central lane. On the contrary, at ramp 1 (Figure 3.4a), the mean speed of the vehicles driving in the merging lane is, $\forall t \in T$, significantly lower than the mean speed of the vehicles driving in the central lane. In the case of ramp 3 (Figure 3.5a), the mean speed of the vehicles driving in the exit lane and in the central lane differs significantly $\forall t \in T$. At ramp 5 (Figure 3.5b), the mean speed of the vehicles driving in the exit lane is significantly lower, compared to the mean speed of the vehicles driving in the central lane, from approximately 800th meter until the end of the ramp.

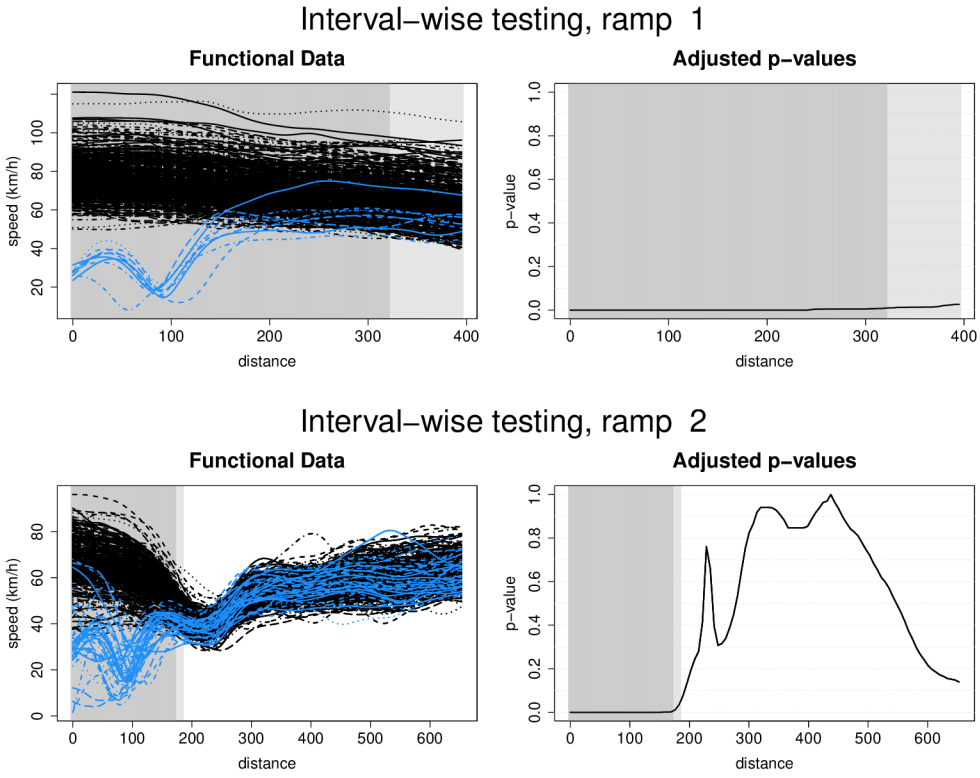


Figure 3.4: The results of testing the hypothesis (3.20) of equality of the means of the driving speed in the merging and the central lane. Left: The observations from the merging lane (blue) and the central lane (black). Right: The adjusted p-values of the test (3.20). Light grey: 0.05 level of significance, dark grey: 0.01 level of significance.

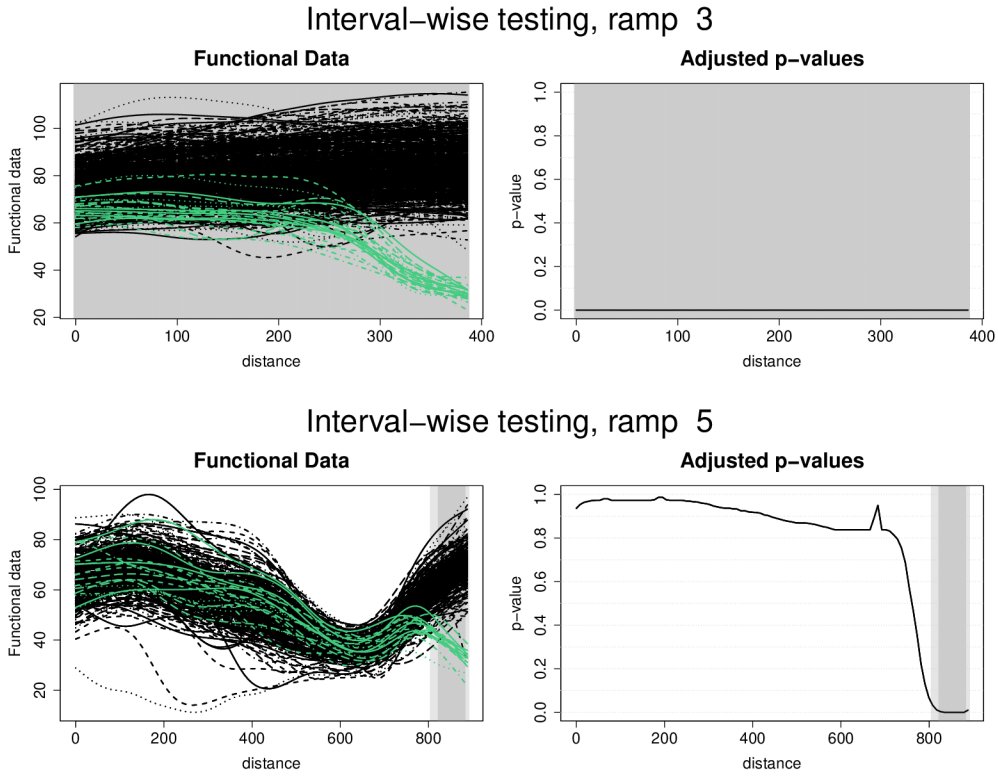


Figure 3.5: The results of testing the hypothesis (3.20) of equality of the means the driving of speed in the exit and the central lane. Left: The observations from the exit lane (green) and the central lane (black). Right: The adjusted p-values of the test (3.20). Light grey: 0.05 level of significance, dark grey: 0.01 level of significance.

3.6.6. Models for the driving speed data

The initial models chosen to describe the vehicle's driving speed include the functional covariate radius(t) and the factor covariate auxiliary lane (either the merging lane or the exit lane). Following Equation (3.2), the corresponding i-th row of the design matrix $\mathbf{F}$ can be symbolically expressed for the i-th functional observation of the driving speed as $(1, \text{radius}_i(t), \text{auxiliary lane}_i)$, from left to right representing the intercept, the function of the radius of the i-th observation, and a factor variable aux_i taking either the value 0 for the vehicle driving in the central lane, or 1 for the vehicle driving in the auxiliary lane. To find an optimal model, the backward elimination method can be used, where at each iteration, a covariate with the largest p-value is removed and the model is re-estimated. The procedure stops when only the significant covariates remain in the model. In all cases, 1000 permutations are performed, considering the level of significance $\alpha = 0.05$. The results for the final models for the global p-values of the $\mathcal{T}$ and $\mathcal{T}_\ell$ tests (Equations (3.17) and (3.19), respectively), as well as the model evaluation characteristics, can be seen in Table 3.2. In more detail, the normalized RMSE takes the values from 0.10 to 0.14 for all analysed ramps, which indicates a sufficiently high accuracy of the fitted models. On the other hand, small values of the R^2 indicate that the analysed covariates (the radius, and the auxiliary lane) explain the behaviour of the driving speed only poorly, indicating that more covariates would be needed to sufficiently model the driving speed. This is further discussed in Section 3.6.7.

Note that, the design matrix can be augmented by a column with an interaction term between the radius and the auxiliary lane allowing to model the effect of the curvature separately for the central and the auxiliary lane. In this case, the i-th row of the matrix $\mathbf{F}$ can be of the form $(1, \text{radius}_i(t), \text{auxiliary lane}_i, \text{auxiliary lane}_i \cdot \text{radius}_i(t))$. Nevertheless, the interaction term is not significant in any of the considered ramps, and therefore it is not presented here in more detail.

For each ramp, the resulting fitted speed and the estimates of the regression

Table 3.2: Left part of the table: The global p-values of the $\mathcal{T}$ and $\mathcal{T}_\ell$ tests for the final models based on 1000 permutations of the driving speed data. The symbol - means the covariate is not included in the final model. Right part of the table: Model evaluation. Note that, ramp 6 has not been included based on the results of the backward elimination in the initial model, as none of the covariates had a significant effect on the driving speed.

	Global p-values of the final models				Model evaluation		
ramp ID	Omnibus test	β_{radius}	$\beta_{\text{merging lane}}$	$\beta_{\text{exit lane}}$	R^2	RMSE	NRMSE
1	0	-	0	-	0.138	201	0.143
2	0	0.015	0	-	0.384	159	0.111
3	0	0	-	0	0.206	186	0.119
4	0.012	0.015	-	-	0.016	176	0.104
5	0	0	0.001	0	0.091	219	0.130
6	0.139	-	-	-	-	-	-

functions, together with the approximate 95% point-wise confidence intervals for $\widehat{\beta}(t)$, are presented in Figures 3.6 - 3.10.

Concern now the results of the regression models for each of the six ramps. Overall, the auxiliary lanes are indeed the important predictors of the driving speed. Note that, this can be observed in the regression models as well as in the interval-wise testing (Section 3.6.5), see ramps 1, 2, 3, and 5. For the more complex ramps, speaking in terms of their curvature (ramps 2, 4, 5, and 6), the radius has a significant effect on the driving speed for ramps 2, 4, and 5. Ramp 3 would be classified as a rather less complex ramp based on its curvature, although the effect of the radius is a significant predictor of the driving speed as well. The individual results of the models for the driving speed are discussed in more detail in the following part of this section.

Starting with the more complex ramps in terms of the curvature, the shape of $\beta_{\text{merging lane}}(t)$ at ramp 2 indicates that the effect of the merging lane on the driving speed is mainly in lowering the driving speed at first 100 m by 30-40 km/h, compared to the central lane. On approximately 100-250 m, the vehicles driving from the merging lane accelerate, and after passing this distance, the effect of the merging lane becomes only marginal. Secondly, the shape of $\beta_{\text{radius}}(t)$ suggests that the effect of this functional covariate manifests in a slight increase of the

driving speed $\forall t \in T$. However, the point-wise confidence interval, for some $t \in T$, includes also the negative values, concluding that there are sections of the ramp where the effect of the radius is not significant. Recall that, for rejecting the global null hypothesis it is sufficient that $\beta(t)$ is significantly different from 0 for at least some $t \in T$. This may indicate that the hypothesis is rejected at some parts of T and not rejected on the rest. As the test is of global nature, the interpretation should be very careful.

The effect of the radius is also significant for ramp 4 where its effect manifests in a slight decrease of the speed at approximately 220th meter, followed by similarly high increase of the driving speed after 250th meter. At approximately 250th meter the radius reaches its lowest values, i.e., the curvature of the road is the highest here (see Figures 1.1 and 3.9).

Regarding the loop ramp 5, the shape of $\beta_{\text{exit lane}}(t)$ suggests that at first 700 metres the speed of the vehicles leaving the ramp in the exit lane is comparable to those driving in the central lane. Then, as exiting the ramp, they quickly decelerate at the last 180 m, where their speed becomes almost 40 km/h lower than the speed of the vehicles driving in the central lane. The loop begins at approximately 500th meter, from where the effect of the radius causes a slight decrease of the driving speed, reaching its local minimum at approximately 630th meter where the curvature of the loop is the highest (i.e., the radius is the lowest), see Figures 1.1 and 3.10 for more details.

Assessing ramp 6, no significant relationship is found between the radius and the driving speed.

Concerning the less complex ramps in terms of the curvature, at ramp 1, the shape of $\beta_{\text{merging lane}}(t)$ shows that the speed of the vehicles driving in the merging lane is approximately 40-50 km/h lower at first 100 m, compared to the speed of the vehicles driving in the central lane. After passing this point the vehicles start to accelerate. The acceleration is most evident at 100-180 m; from there, the vehicles drive at almost constant speed, accelerating only mildly, although still being approximately 10 km/h slower compared to the vehicles driving in the

central lane. Overall, the shape of $\beta_{\text{merging lane}}(t)$ shows that the vehicles driving in the merging lane drive slower throughout the whole ramp, compared to the vehicles driving in the central lane (Figure 3.6). The effect of the radius is not significant.

Ramp 3 is located on the same road as ramp 1 but in the opposite direction. The final model for ramp 3 includes the exit lane and the radius. Concluding from the shape of $\beta_{\text{exit lane}}(t)$, the effect of the exit lane in the model is negative at the whole ramp and the vehicles are decelerating even more as t increases. Their driving speed changes only slightly at first 200 m, being 10-20 km/h slower compared to the vehicles driving in the central lane. After passing 200th meter, the drivers begin to decelerate sharply, reaching a minimum speed and, compared to the vehicles driving in the central lane, their mean speed is 60 km/h lower at the end of the ramp. The effect of the radius is significant here, causing a slight decrease of the speed starting from approximately 170th meter until the end of the ramp. Note that, although ramps 1 and 3 are located on the same road, their shapes are not symmetrical. Ramp 1 is more curved in the beginning and straighter in the end of the ramp; on contrary, ramp 3 starts with the straightest part and then becomes more curved in its second half. At ramp 3, the significance of the effect of the radius on the driving speed is shown by the global permutation test presented in Section 3.4. This difference between the curvature of ramps 1 and 3 is also evident from the behaviour of the functions of the radius (see Figures 3.6 and 3.8 for comparison).

The precision of the regression models for the driving speed can be assessed through the precision of the estimators of the regression functions using the pointwise confidence intervals (see Figures 3.6 - 3.10). Firstly, the precision of the estimates of $\beta_0(t)$ is discussed. The confidence interval at ramp 1 has almost the same width over the whole functional domain T (the margin of error is approximately 1 km/h). Similarly, the confidence interval for the estimate of $\beta_0(t)$ at ramp 3 has almost constant width (the margin of error is approximately 5 km/h) but is slightly narrower in the middle part of the ramp (the margin of error is

approximately 3 km/h). At ramp 2, the precision of the estimate of $\beta_0(t)$ is the highest at approximately 250th meter (the margin of error is less than 1 km/h); this part of the ramp corresponds to the curve. From here towards the beginning and the end of the ramp, the margin of error increases to 5 km/h. Concerning ramp 4, the width of the point-wise confidence interval for the estimate of $\beta_0(t)$ is estimated with a higher precision in the first half of ramp 4 (the margin of error is approximately 5 km/h), while after 500th meter, the margin of error increases to 40 km/h and then it sharply decreases. In the case of ramp 5, the width of the confidence interval is higher in the first half of the ramp (the margin of error is approximately 50 km/h) and then $\beta_0(t)$ is estimated with more precision in the part of ramp 5 corresponding to the loop.

Secondly, we concern the precision of the estimates of $\beta_{\text{radius}}(t)$. At ramp 2, the margin of error increases from 0.005 to 0.01 with the increasing distance. A similar pattern is observed at ramp 3 with a similar shape and magnitude of the margin of error. In the case of ramp 4, the margin of error varies from the very low values in the beginning and the end of the ramp (0.1 and 0.2, respectively) to approximately 0.8 at 220th meter and 1.5 at 300th meter where the margin of error amplitudes and then becomes lower again. Concerning ramp 5, the point-wise confidence interval for the estimate of $\beta_{\text{radius}}(t)$ is narrower in the beginning of the ramp (the margin of error increases from 0.01 to 0.025), from where it increases to the maximum of almost 0.1 at 520th meter. From this point, the width of the confidence interval for the estimate of $\beta_{\text{radius}}(t)$ begins to decrease very sharply to 0.01. It should be noted that the radius is shown in metres, which allows to capture even small changes in the curvature. Therefore, based on the units chosen for the radius, the effect of the estimate of $\beta_{\text{radius}}(t)$ and the point-wise confidence interval limits manifests in small values only.

Lastly, the point-wise confidence intervals for the estimates of the regression functions for the auxiliary lane are discussed. One can see that they have similar width over the whole functional domain T. The most precise estimate of the regression function for the auxiliary lane, relative to the units, can be observed

at ramp 2 where the margin of error varies from 3 km/h at first 100 m, 1 km/h until 200th meter, and from 200th meter until the end of the ramp increases up to 2.5 km/h (see Figure 3.7). At ramp 1, the margin of error of the estimate of $\beta_{\text{merging lane}}(t)$ is the highest, approximately 5 km/h over the whole ramp (Figure 3.6). As can be seen at ramp 5, the precision of the estimate of $\beta_{\text{exit lane}}(t)$ is slightly less accurate at first 550 m (the margin of error is approximately 5 km/h). From there until the end of the ramp the confidence intervals become narrower (the margin of error decreases to 2 km/h), probably as the vehicles drive through the loop which is followed by the exit lane. In the first, more straight part of ramp 5, their speeds can vary more, while after entering the loop, they all need to drive similarly, which allows for a greater precision of the estimate of $\beta_{\text{exit lane}}(t)$ (see Figure 3.10). In contrast, the confidence intervals for the estimate of $\beta_{\text{exit lane}}(t)$ at ramp 3 are narrower in the first part of the ramp (the margin of error is approximately 4 km/h) and become slightly wider after 200th meter. After passing this point, the vehicles start to decelerate to exit the central lane and the margin of error increases from 4 to 8 km/h (Figure 3.8).

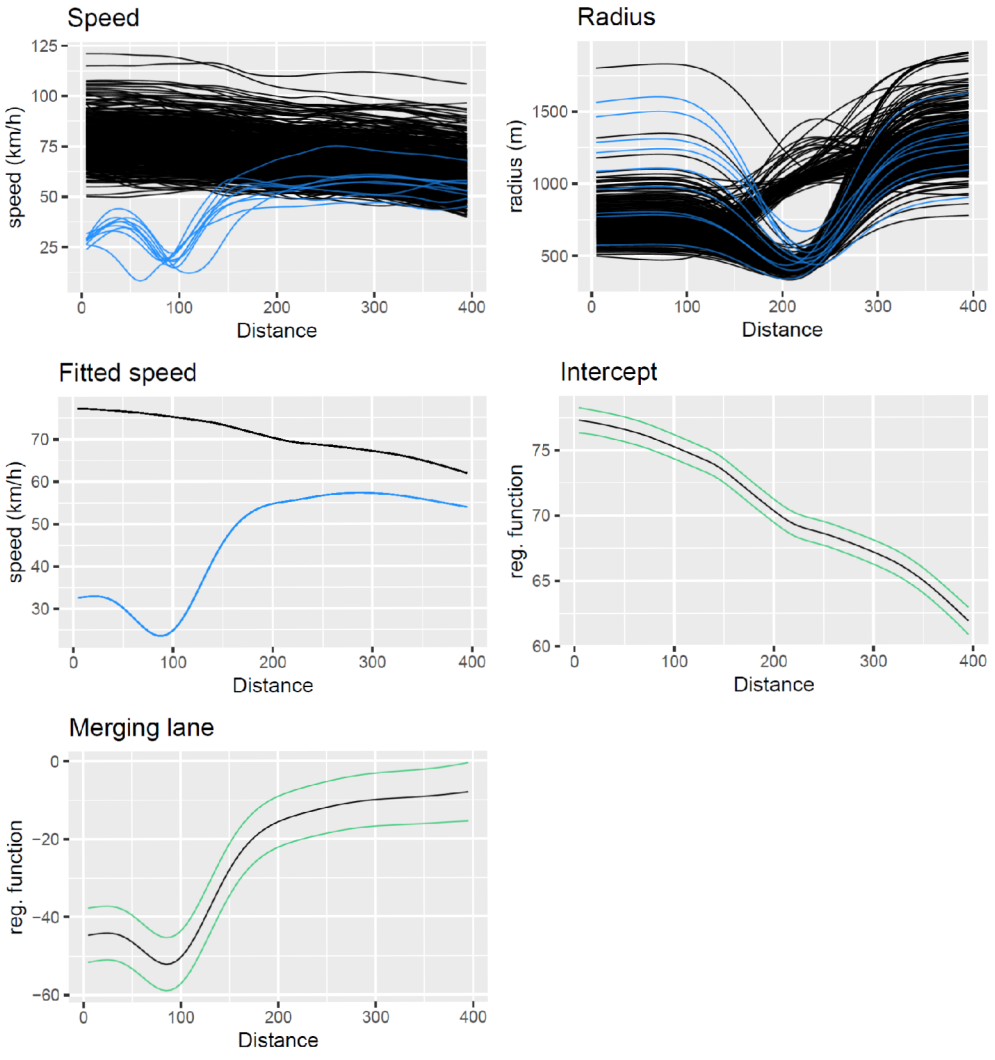


Figure 3.6: Ramp 1: The functional observations of the speed and the radius, the fitted speed, the estimates of the regression functions together with the point-wise confidence intervals. The colours distinguish the observations from the central lane (black) and from the merging lane (blue).

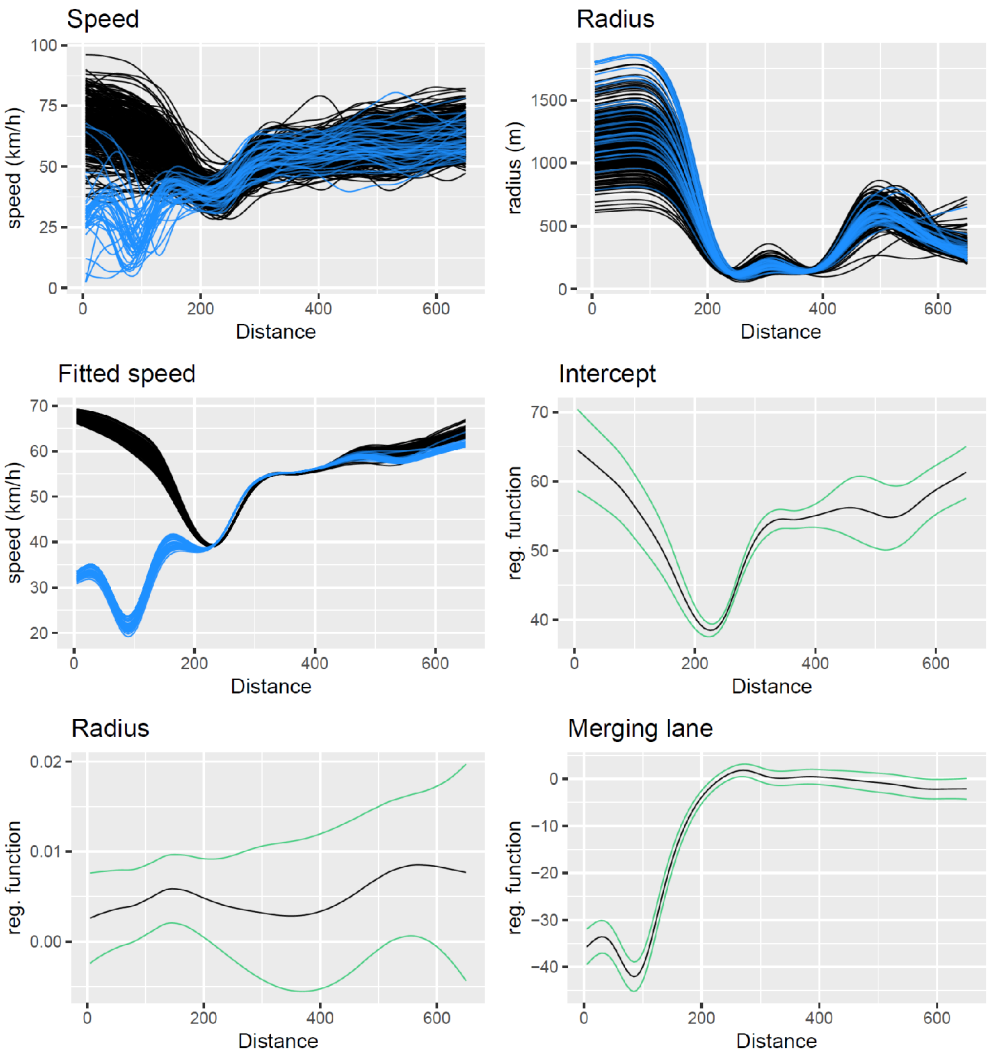


Figure 3.7: Ramp 2: The functional observations of the speed and the radius, the fitted speed, the estimates of the regression functions together with the point-wise confidence intervals. The colours distinguish the observations from the central lane (black) and from the merging lane (blue).

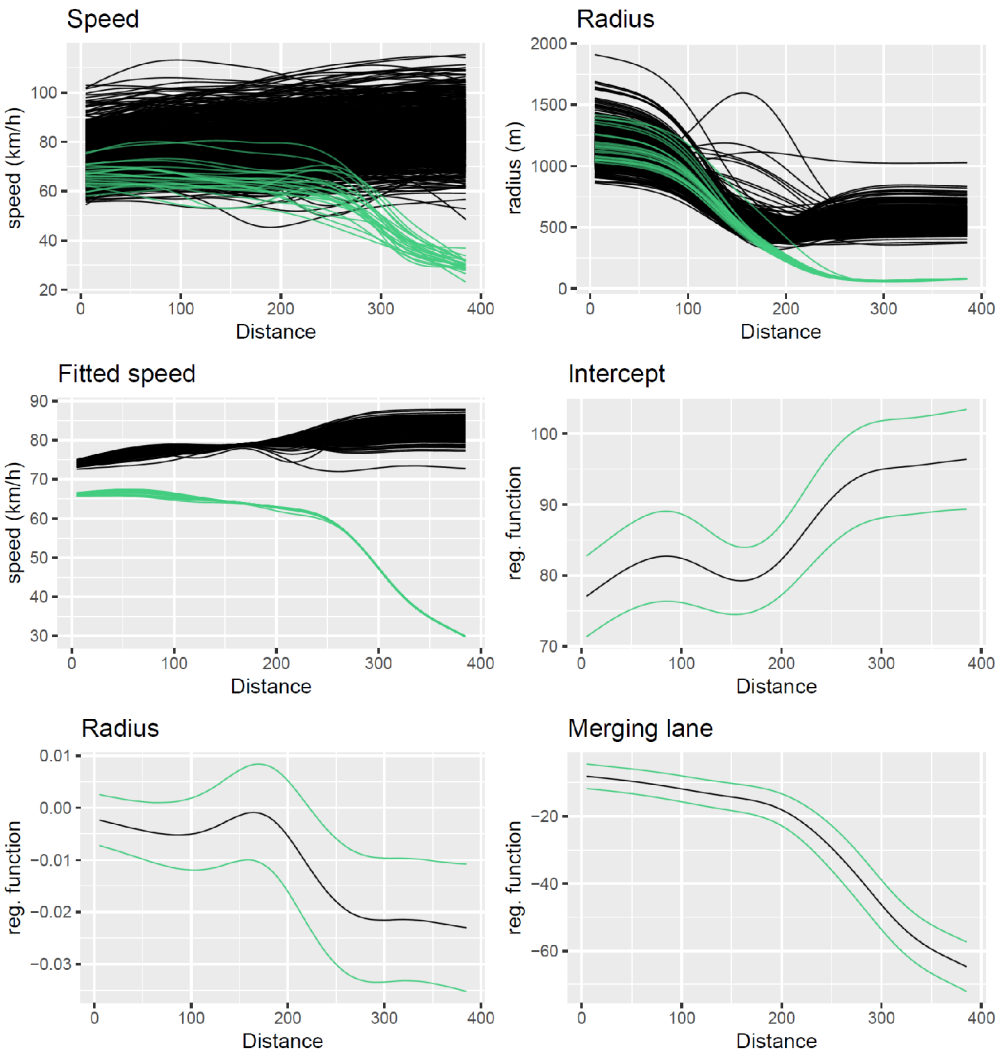


Figure 3.8: Ramp 3: The functional observations of the speed and the radius, the fitted speed, the estimates of the regression functions together with the point-wise confidence intervals. The colours distinguish the observations from the central lane (black) and from the exit lane (green).

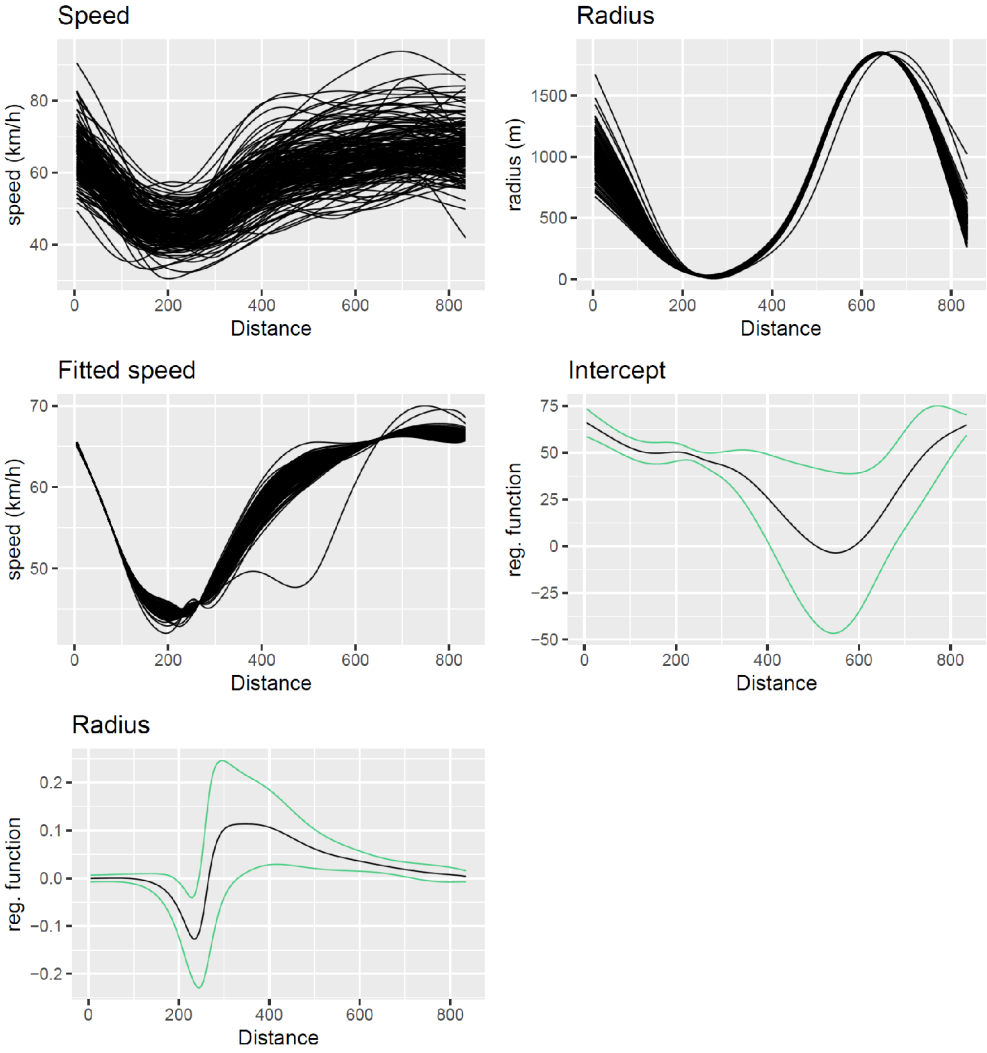


Figure 3.9: Ramp 4: The functional observations of the speed and the radius, the fitted speed, the estimates of the regression functions together with the point-wise confidence intervals.

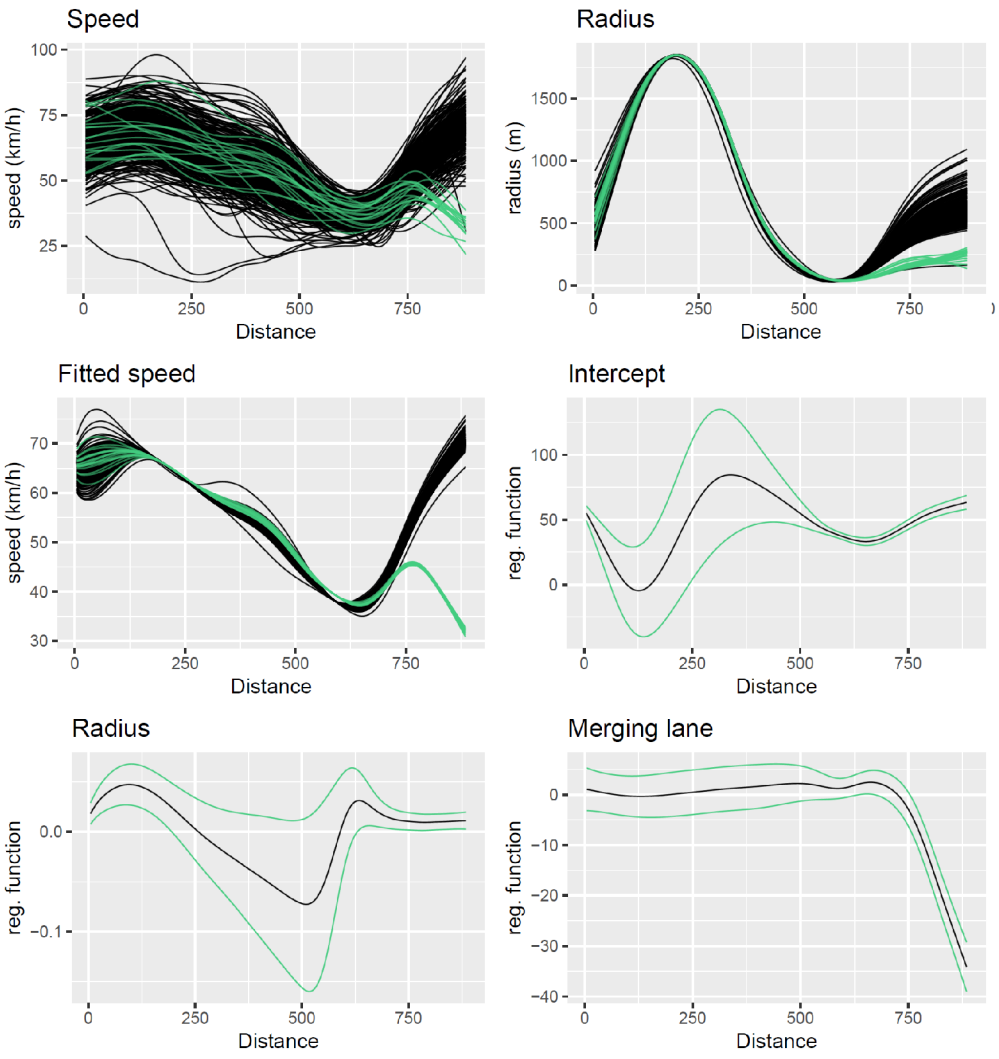


Figure 3.10: Ramp 5: The functional observations of the speed and the radius, the fitted speed, the estimates of the regression functions together with the point-wise confidence intervals. The colours distinguish the observations from the central lane (black) and from the exit lane (green).

3.6.7. Model limitations

Firstly, from the transportation research point of view, more variables related to the driving speed may be used in the future studies, as numerous factors are known to influence the driving speed [24, 29, 72]. These factors may fall into three domains: the road environment, the drivers, and the vehicles. Since FCD, due to the privacy regulations, do not usually contain information regarding the drivers and the vehicles, the road environment factors are usually the only ones available, including, e.g., the cross-section (road width, lane width) or the road-side characteristics (barriers, fixed obstacles, vegetation). For example, a more generous road width, i.e., more lanes, wider lanes, and a wider obstacle-free zone, was associated with a higher speed in the previous studies. On the other hand, the visual constraints, such as buildings or vegetation along the road or more curves, which limit the sight length, were associated with a lower speed [70].

The second limitation arises from the data source. Since FCD are only sourced from the fleet vehicles, the received data represent this subgroup and its characteristics only. The most influential characteristic is probably the type of the source vehicles. In the case of the data concerned in this application, these are the company vehicles and it is likely that their drivers may have different driving patterns compared to the general population of the drivers.

Thirdly, from the methodological point of view, the inference procedures presented in this chapter assume homoscedasticity among the observations. This assumption can be verified, e.g., by the global test for homoscedasticity [66] (discussed in Section 4.5) or by the two-sample interval-wise t-test [60] (discussed in Section 3.5) generalized for testing the variances.

The last limitation can arise from the computational demand. In case of this application, the data preparation and the model fitting were running fast, but the inference in regression models, the model selection and the estimates of point-wise confidence intervals required more computational time. In particular, the model selection for all six ramps ran approximately 8 hours (parallel computations on a Windows 10 Home machine with AMD Ryzen 7 2700U, 2.2 GHz, 8-GB RAM).

Chapter 4

Regression models for spatial covariates and functional response

In this chapter, a definition of the spatial-on-functional regression model, together with the methods for the analysis of spatial correlation are provided in Sections 4.1 and 4.2. Then, the estimation of the regression functions is overviewed in Section 4.3, detailing the estimates of regression functions for the cases of homoscedastic, heteroscedastic, and spatially correlated functional observations. Next, the permutation-based tests for the general linearly independent linear hypotheses about the spatial covariates and their special cases, the omnibus and the single-parameter hypothesis tests for the lack of the effect of the spatial covariates, are introduced in Section 4.4. In Section 4.5, the methods for assessing the model assumptions for the case of heteroscedasticity are introduced. The performance of the proposed permutation-based tests is evaluated in Sections 4.6 and 4.7 through the extensive simulation studies. Lastly, all the proposed methods are demonstrated on the available real-world data sets from the fields of geochemistry and tourism (Sections 4.8 and 4.9).

4.1. Model specification

Let the spatial domain of interest be denoted as $D \subset \mathbb{R}^d$. Let s be a point in D and denote by $X_s(t)$ an observation at location $s \in D$ at time $t \in T = [t_1, t_2]$. From this point, we assume that for the fixed location s, the random function

$\mathcal{X}_s(\cdot)$ is an element of the Hilbert space $L^2(T)$ of the square-integrable functions. The set

$$\{\mathcal{X}_s(t), t \in T, s \in D \subset \mathbb{R}^d\} \tag{4.1}$$

is called the functional random field of the time-varying entity, defined on L^2.

In the following, we assume that the functional random field (4.1) is second-order stationary and isotropic in the sense of the work of [52], i.e.,

1. Second-order stationarity
   - $\mathrm{E}(\mathcal{X}_s(t)) = \mu_s(t), \forall s \in D$,
   - $\mathrm{Cov}(\mathcal{X}_{s_i}, \mathcal{X}_{s_j}) = \mathrm{E}(\langle \mathcal{X}_{s_i} - m_{s_i}, \mathcal{X}_{s_j} - m_{s_j} \rangle) = C(h), \forall s_i, s_j \in D$, $h = s_i - s_j, h \in D$, and s_i, s_j, h are d-dimensional vectors,
2. Isotropy
   - $\mathrm{Cov}(\mathcal{X}_{s_i}, \mathcal{X}_{s_j}) = \mathrm{E}(\langle \mathcal{X}_{s_i} - \mu_{s_i}, \mathcal{X}_{s_j} - \mu_{s_j} \rangle) = C(\|h\|_D), \forall s_i, s_j \in D$, $h = s_i - s_j, h \in D$, and s_i, s_j, h are d-dimensional vectors,

where $\|\cdot\|_D$ is a norm over D and $\langle \cdot, \cdot \rangle$ denotes the inner product in L^2. This means that, only the mean $\mathrm{E}(\mathcal{X}_s(t))$ is a function of the location s, whereas the global covariances $\mathrm{Cov}(\mathcal{X}_{s_i}, \mathcal{X}_{s_j})$ do not depend on the locations s_i, s_j, but only on $\|h\|_D$, the distance between these locations [21, 39, 52].

Following the work of [52], we assume that the random field can be modelled as

$$\mathcal{X}_s(t) = m_s(t) + \delta_s(t), s \in D, t \in T, \tag{4.2}$$

where $m_s(t)$ is the drift and $\delta_s(t)$ are the residuals, the realizations of a zero-mean, second-order stationary and isotropic random process. The drift captures a non-constant mean variation in space D and can be represented by a linear model

$$m_s(t) = \beta_0(t) + \sum_{\ell=1}^{L} \beta_\ell(t) f_\ell(s), s \in D, t \in T, \tag{4.3}$$

where $\beta_\ell(t), \ell \in \{0, \ldots, L\}, t \in T$, are the unknown regression functions, independent on the location in space D, and $f_\ell(s)$, $\ell \in \{1, \ldots, L\}, s \in D$, are the known functions of the spatial variable $s \in D$.

Model (4.2) can be rewritten into the matrix form as

$$\boldsymbol{\mathcal{X}} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\delta}, \tag{4.4}$$

where $\boldsymbol{\mathcal{X}} = (\mathcal{X}_{s_1}(t), \ldots, \mathcal{X}_{s_n}(t))', t \in T$, is a matrix of the functional observations at the spatial points $s_1, \ldots, s_n \in D, \mathbf{F}_i = (1, f_1(s_i), \ldots, f_L(s_i)), s_i \in D$, $i \in \{1, \ldots, n\}$, is an i-th row of a known $n \times (L+1)$ design matrix $\mathbf{F}$, $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_L(t))', t \in T$, are the unknown regression functions, and $\boldsymbol{\delta} = (\delta_{s_1}(t), \ldots, \delta_{s_n}(t))', s_1, \ldots, s_n \in D, t \in T$, are the spatially correlated residuals with an unknown variance-covariance structure represented by the matrix $\boldsymbol{\Sigma}$. Note that, $\boldsymbol{\Sigma}$ is constant over T.

4.2. Analysis of the spatial correlation

From here, to avoid a heavy notation, we directly set $h = \|s_i - s_j\|_D$. A covariance function of a second-order stationary and isotropic random field (4.1) can be typically parametrized as

$$C(h) = \sigma^2 \rho(h), h \geq 0, \rho(0) = 1, \tag{4.5}$$

where $\rho(h)$ is a correlation function measuring the linear relationship between the observations at the distance h. Under these assumptions, the (i, j)-th entry of the variance-covariance matrix $\boldsymbol{\Sigma}$ is given as

$$\Sigma_{ij} = \mathrm{Cov}(\delta_{s_i}, \delta_{s_j}) = C(h), s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}. \tag{4.6}$$

The spatial dependence among the residuals can be also specified by means of the (functional) semivariogram

$$\gamma(h) = \frac{1}{2}\,\mathrm{Var}\left(\delta_{s_i} - \delta_{s_j}\right) = \frac{1}{2}\,\mathrm{E}[\|\delta_{s_i} - \delta_{s_j}\|^2], s_i, s_j \in D, h = \|s_i - s_j\|_D\,, \tag{4.7}$$

where $\|\cdot\|$ is the L^2-norm. Note that, one can define the variogram as $2\gamma(h)$. The relationship between the semivariogram (4.7) and the covariance function (4.5) can be expressed as

$$\gamma(h) = C(0) - C(h), s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}. \tag{4.8}$$

The functional semivariogram is estimated by the empirical functional semivariogram [51, 52]

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \|\delta_{s_i} - \delta_{s_j}\|^2, s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}, \tag{4.9}$$

where $N(h)$ is the set of all pairs of the observations at a distance approximately h and $|N(h)|$ is the cardinality of this set. As was pointed out in [52], calculating $\widehat{\gamma}(h)$ for every distance h may not be feasible in real applications. Instead, the distances can be binned into K intervals centred in $h_1, \ldots, h_K$, computing a discretized version of the semivariogram $\widehat{\gamma}(h) = (\widehat{\gamma}(h_1), \ldots, \widehat{\gamma}(h_K))$. In the next step, a parametric model is fitted to the empirical semivariogram. To find an optimal model, a distance between the empirical estimate $\widehat{\gamma}(h)$ and a parametric model $\gamma(h)$ is minimized.

Similarly to the scalar case [19], the shape of the trace-variogram can be used to determine if $\mathbf{\Sigma}$ is diagonal. Indeed, the uncorrelated residuals δ_{s_i} are associated with a pure nugget model for a variogram, that is, a constant trace-variogram function.

The more detailed information regarding the variography and the other approaches to the estimation of the variance-covariance matrix $\mathbf{\Sigma}$ can be found, e.g., in [19, 51, 52].

4.3. Estimation of the regression functions

Consistently with the case studies which shall be concerned in this chapter, we assume that the residuals are either homoscedastic, heteroscedastic, or spatially correlated, and that their spatial variance-covariance matrix is invertible.

When the conditions $\text{rank}(\mathbf{F}) = L + 1 \leq n$ and $\text{rank}(\boldsymbol{\Sigma}) = n$ hold, the regression functions $\boldsymbol{\beta}$ in the model (4.4) can be estimated as follows.

For the case of homoscedastic and uncorrelated residuals, the variance-covariance matrix is of the form $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix, and the estimator of the regression functions $\boldsymbol{\beta}$ simplifies to the ordinary least-squares (OLS) one [52] as

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\mathcal{X}}. \tag{4.10}$$

The OLS estimator is the best linear unbiased estimator and is associated with the variance-covariance matrix $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2(\mathbf{F}'\mathbf{F})^{-1}$. The parameter σ^2 is usually unknown and is simply estimated from the estimated residuals by Equation (3.9), where $\boldsymbol{\kappa} = \widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is a vector of the estimated functional residuals.

For the case of heteroscedastic and uncorrelated residuals, the variance-covariance matrix $\boldsymbol{\Sigma}$ is a diagonal matrix. Let $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ be a diagonal matrix of the weights. In general, the weight $w_{ii}, i \in \{1, \ldots, n\}$, can differ for each observation. However, in the following, based on the example outlined in Section 1.2, we shall assume that the weights are constant within the groups and possibly different between the groups in the data. The regression functions $\boldsymbol{\beta}$ can be estimated through the weighted least-squares (WLS) estimator as

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{F}'\mathbf{W}\mathbf{F})^{-1}\mathbf{F}'\mathbf{W}\boldsymbol{\mathcal{X}}. \tag{4.11}$$

Assuming $\mathbf{W}$ is a diagonal matrix, the WLS estimator is the best linear unbiased estimator and is associated with the variance-covariance matrix $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{WLS}}) = (\mathbf{F}'\mathbf{W}\mathbf{F})^{-1}$.

For the case of spatially correlated residuals, a general variance-covariance matrix $\boldsymbol{\Sigma}$ is associated with the residuals, and the regression functions $\boldsymbol{\beta}$ can be estimated through the generalized least-squares (GLS) estimator [52] as

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{X}}, \tag{4.12}$$

The GLS estimator is the best linear unbiased estimator and is associated with the variance-covariance matrix $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{GLS}}) = (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}$.

In practice, the variance-covariance matrix $\mathbf{\Sigma}$ describing the spatial dependence is usually unknown, so that the GLS estimator is not available. The general variance-covariance matrix $\mathbf{\Sigma}$ can be estimated, e.g., from the estimated residuals by the semivariogram through an iterative two-step procedure for the GLS estimates of the regression functions (Algorithm 2). Particularly, with respect to Equations (4.5) and (4.8), the (i, j)-th entry of $\mathbf{\Sigma}$ is estimated as

$$\widehat{\mathbf{\Sigma}}_{i,j} = \widehat{\sigma}^2 - \widehat{\gamma}(h), s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}, \tag{4.13}$$

where $\widehat{\sigma}^2$ and $\widehat{\gamma}(h)$ are given by Equations (3.9) and (4.9), respectively.

In the case of the diagonal variance-covariance matrix $\mathbf{\Sigma} = \mathbf{W}^{-1}$, under the presence of groups in the data where the variances are assumed to be constant within each group and possibly different between the groups, the estimation of the diagonal matrix $\mathbf{W}$ is as follows. Assume $\delta_{s_i(m)}(t), s_i \in D, i \in \{1, \ldots, n_m\}$, $m \in \{A, B\}$, are the two groups of the functional residuals from the model specified in Equations (4.2) and (4.3), where i is the unit index and m is the population index. Let $\sigma^2_{(A)}$ and $\sigma^2_{(B)}$ be the global variances (constant over T) of these two populations. The variances are estimated separately for each group by Equation (4.27), and their estimates $\widehat{\sigma}^2_{(A)}$ and $\widehat{\sigma}^2_{(B)}$ are indeed the estimates of the reciprocal diagonal elements $w_{ii(A)}$ and $w_{ii(B)}$ of the matrix of weights $\mathbf{W}$, i.e., $\widehat{w}_{ii(m)} = 1/\widehat{\sigma}^2_{(m)}, i \in \{1, \ldots, n\}, m \in \{A, B\}$. The rest of the procedure is analogous to the GLS case and is described in Algorithm 2.

In both cases, the iterative estimation procedure of $\boldsymbol{\beta}$ should be repeated few times until the convergence is reached [52]. The iterative procedure is stopped when the change of the estimates of the residuals in two consecutive steps is sufficiently small, i.e., if

$$\|\widehat{\delta}_{s_i,l} - \widehat{\delta}_{s_i,l-1}\| < \text{given value}, s_i, s_j \in D, i \in \{1, \ldots, n\}. \tag{4.14}$$

Here, $\widehat{\delta}_{s_i,l}$ denotes the estimate of the residual δ_{s_i} in the l-th step.

Algorithm 2 Iterative two-step estimation procedure of the regression functions.

1. Estimate the regression functions $\boldsymbol{\beta}$ by the OLS (Equation (4.10)).
2. Determine the estimated residuals $\widehat{\boldsymbol{\delta}}$ as $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{\text{OLS}}$.
3. Estimate the variance-covariance matrix $\boldsymbol{\Sigma}$ (or $\mathbf{W}$) from the estimated residuals $\widehat{\boldsymbol{\delta}}$.
4. Estimate the regression functions $\boldsymbol{\beta}$ by the GLS (or the WLS) by plugging in $\widehat{\boldsymbol{\Sigma}}$ (or $\widehat{\mathbf{W}}$) into Equation (4.12) (or Equation (4.11)).
5. Estimate the residuals $\boldsymbol{\delta}$ as $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{\text{GLS}}$ (or as $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{\text{WLS}}$).
6. Repeat steps 3-5 until the condition (4.14) is met.

4.4. The global permutation tests for the effect of the covariates

This section focuses on the problem of testing for the significance of the regression functions in model (4.3). Recall a compact matrix form of the linear hypotheses about the effects of covariates as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}, \text{ against } H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}, \tag{4.15}$$

where $\mathbf{C}$ is an $h \times (L+1)$ matrix of constants and $1 \leq h \leq L+1$ is a number of linearly independent hypotheses. An important special case is the omnibus test considering

$$H_0 : \beta_1(t) = \ldots = \beta_L(t) = 0 \, \forall t \in T, \text{ against} \tag{4.16}$$

$$H_1 : \beta_\ell(t) \neq 0 \text{ for some } \ell \in \{1, \ldots, L\} \text{ and some } t \in T, \tag{4.17}$$

to compare the full and the null models. In this case, $\mathbf{C}$ is an $L \times (L+1)$ block matrix partitioned as $\mathbf{C} = (\mathbf{0}|\mathbf{I})$, where $\mathbf{0}$ is a vector of zeros of length L and $\mathbf{I}$ is an $(L \times L)$ identity matrix. In the case of testing the effect of a single regression function $\beta_\ell(t), \ell \in \{1, \ldots, L\}$, through the hypothesis

$$H_0^\ell : \beta_\ell(t) = 0 \, \forall t \in T, \text{ against } H_1^\ell : \beta_\ell(t) \neq 0 \text{ for some } t \in T, \tag{4.18}$$

the matrix $\mathbf{C}$ is reduced to a unit row vector with 1 on the ℓ-th entry and zeros otherwise, i.e, $\mathbf{C}_\ell = (0, \ldots, 0, 1, 0, \ldots, 0)$.

If the variance-covariance matrix has the form $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, one can use the test statistic [66]

$$\mathrm{T}_{\mathrm{OLS}} = \int_T \frac{1}{\widehat{\sigma}^2}\big(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\big)' \left[\mathbf{C}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \big(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\big)dt \tag{4.19}$$

for testing the hypothesis (4.15). The test of a single regression function $\beta_\ell(t)$, $\ell \in \{1, \ldots, L\}$, expressed through the hypothesis (4.18), can be run using the test statistic

$$\mathrm{T}^{\ell}_{\mathrm{OLS}} = \int_T \frac{\left(\mathbf{C}_\ell\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)^2}{\widehat{\sigma}^2\mathbf{C}_\ell(\mathbf{F}'\mathbf{F})^{-1}\mathbf{C}'_\ell}dt. \tag{4.20}$$

Without the presence of spatial correlation and under homoscedasticity, a decision about the hypotheses (4.15) and (4.18) can be made by the Freedman and Lane permutation scheme (see [31]) suitably adapted to the functional data as in [1]. The principle of this permutational scheme is outlined in Section 3.4. Algorithm 3 provides a detailed description of the scheme.

Under the presence of heteroscedasticity or spatial correlation, a modification of the testing procedure provided in Algorithm 3 is necessary to account for the heterogeneous structure of the estimated residuals $\widehat{\boldsymbol{\delta}}$ which are no longer asymptotically exchangeable if characterized by a heteroscedastic or spatial structure.

Without the presence of spatial correlation and under heteroscedasticity, the hypothesis (4.15) can be tested using a test statistic accounting for the covariance structure of the residuals,

$$\mathrm{T}_{\mathrm{WLS}} = \int_T \big(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}\big)' \left[\mathbf{C}(\mathbf{F}'\widehat{\mathbf{W}}\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \big(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}\big)dt, \tag{4.21}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ is the weighted least-squares estimator of $\boldsymbol{\beta}$, obtained by Algorithm 2. One can proceed analogously in the case of the test of a single regression function

Algorithm 3 Freedman and Lane permutation scheme for the omnibus test in the case $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$.

1. Compute the test statistic $\mathrm{T}_{\mathrm{OLS}}^{\mathrm{data}}$ from the original data.
2. During the e-th permutation, $e \in \{1, \ldots, \mathcal{E}\}$,
   (a) Determine the estimated residuals $\widehat{\delta}_{s_i,r}$ of the reduced model $\mathcal{X}_{s_i,r} = \beta_0 + \delta_{s_i,r}$. Obtain $\widehat{\beta}_0, \widehat{\delta}_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (b) Permute the estimated residuals $\widehat{\delta}_{s_1,r}, \ldots, \widehat{\delta}_{s_n,r}$, obtaining $\widehat{\delta}^*_{s_i,r}$, $s_i \in D, i \in \{1, \ldots, n\}$.
   (c) Compute the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ through the fitted reduced model and the permuted estimated residuals $\widehat{\boldsymbol{\delta}}^*_r$ as $\mathcal{X}^*_{s_i,r} = \widehat{\beta}_0 + \widehat{\delta}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (d) Estimate the regression functions $\boldsymbol{\beta}$ of the full model from the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ by the OLS (Equation (4.10)). Obtain $\widehat{\boldsymbol{\beta}}^*_{\mathrm{OLS}}, \widehat{\boldsymbol{\delta}}^*_p$.
   (e) Estimate the parameter σ^2_p from the estimated residuals $\widehat{\boldsymbol{\delta}}^*_p$ from step (d), obtaining $\widehat{\sigma}^2_p$.
   (f) Compute the test statistic $\mathrm{T}^{*e}_{\mathrm{OLS}}$ as $\mathrm{T}^{*e}_{\mathrm{OLS}} = \int_T \frac{1}{\widehat{\sigma}^2_p} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{OLS}}\right)' \left[\mathbf{C}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{OLS}}\right) dt$.
3. Compute the global p-value of the $\mathrm{T}_{\mathrm{OLS}}$ test as the proportion of all permutations for which $\mathrm{T}^{*e}_{\mathrm{OLS}} \geq \mathrm{T}^{\mathrm{data}}_{\mathrm{OLS}}, e \in \{1, \ldots, \mathcal{E}\}$.

(see hypothesis (4.18)), using the test statistic

$$\mathrm{T}^{\ell}_{\mathrm{WLS}} = \int_T \frac{\left(\mathbf{C}_\ell \widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}\right)^2}{\mathbf{C}_\ell (\mathbf{F}'\widehat{\mathbf{W}}\mathbf{F})^{-1}\mathbf{C}'_\ell} dt. \tag{4.22}$$

Define the vector $\boldsymbol{\varepsilon} = \mathbf{W}^{1/2}\boldsymbol{\delta}$, as

$$\boldsymbol{\varepsilon}_{s_i} = w_{ii}^{1/2}\, \delta_{s_i}, s_i \in D, i \in \{1, \ldots, n\}. \tag{4.23}$$

Clearly, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathbf{I}$, and $\{\varepsilon_{s_i}\}, s_i \in D, i \in \{1, \ldots, n\}$, are exchangeable. In principle, one may specify any model for the weights $w_{ii}, i \in \{1, \ldots, n\}$, without substantial modifications of the method here proposed.

In this section, we propose to consider for the permutation scheme the standardized residuals $\boldsymbol{\varepsilon}$, which can be obtained from the estimated residuals $\widehat{\boldsymbol{\delta}}$ as

$\widehat{\boldsymbol{\varepsilon}} = \widehat{\mathbf{W}}^{1/2}\widehat{\boldsymbol{\delta}}$. Assuming the homoscedasticity within the groups and the heteroscedasticity between the groups, the variance $\sigma^2_{(m)}$ in the m-th group can be estimated via the sample variance within this group (Equation (4.27)). The estimated standardized residuals $\widehat{\boldsymbol{\varepsilon}}$ are then randomly permuted, and the corresponding permuted responses are used similarly as in the case of homoscedasticity. The permutation scheme is described in detail in Algorithm 4. This algorithm is a new proposal for the case of heteroscedasticity among the observations.

Algorithm 4 Freedman and Lane permutation scheme in the case $\boldsymbol{\Sigma} = \mathbf{W}^{-1}$.

1. Compute the test statistic $\mathrm{T}^{\mathrm{data}}_{\mathrm{WLS}}$ from the original data.
2. During the e-th permutation, $e \in \{1, \ldots, \mathcal{E}\}$,
   (a) Determine the estimated residuals $\widehat{\delta}_{s_i,r}$ of the reduced model $\mathcal{X}_{s_i,r} = \beta_0 + \delta_{s_i,r}$ using Algorithm 2. Obtain $\widehat{\beta}_0, \widehat{\mathbf{W}}_r, \widehat{\delta}_{s_i,r}, s_i \in D$, $i \in \{1, \ldots, n\}$.
   (b) Determine the exchangeable estimated residuals $\widehat{\boldsymbol{\varepsilon}}_r$ from $\widehat{\boldsymbol{\delta}}_r$ as $\widehat{\boldsymbol{\varepsilon}}_r = \widehat{\mathbf{W}}_r^{1/2}\widehat{\boldsymbol{\delta}}_r$.
   (c) Permute the estimated exchangeable residuals $\widehat{\varepsilon}_{s_1,r}, \ldots, \widehat{\varepsilon}_{s_n,r}$, obtaining $\widehat{\varepsilon}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (d) Compute the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ through the fitted reduced model and the permuted estimated heteroscedastic residuals $\widehat{\boldsymbol{\delta}}^*_r = \widehat{\mathbf{W}}_r^{-1/2}\widehat{\boldsymbol{\varepsilon}}^*_r$ as $\mathcal{X}^*_{s_i,r} = \widehat{\beta}_0 + \widehat{\delta}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (e) Estimate the regression functions $\boldsymbol{\beta}$ of the full model from the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ by the WLS (Equation (4.11)) using Algorithm 2. Obtain $\widehat{\boldsymbol{\beta}}^*_{\mathrm{WLS}}$, $\widehat{\boldsymbol{\delta}}^*_p$, $\widehat{\mathbf{W}}_p$.
   (f) Compute the test statistic $\mathrm{T}^{*e}_{\mathrm{WLS}}$ as $\mathrm{T}^{*e}_{\mathrm{WLS}} = \int_T \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{WLS}}\right)' \left[\mathbf{C}(\mathbf{F}'\widehat{\mathbf{W}}_p\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{WLS}}\right) dt$.
3. Compute the global p-value of the $\mathrm{T}_{\mathrm{WLS}}$ test as the proportion of all permutations for which $\mathrm{T}^{*e}_{\mathrm{WLS}} \geq \mathrm{T}^{\mathrm{data}}_{\mathrm{WLS}}, e \in \{1, \ldots, \mathcal{E}\}$.

Similarly to the case of heteroscedasticity, the residuals $\widehat{\boldsymbol{\delta}}$ are no longer asymptotically exchangeable if characterized by a spatial correlation. As an approach to dealing with this issue, we propose a modification of Algorithm 3, which

will be here referred to as the spatial functional Freedman and Lane Permutation scheme (SF-FLPerm). In the spatial functional setting, assuming a global variance-covariance matrix $\mathbf{\Sigma}$, the test statistic can be formulated as

$$\mathrm{T}_{\mathrm{GLS}} = \int_T \left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)' \left[\mathbf{C}(\mathbf{F}'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right) dt, \tag{4.24}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is the generalized least-squares estimator of $\boldsymbol{\beta}$, obtained by Algorithm 2. One can analogously proceed when testing the hypothesis (4.18), using the test statistic

$$\mathrm{T}^{\ell}_{\mathrm{GLS}} = \int_T \frac{\left(\mathbf{C}_{\ell}\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)^2}{\mathbf{C}_{\ell}(\mathbf{F}'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{F})^{-1}\mathbf{C}'_{\ell}} dt. \tag{4.25}$$

Define the vector of spatially filtered residuals $\boldsymbol{\varepsilon}$ as

$$\boldsymbol{\varepsilon} = \mathbf{\Sigma}^{-1/2}\,\boldsymbol{\delta}, \tag{4.26}$$

where $\mathbf{\Sigma}^{-1/2}$ is a symmetric square root of $\mathbf{\Sigma}^{-1}$ [42]. Clearly, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathbf{I}$, and $\{\varepsilon_{s_i}\}, s_i \in D, i \in \{1, \ldots, n\}$, are exchangeable.

The key difference of the SF-FLPerm scheme, compared to the Freedman and Lane scheme (see [31]) is that the inference is based on the permutations of the spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}$ instead of on the permutations of the correlated residuals $\widehat{\boldsymbol{\delta}}$. In more detail, at each iteration, the residuals $\widehat{\boldsymbol{\delta}}$ are decorrelated by using the inverse-square root of the residual spatial covariance $\widehat{\mathbf{\Sigma}}$, i.e., $\widehat{\boldsymbol{\varepsilon}} = \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\boldsymbol{\delta}}$. The spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}$ are permuted and then they are correlated back by using the square root of $\widehat{\mathbf{\Sigma}}$, i.e., $\widehat{\boldsymbol{\delta}} = \widehat{\mathbf{\Sigma}}^{1/2}\widehat{\boldsymbol{\varepsilon}}$, to create the permuted data set. That is, the distribution of the $\mathrm{T}_{\mathrm{GLS}}$ ($\mathrm{T}^{\ell}_{\mathrm{GLS}}$) test statistic under the permutations is estimated in a similar manner as in the homoscedasticity or heteroscedasticity cases, but through the estimated approximately exchangeable spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}$. The detailed overview of the spatial functional Freedman and Lane Permutation schemes is provided by Algorithms 5 and 6 which are designed for testing the general linear hypotheses about the effect of the covariates and the single parameter hypothesis, respectively. Algorithms 5 and 6 extend the scheme

for heteroscedastic observations proposed in [66] to the more complex structures of the data. To avoid repetition, a single hypothesis algorithm is not shown for OLS and WLS cases, but Algorithm 6 can be adapted to these cases as well.

Algorithm 5 Spatial functional Freedman and Lane permutation scheme for hypothesis (4.15).

1. Compute the test statistic $\mathrm{T}_{\mathrm{GLS}}^{\mathrm{data}}$ from the original data.
2. During the e-th permutation, $e \in \{1, \ldots, \mathcal{E}\}$,
   (a) Determine the estimated residuals $\widehat{\delta}_{s_i,r}$ of the reduced model $\mathcal{X}_{s_i,r} = \beta_0 + \delta_{s_i,r}$ using Algorithm 2. Obtain $\widehat{\beta}_0, \widehat{\boldsymbol{\Sigma}}_r, \widehat{\delta}_{s_i,r}, s_i \in D$, $i \in \{1, \ldots, n\}$.
   (b) Determine the exchangeable estimated spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}_r$ from $\widehat{\boldsymbol{\delta}}_r$ as $\widehat{\boldsymbol{\varepsilon}}_r = \widehat{\boldsymbol{\Sigma}}_r^{-1/2}\widehat{\boldsymbol{\delta}}_r$.
   (c) Permute the estimated exchangeable spatially filtered residuals $\widehat{\varepsilon}_{s_1,r}, \ldots, \widehat{\varepsilon}_{s_n,r}$, obtaining $\widehat{\varepsilon}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (d) Compute the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ through the fitted reduced model and the permuted estimated spatially correlated residuals $\widehat{\boldsymbol{\delta}}^*_r = \widehat{\boldsymbol{\Sigma}}_r^{1/2}\widehat{\boldsymbol{\varepsilon}}^*_r$ as $\mathcal{X}^*_{s_i,r} = \widehat{\beta}_0 + \widehat{\delta}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (e) Estimate the regression functions $\boldsymbol{\beta}$ of the full model from the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ by the GLS (Equation (4.12)) using Algorithm 2. Obtain $\widehat{\boldsymbol{\beta}}^*_{\mathrm{GLS}}, \widehat{\boldsymbol{\delta}}^*_p(t)$, $\widehat{\boldsymbol{\Sigma}}_p$.
   (f) Compute the test statistic $\mathrm{T}^{*e}_{\mathrm{GLS}}$ as
   $\mathrm{T}^{*e}_{\mathrm{GLS}} = \int_T \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{GLS}}\right)' \left[\mathbf{C}(\mathbf{F}'\widehat{\boldsymbol{\Sigma}}_p^{-1}\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{GLS}}\right) dt.$
3. Compute the global p-value of the $\mathrm{T}_{\mathrm{GLS}}$ test as the proportion of all permutations for which $\mathrm{T}^{*e}_{\mathrm{GLS}} \geq \mathrm{T}^{\mathrm{data}}_{\mathrm{GLS}}, e \in \{1, \ldots, \mathcal{E}\}$.

Algorithm 6 Spatial functional Freedman and Lane permutation scheme for hypothesis (4.18).

1. Compute the test statistic $\mathrm{T}^{\ell\text{data}}_{\text{GLS}}$ from the original data.
2. During the e-th permutation, $e \in \{1, \ldots, \mathcal{E}\}$,
   (a) Determine the estimated residuals $\widehat{\delta}_{s_i,r}$ of the reduced model $\mathcal{X}_{s_i,r} = \beta_0 + \sum_{k\neq\ell} \beta_k f_k(s_i) + \delta_{s_i,r}$ using Algorithm 2. Obtain $\widehat{\beta}_0, \widehat{\beta}_k$, $k \in \{1, \ldots, L\}, k \neq \ell, \widehat{\boldsymbol{\Sigma}}_r, \widehat{\delta}_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (b) Determine the exchangeable estimated spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}_r$ from $\widehat{\boldsymbol{\delta}}_r$ as $\widehat{\boldsymbol{\varepsilon}}_r = \widehat{\boldsymbol{\Sigma}}_r^{-1/2}\widehat{\boldsymbol{\delta}}_r$.
   (c) Permute the estimated exchangeable spatially filtered residuals $\widehat{\varepsilon}_{s_1,r}, \ldots, \widehat{\varepsilon}_{s_n,r}$, obtaining $\widehat{\varepsilon}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (d) Compute the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ through the fitted reduced model and the permuted estimated spatially correlated residuals $\widehat{\boldsymbol{\delta}}^*_r = \widehat{\boldsymbol{\Sigma}}_r^{1/2}\widehat{\boldsymbol{\varepsilon}}^*_r$ as $\mathcal{X}^*_{s_i,r} = \widehat{\beta}_0 + \sum_{k\neq\ell} \widehat{\beta}_k f_k(s_i) + \widehat{\delta}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (e) Estimate the regression functions $\boldsymbol{\beta}$ of the full model from the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ by the GLS (Equation (4.12)) using Algorithm 2. Obtain $\widehat{\boldsymbol{\beta}}^*_{\text{GLS}}, \widehat{\boldsymbol{\delta}}^*_p, \widehat{\boldsymbol{\Sigma}}_p$.
   (f) Compute the test statistic $\mathrm{T}^{\ell*e}_{\text{GLS}}$ as $\mathrm{T}^{\ell*e}_{\text{GLS}} = \int_T \frac{\left(\mathbf{C}_\ell \widehat{\boldsymbol{\beta}}^*_{\text{GLS}}\right)^2}{\mathbf{C}_\ell(\mathbf{F}'\widehat{\boldsymbol{\Sigma}}_p^{-1}\mathbf{F})^{-1}\mathbf{C}'_\ell} dt$.
3. Compute the global p-value of the $\mathrm{T}^{\ell}_{\text{GLS}}$ test as the proportion of all permutations for which $\mathrm{T}^{\ell*e}_{\text{GLS}} \geq \mathrm{T}^{\ell\text{data}}_{\text{GLS}}, e \in \{1, \ldots, \mathcal{E}\}$.

4.5. Assessment of the model assumptions under heteroscedasticity

As was outlined earlier in this chapter, three possible scenarios may occur under the spatial setting. Which approach to choose can be decided by the shape of the semivariogram. Firstly, the concave shape, where the semivariogram increases in the beginning part and then becomes constant or non-decreasing (see the example in Figure 4.1b), is associated with the spatial correlation and a general variance-covariance matrix $\boldsymbol{\Sigma}$. Secondly, the constant shape of the semi-

variogram (i.e., the pure nugget structure, see the example in Figure 4.1a) is associated with a diagonal variance-covariance matrix ($\boldsymbol{\Sigma}$ equals either $\sigma^2\mathbf{I}$, or $\mathbf{W}$). However, even for the case of spatially uncorrelated residuals, one still needs to verify the assumption of homoscedasticity, as this determines the exchangeability of the residuals. For this reason, we employ a test based on the same permutations as described in Sections 3.4, 3.5, and 4.4, but with a test statistic specifically aiming at detecting the differences in variances. Note that, using the multiple comparison principles, the test can be extended to the case of more populations.

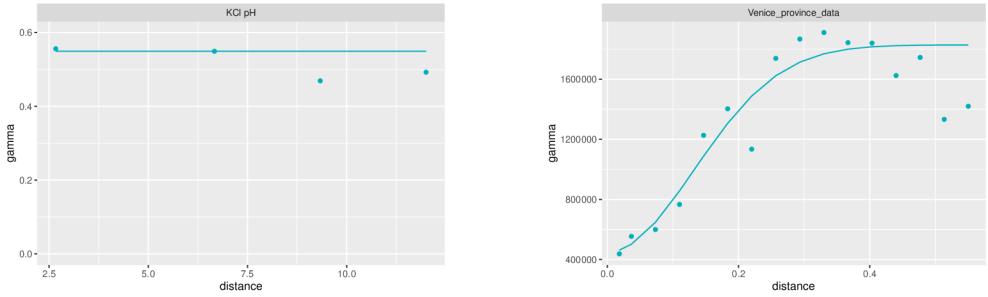


Figure 4.1: Examples of the nugget (left) and exponential (right) models for the semivariogram.

Let $\delta_{s_i(m)}(t), s_i \in D, i \in \{1,\dots,n_m\}, m \in \{A,B\}$, be two groups of the functional residuals from model in Equations (4.2) and (4.3), where i is the unit index and m is the population index. Let $\sigma^2_{(A)}$ and $\sigma^2_{(B)}$ be the global variances (constant over T) of these two populations, that is, $\sigma^2_{(m)} = \mathrm{E}[\|\delta_{s_i(m)}\|^2]$. The latter can be estimated as

$$\widehat{\sigma}^2_{(m)} = \frac{1}{n_m}\sum_{i=1}^{n_m} \|\widehat{\delta}_{s_i(m)}(t)\|^2, \tag{4.27}$$

where $s_i \in D$, $i \in \{1,\dots,n_m\}$, $m \in \{A,B\}$, $t \in T$, and $\|\cdot\|$ is the L^2-norm. Note that, by construction, the overall residual sample mean is zero; furthermore, the residual sample mean within the groups is zero if the dummy variables are included among the covariates (i.e., in a functional analysis of covariance setting).

The goal is now to test the hypothesis

$$H_0 : \sigma^2_{(A)} = \sigma^2_{(B)}, \text{ against } H_1 : \sigma^2_{(A)} \neq \sigma^2_{(B)}. \tag{4.28}$$

For this purpose, we propose as a test statistic the absolute value of the log-proportion of variances, as follows:

$$\mathrm{T}_{\mathrm{Var}} = \left| \log\left(\frac{\widehat{\sigma}^2_{(A)}}{\widehat{\sigma}^2_{(B)}} \right) \right|. \tag{4.29}$$

To perform the test, we consider a permutation scheme similar to those discussed in Sections 3.4, 3.5, and 4.4. Under H_0, the residuals are approximately exchangeable, and the permutation procedure from [1] can be applied. The global p-value of the $\mathrm{T}_{\mathrm{Var}}$ test (4.28) is computed as the proportion of the permutations leading to a value of $\mathrm{T}_{\mathrm{Var}}$ higher than or equal to the one observed in the data. The testing procedure is described in Algorithm 7. Its results can be used to establish which permutation strategy, among those detailed in Section 4.4, should be used to test the significance of the regression functions. Note that, the proposed statistical inference concerning homoscedasticity is only approximate due to its definition on the estimated residuals.

4.6. Simulation studies under heteroscedasticity

In this section, we report the results of a simulation study intended to evaluate the empirical size and the power of the proposed permutation-based global tests for the significance of the effect of covariates in the spatial regression model under heteroscedasticity (Section 4.1). In the following, we consider two simulation scenarios. Firstly, the empirical size and the power of the $\mathrm{T}_{\mathrm{OLS}}$ and $\mathrm{T}_{\mathrm{WLS}}$ tests are assessed for a simple functional linear model with one covariate. Secondly, the empirical size of both tests is evaluated for a model with more covariates. The design of the simulation studies is inspired by the real-world application concerned in Sections 1.2 and 4.8.

Algorithm 7 The permutation scheme for the homoscedasticity test for hypothesis (4.28).

1. Compute the test statistic $\mathrm{T}_{\mathrm{Var}}^{\mathrm{data}}$ from the original data.
2. During the e-th permutation, $e \in \{1, \ldots, \mathcal{E}\}$,
   (a) Sort function's indices as $\underbrace{1, 2, \ldots, n_A}_{\text{group } A}, \underbrace{n_A + 1, n_A + 2, \ldots, n_A + n_B}_{\text{group } B}$.
   (b) Permute the estimated residuals $\widehat{\delta}_{s_i(m)}, s_i \in D$, $i \in \{1, \ldots, n\}$, $m \in \{A, B\}$, from both groups together. Obtain
   $$\underbrace{\widehat{\delta}^*_{s_1}, \ldots, \widehat{\delta}^*_{s_{n_A}}}_{\text{group } A}, \underbrace{\widehat{\delta}^*_{s_{n_A+1}}, \ldots, \widehat{\delta}^*_{s_{n_A+n_B}}}_{\text{group } B}$$
   (c) Using the rearranged functions $\widehat{\delta}^*_{s_i(m)}, s_i \in D$, $i \in \{1, \ldots, n\}$, $m \in \{A, B\}$, estimate the group variances $\widehat{\sigma}^{*2}_{(A)} = \frac{1}{n_A} \sum_{i=1}^{n_A} \|\widehat{\delta}^*_{s_i(A)}\|^2$ and $\widehat{\sigma}^{*2}_{(B)} = \frac{1}{n_B} \sum_{i=n_A+1}^{n_A+n_B} \|\widehat{\delta}^*_{s_i(B)}\|^2$.
   (d) Compute the test statistic $\mathrm{T}^{*e}_{\mathrm{Var}} = \left| \log\left(\frac{\widehat{\sigma}^{*2}_{(A)}}{\widehat{\sigma}^{*2}_{(B)}} \right) \right|$.
3. Compute the global p-value of the $\mathrm{T}_{\mathrm{Var}}$ test as the proportion of all permutations for which $\mathrm{T}^{*e}_{\mathrm{Var}} \geq \mathrm{T}^{\mathrm{data}}_{\mathrm{Var}}, e \in \{1, \ldots, \mathcal{E}\}$.

The data are divided into two groups, group A with a variance $\sigma^2_{(A)}$, and group B with a variance $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}, a \in (0, \infty)$. The sample sizes of groups A and B are chosen as follows:

1. unbalanced design I, where $n_B = 3n_A$,
2. balanced design II, where $n_A = n_B$,
3. unbalanced design III, where $n_A = 3n_B$.

The heteroscedastic residuals $\delta_{s_1}(t), \ldots, \delta_{s_n}(t), s_1, \ldots, s_n \in D, t \in [3, 10]$, are generated by the following procedure. Denote by $\{\phi_q,\ q \in \{1, \ldots, 10\}\}$ the cubic B-spline basis, where the knots are placed at the data points, and by $\mathbf{v} = (v_1, \ldots, v_{10})'$ the corresponding coefficient vector. The residuals $\delta_{s_i}(t)$,

$s_i \in D, i \in \{1, \ldots, n\}, t \in [3, 10]$, are built on the same cubic B-spline basis expansion $\{\phi_q,\ q \in \{1, \ldots, 10\}\}$ as

$$\delta_{s_i}(t) = \sum_{q=1}^{10} v_q^{*i} \phi_q(t), s_i \in D, i \in \{1, \ldots, n\}, t \in [3, 10], \tag{4.30}$$

where the basis coefficients v_q^{*i} are drawn from a multivariate normal distribution with zero mean and a variance-covariance matrix $\mathbf{S_v}$. For group A, the matrix $\mathbf{S}_\mathbf{v}^{(A)}$ is set to the sample variance-covariance matrix of $\mathbf{v}$. To achieve heteroscedasticity, the variance-covariance matrix for the group B is multiplied by a scalar a, that is, $\mathbf{S}_\mathbf{v}^{(B)} = a \cdot \mathbf{S}_\mathbf{v}^{(A)}, a \in (0, \infty)$. In the next step, the simulated residuals $\delta_{s_1}(t), \ldots, \delta_{s_n}(t), s_1, \ldots, s_n \in D,\ t \in [3, 10]$, are added to the model for the drift (4.3), obtaining the functional observations $\mathcal{X}_{s(m)}(t), s \in D,\ m \in \{A, B\}$.

4.6.1. Assessing the empirical size and the power of the T_{OLS} and T_{WLS} tests in a simple model with an indicator

The first simulation study is based on a functional linear model with one covariate. To evaluate the empirical size and the power of the T_{OLS} and T_{WLS} tests, a hypothesis of no effect of the covariate is tested. The nominal size of the test is set to $\alpha = 0.05$. The sample sizes $n \in \{20, 40, 200, 500\}$ are considered. The relationship among the group variances is modelled through $a \in \{3, 9\}$, such that $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}$.

For each scenario, 2000 simulations are performed. In each case, the model is built as

$$\mathcal{X}_{s(m)}(t) = \beta_0(t) + b \cdot \beta_1(t)\text{soil}(s) + \delta_{s(m)}(t),\ s \in D,\ t \in [3, 10], \tag{4.31}$$

where $m \in \{A, B\}$,

$$\text{soil}(s) = \begin{cases} 0 \text{ for group } A, \\ 1 \text{ for group } B, \end{cases} \tag{4.32}$$

and $b \in [0, 5]$ controls the severity of the deviation between the means of groups A and B. The true regression functions $\beta_0(t)$ and $\beta_1(t)$ are shown in Figure 4.2.

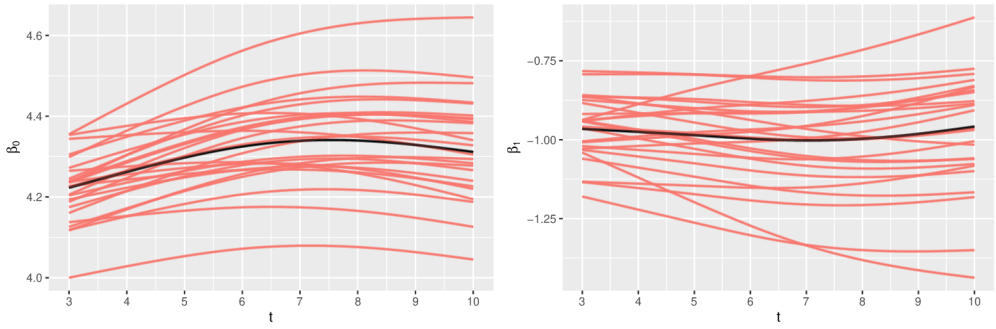


Figure 4.2: The true regression functions $\beta_0(t)$ and $\beta_1(t)$ (black) and their estimates $\widehat{\beta}_0(t)$ and $\widehat{\beta}_1(t)$ (red) for the simple model with an indicator (4.31). For the sake of clarity, only 25 estimates are shown.

The results of the simulation study for the size of the $\mathrm{T_{OLS}}$ and $\mathrm{T_{WLS}}$ tests are presented in Table 4.1; the power functions are shown in Figure 4.3. One can see that, for the balanced design II, the $\mathrm{T_{OLS}}$ and $\mathrm{T_{WLS}}$ tests show similar behaviour regarding the empirical size and the power; here, the empirical size of both tests is close to the nominal one, $\alpha = 5\,\%$. In the case of the unbalanced design I ($n_B = 3n_A$ and $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}, a \in \{3, 9\}$), the $\mathrm{T_{OLS}}$ test is very conservative for each considered sample size $n \in \{20, 40, 200, 500\}$. Its empirical size varies from 0.2 % to 1.3 % (95% confidence intervals: (0.1;0.5) and (0.9;1.9), respectively). In contrast, for the case of the unbalanced design III ($n_A = 3n_B$ and $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}, a \in \{3, 9\}$), the $\mathrm{T_{OLS}}$ test is, for each sample size $n \in \{20, 40, 200, 500\}$, very liberal. Its empirical size ranges from 12.4 % to 23.9 % (95% confidence intervals: (11.0;13.9) and (22.1;25.8), respectively). In general, under heteroscedasticity, the $\mathrm{T_{WLS}}$ test performs better than the $\mathrm{T_{OLS}}$ test whenever the design is unbalanced. The empirical size of the $\mathrm{T_{WLS}}$ test is close to the nominal size of $\alpha = 5\,\%$ in the majority of cases. However, for design III and a small sample size $n \in \{20, 40\}$ the empirical size of $\mathrm{T_{WLS}}$ is slightly higher than the nominal one, with the maximum empirical size of the $\mathrm{T_{WLS}}$ test being 9.6 % (95% confidence interval: (8.4;11.0)). This effect is related to design III, where there is necessary to estimate a high variance ($a \cdot \sigma^2_{(A)}, a \in \{3, 9\}$) from a relatively small number of the observations ($n_B = n/4$).

Table 4.1: The empirical sizes (in percentages) together with the 95% confidence intervals of the T_{OLS} and T_{WLS} tests for model (4.31) under different conditions based on 2000 simulations and 1000 permutations. The nominal size of the tests is set to $\alpha = 5\,\%$. The sample size is denoted by n and the data are divided into groups A and B, such that: unbalanced design I: $n_B = 3n_A$, balanced design II: $n_A = n_B$ and unbalanced design III: $n_A = 3n_B$. The relationship among the group variances is modelled through $a \in \{3, 9\}$, where $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}$.

		Design					
		I	I	II	II	III	III
n	Test	$a = 3$	$a = 9$	$a = 3$	$a = 9$	$a = 3$	$a = 9$
20	OLS	1.3 (0.9;1.9)	0.3 (0.1;0.7)	6.0 (5.0;7.1)	5.9 (5.0;7.0)	13.1 (11.7;14.7)	23.9 (22.1;25.8)
20	WLS	6.2 (5.2;7.3)	5.5 (4.6;6.6)	6.8 (5.8;8.0)	7.0 (6.0;8.2)	8.0 (6.9;9.3)	9.6 (8.4;11.0)
40	OLS	1.2 (0.8;1.8)	0.2 (0.1;0.5)	4.9 (4.0;5.9)	5.4 (4.5;6.5)	12.5 (11.1;14.0)	21.5 (19.8;23.4)
40	WLS	5.7 (4.8;6.8)	4.5 (3.7;5.5)	5.5 (4.6;6.6)	6.5 (5.5;7.7)	6.1 (5.1;7.2)	8.2 (7.1;9.5)
200	OLS	1.2 (0.8;1.8)	0.3 (0.1;0.7)	5.2 (4.3;6.3)	4.7 (3.9;5.7)	13.1 (11.7;14.7)	20.3 (18.7;22.1)
200	WLS	6.5 (5.5;7.7)	5.4 (4.5;6.5)	5.8 (4.9;6.9)	5.4 (4.5;6.5)	5.5 (4.6;6.6)	6.9 (5.9;8.1)
500	OLS	1.1 (0.7;1.6)	0.5 (0.2;0.9)	4.9 (4.0;5.9)	4.6 (3.8;5.6)	12.4 (11.0;13.9)	19.9 (18.2;21.7)
500	WLS	5.8 (4.9;6.9)	6.5 (5.5;7.6)	5.4 (4.5;6.4)	5.3 (4.4;6.3)	5.6 (4.7;6.7)	6.2 (5.2;7.3)

Focusing on the size of the difference among the group variances, where $\sigma^2_{(B)}$ is three or nine times higher than $\sigma^2_{(A)}$, one can see that, as a increases, the empirical power of both T_{OLS} and T_{WLS} tests converges to 1 for a bigger difference among the group means (compare the power functions for $a = 3$ and $a = 9$ in Figure 4.3). In the case of design III, the T_{OLS} test may be mistakenly considered more powerful than the T_{WLS} test. Nevertheless, this is a consequence of a highly liberal behaviour of the T_{OLS} test. As one could naturally expect, the power of both tests increases with a higher sample size n, as can be seen for each simulation scenario in Figure 4.3.

4.6.2. Assessing the empirical size of the T_{OLS} and T_{WLS} tests in a more complex model with more covariates

In this simulation scenario, the T_{OLS} and T_{WLS} tests are evaluated through a functional linear model with three covariates; an indicator function, comparable to model (4.31) from Section 4.6; a linear function of the distance; and an interaction term of these two. In total, 2000 simulations are run for a sample size $n = 200$.

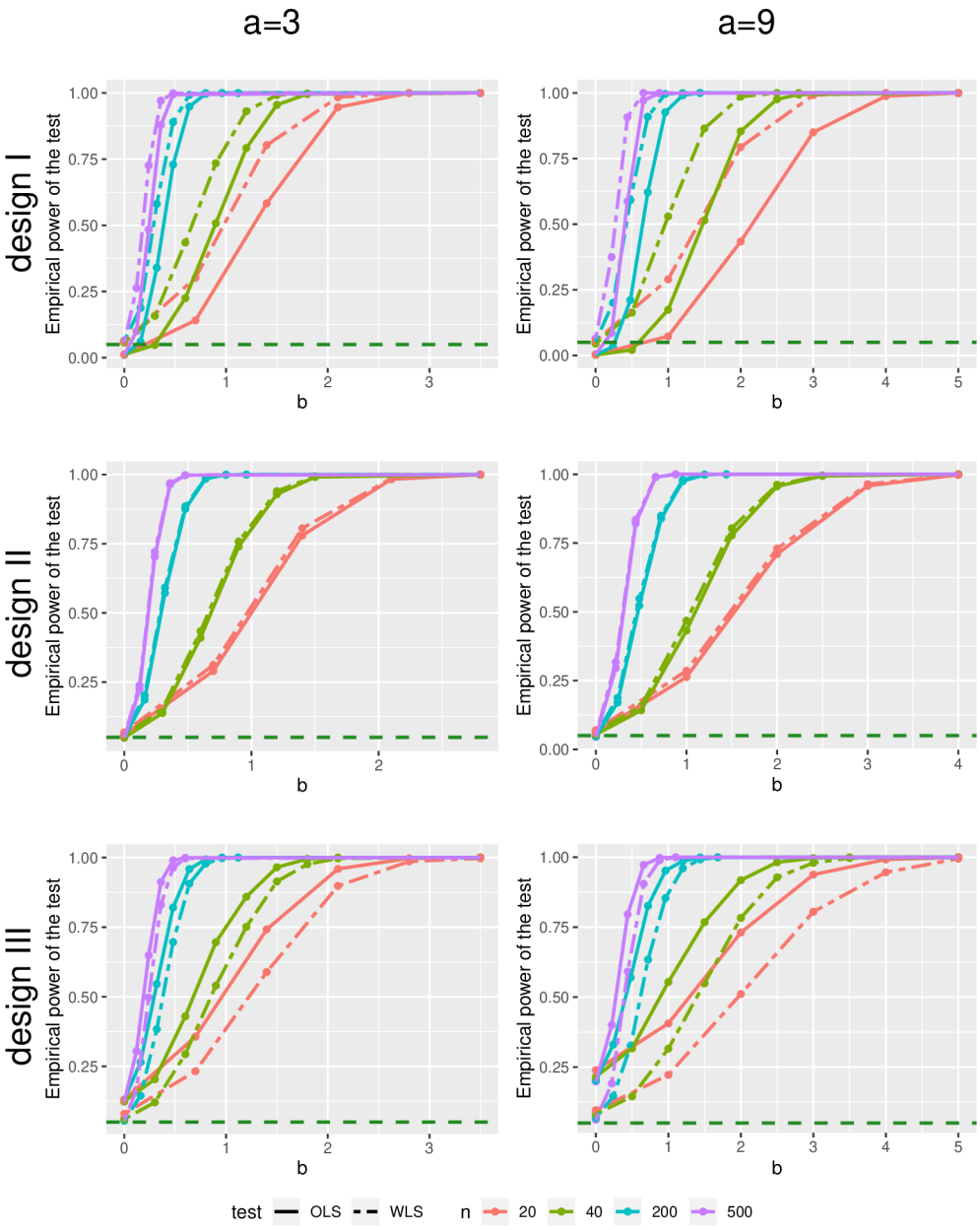


Figure 4.3: The empirical power of the T_{OLS} (solid line) and T_{WLS} (dot-dashed line) tests under different $a = \sigma^2_{(B)}/\sigma^2_{(A)}$ and design, based on 2000 simulations and 1000 permutations. The mean difference among the groups is modelled as $b \cdot \beta_1(t)$. The dashed horizontal line represents the nominal size of the test $\alpha = 0.05$.

Inspired by the case study from Sections 1.2 and 4.8, the model is built as

$$\mathcal{X}_{s(m)}(t) = \beta_0(t) + b \cdot \beta_1(t)\text{soil}(s) + \beta_2(t)\text{dist}(s) + b \cdot \beta_3(t)\text{soil}(s) \cdot \text{dist}(s) + \delta_{s(m)}(t), \tag{4.33}$$

where $s \in D, m \in \{A, B\}$, and $t \in [3, 10]$. The term $\text{dist}(s)$ denotes the distance from the central sampling point (the ecotone), where $\text{dist}(s) \in \{3, 6, \ldots, 3n_m\}$, and $m \in \{A, B\}$. In this case, evaluating the empirical size of the T_{OLS} and T_{WLS} tests, b is equal to 0. In this simulation study, the T_{OLS} and T_{WLS} tests are evaluated under both homoscedasticity and heteroscedasticity, i.e., for the group variances it holds $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}, a \in \{1, 3, 9\}$. The true regression functions $\beta_\ell(t), \ell \in \{0, 1, 2, 3\}$, are presented in Figure 4.4.

We aim at examining the empirical size of the T_{OLS} and T_{WLS} tests for the global null hypothesis $H_0 : \beta_1(t) = \beta_2(t) = \beta_3(t) = 0 \, \forall t \in [3, 10]$, and comparing it to the empirical sizes for the simple model with an indicator (4.31). The results (see Table 4.2) suggest that the T_{OLS} and T_{WLS} tests behave similarly for both the simple model with an indicator (4.31) and the more complex model (4.33). Under heteroscedasticity, the T_{OLS} approach in model (4.33) is, compared to the results for model (4.31), less conservative for design I and even more liberal for design III. On contrary, under homoscedasticity, the empirical sizes of both T_{OLS} and T_{WLS} tests are close to the nominal one, $\alpha = 5\,\%$. The T_{WLS} test, despite being applied to homoscedastic data, gives results comparable to the (proper) T_{OLS} approach. The precision of the OLS and WLS estimators of regression functions is more detailed in Section 4.6.3.

The results from the simulation studies from Sections 4.6.1 and 4.6.2 show that the T_{WLS} approach introduced in this section performs well under the setting of a complex functional regression model and heteroscedasticity at the same time. Concerning the either liberal or conservative behaviour of the T_{OLS} test, the results presented in Sections 4.6.1 and 4.6.2 are consistent with those obtained, in the scalar case, by [40].

Table 4.2: The empirical sizes (in percentages) together with the 95% confidence intervals of the T_{OLS} and T_{WLS} tests for models (4.31) and (4.33) and the sample size $n = 200$ based on 2000 simulations and 1000 permutations. The data are divided into groups A and B, such that: unbalanced design I: $n_B = 3n_A$, balanced design II: $n_A = n_B$, and unbalanced design III: $n_A = 3n_B$. The relationship among the group variances is achieved through $a \in \{1, 3, 9\}$, such that $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}$.

			a		
Design	Model	Test	1	3	9
I	(4.31)	OLS	5.2 (4.3;6.3)	1.2 (0.8;1.8)	0.3 (0.1;0.7)
I	(4.31)	WLS	5.7 (4.8;6.8)	6.5 (5.5;7.7)	5.4 (4.5;6.5)
I	(4.33)	OLS	6.2 (5.2;7.3)	2.5 (1.9;3.3)	2 (1.5;2.7)
I	(4.33)	WLS	6.5 (5.5;7.7)	5.4 (4.5;6.5)	5.6 (4.7;6.7)
II	(4.31)	OLS	5.2 (4.3;6.3)	5.2 (4.3;6.3)	4.7 (3.9;5.7)
II	(4.31)	WLS	5.4 (4.5;6.5)	5.8 (4.9;6.9)	5.4 (4.5;6.5)
II	(4.33)	OLS	4.3 (3.5;5.3)	6.3 (5.3;7.5)	7.3 (6.2;8.5)
II	(4.33)	WLS	4.1 (3.3;5.1)	5.7 (4.8;6.8)	6.4 (5.4;7.6)
III	(4.31)	OLS	5.6 (4.7;6.7)	13.1 (11.7;14.7)	20.3 (18.6;22.1)
III	(4.31)	WLS	6.0 (5.0;7.1)	5.5 (4.6;6.6)	6.9 (5.9;8.1)
III	(4.33)	OLS	4.3 (3.5;5.3)	15.6 (14.1;17.3)	24.7 (22.9;26.6)
III	(4.33)	WLS	5.0 (4.1;6.0)	6.5 (5.5;7.7)	5.9 (5.0;7.0)

4.6.3. Evaluating the precision of the estimates of the regression functions

In this section, the precision of the OLS and WLS estimators of regression functions $\beta_\ell(t), \ell \in \{0, 1, \ldots, L\}$, is studied. As an optimality criterion, the L^2-norm of the difference between the true $\beta_\ell(t)$ and its estimate $\widehat{\beta}_\ell(t)$ is computed for each iteration of the simulations and then the mean norm of the difference is used as an overall criterion. In other words, for each ℓ, we compute

$$\bar{L}^2_{\text{dif}} = \frac{1}{N_{\text{iter}}} \sum_{i=1}^{N_{\text{iter}}} \|\beta_\ell(t) - \widehat{\beta}_\ell(t)\|^2, \tag{4.34}$$

where $\ell \in \{0, \ldots, L\}$ and N_{iter} represents the number of iterations. Naturally, the closer $\bar{L}^2_{\text{dif}}$ is to zero, the better the regression function is estimated. Indeed, for $\bar{L}^2_{\text{dif}} = 0$, the true regression function and its estimate are exactly the same.

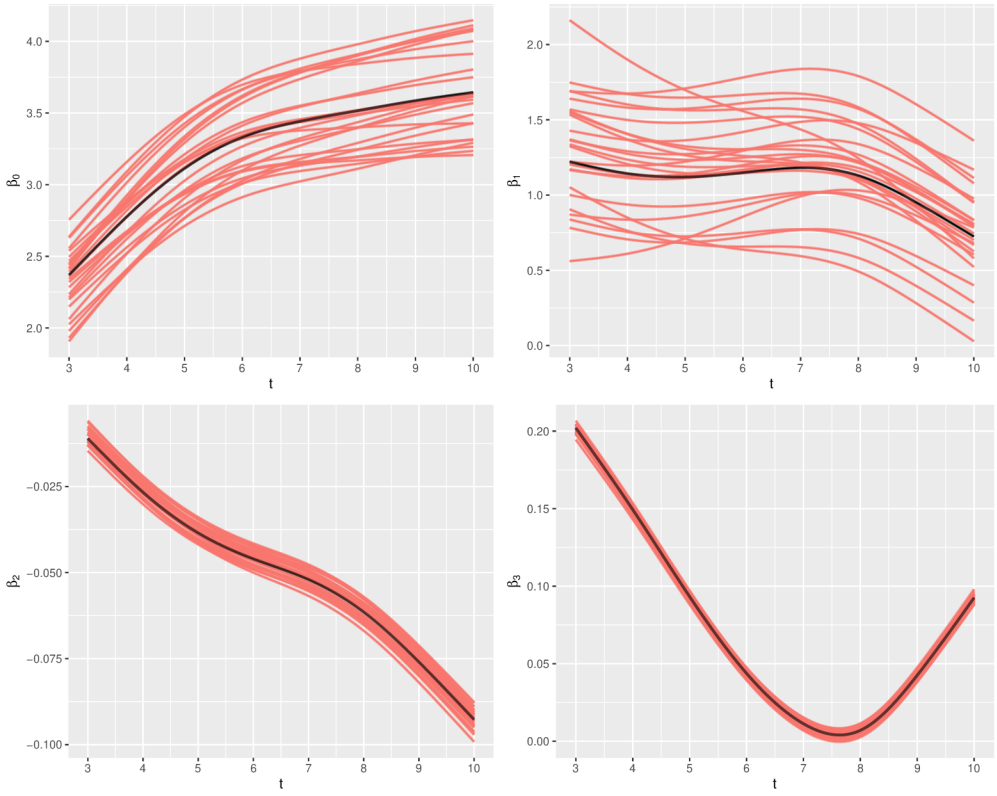


Figure 4.4: The true regression functions $\beta_\ell(t), \ell \in \{0,1,2,3\}$, (black) and their estimates $\widehat{\beta}_\ell(t), \ell \in \{0,1,2,3\}$, (red) for the more complex model (4.33). For the sake of clarity, only 25 estimates are shown.

The results obtained from 200 simulations for the sample sizes $n \in \{20, 40, 200, 500\}$ for designs I, II, and III for the model

$$\mathcal{X}_{s(m)}(t) = \beta_0(t) + \beta_1(t)\text{soil}(s) + \beta_2(t)\text{dist}(s) + \beta_3(t)\text{soil}(s)\cdot\text{dist}(s) + \delta_{s(m)}(t), \quad (4.35)$$

where $m \in \{A, B\}, s \in D$, and $t \in [3, 10]$, are presented in Table 4.3. The heteroscedasticity in the data is achieved through $\sigma^2_{(B)} = 3 \cdot \sigma^2_{(A)}$.

It can be clearly seen from the results in Table 4.3 that $\bar{L}^2_{\text{dif}}$ tends to decrease with the increasing sample size. Focusing on the comparison of the regression functions, $\widehat{\beta}_2(t)$ and $\widehat{\beta}_3(t)$ are very close to the true regression functions. On the other hand, the regression function $\beta_1(t)$ seems to be the most difficult to estimate, probably because the covariate soil(s) is a group indicator only taking

values 0 and 1. Despite that, the main shape of the estimated regression functions is preserved for each regression function, including $\beta_1(t)$.

Under heteroscedasticity, the correct estimates of the variance-covariance matrices of the OLS and WLS estimators are

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\widehat{\mathbf{W}}^{-1}\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}, \tag{4.36}$$

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = (\mathbf{F}'\widehat{\mathbf{W}}\mathbf{F})^{-1}, \tag{4.37}$$

respectively, and the difference $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) - \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}})$ is a positive semidefinite matrix. For the combination of predictors chosen in the simulation study presented in this section, the accuracy of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ is very close to the accuracy of $\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$.

On contrary, under homoscedasticity, we have

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \widehat{\sigma}^2(\mathbf{F}'\mathbf{F})^{-1}, \tag{4.38}$$

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}) = \widehat{\sigma}^2(\mathbf{F}'\widehat{\mathbf{W}}\mathbf{F})^{-1}\mathbf{F}'\widehat{\mathbf{W}}^2\mathbf{F}(\mathbf{F}'\widehat{\mathbf{W}}\mathbf{F})^{-1}, \tag{4.39}$$

and the difference $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}) - \widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})$ is a positive semidefinite matrix. Again, in our case, the accuracy of $\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ is very close to the accuracy of $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$. Thus, the differences between the variance-covariance matrices are very small for both homoscedastic and heteroscedastic case and are not reported here.

Note that, these findings do not represent a general rule. Consider for example the model without an interaction term

$$\mathcal{X}_{s(m)}(t) = \beta_0(t) + \beta_1(t)\mathrm{soil}(s) + \beta_2(t)\mathrm{dist}(s) + \delta_{s(m)}(t), \tag{4.40}$$

where $m \in \{A, B\}, s \in D$, and $t \in [3, 10]$. Under heteroscedasticity, the sum of the variances of the OLS estimators is approximately 1.07 times higher than for the WLS estimators. On contrary, under homoscedasticity, the sum of the variances of the WLS estimators is almost 1.09 times higher than for the OLS estimators.

Table 4.3: the mean norms of the regression functions for model (4.35) under different conditions for 200 simulations. The data are divided into groups A and B, such that: unbalanced design I: $n_B = 3n_A$, balanced design II: $n_A = n_B$, and unbalanced design III: $n_A = 3n_B$. The relationship among the group variances is achieved through $a \in \{1, 3, 9\}$, such that $\sigma^2_{(B)} = a \cdot \sigma^2_{(A)}$.

design	n	$\beta_0(t)$	$\beta_1(t)$	$\beta_2(t)$	$\beta_3(t)$
I	20	79.315	128.092	0.764	0.845
I	40	36.289	71.85	0.102	0.118
I	200	6.572	12.685	0.001	0.001
I	500	2.377	5.028	$< 10^{-4}$	$< 10^{-4}$
II	20	37.182	131.506	0.112	0.426
II	40	17.311	70.344	0.013	0.055
II	200	3.164	12.286	$< 10^{-4}$	$< 10^{-4}$
II	500	2.377	5.028	$< 10^{-4}$	$< 10^{-4}$
III	20	21.521	255.984	0.031	2.258
III	40	9.252	128.681	0.003	0.363
III	200	2.046	21.917	$< 10^{-4}$	0.003
III	500	2.377	5.028	$< 10^{-4}$	$< 10^{-4}$

4.7. Simulation studies under spatial correlation

In this section, we report the results of a simulation study intended to evaluate the empirical size and the power of the permutation-based global tests $\mathrm{T}_{\mathrm{OLS}}$ and $\mathrm{T}_{\mathrm{GLS}}$ for the significance of the effect of the covariates in the spatial functional regression model under the presence of spatial correlation (Section 4.1).

The spatially correlated residuals are generated by a two-step procedure. In the first step, the uncorrelated residuals $\varepsilon_{s_1}(t), \ldots, \varepsilon_{s_n}(t), s_1, \ldots, s_n \in D$, $t \in [1, 8]$, are simulated as follows. Denote by $\{\phi_q,\ q \in \{1, \ldots, 10\}\}$ the cubic B-spline basis with knots at the data points, used to represent the data, and by $\mathbf{u} = (u_1, \ldots, u_{10})'$ the corresponding coefficients vector. The residuals $\varepsilon_{s_i}(t)$, $s_i \in D, i \in \{1, \ldots, n\}, t \in [1, 8]$, are built on the same cubic B-spline basis expansion as

$$\varepsilon_{s_i}(t) = \sum_{q=1}^{10} u_q^i \phi_q(t), s_i \in D,\ i \in \{1, \ldots, n\}, t \in [1, 8], \tag{4.41}$$

where the basis coefficients u_q^i are generated using the standard normal distribution for every location $s_i = (x_i, y_i)$ in the space $D = [0, 1] \times [0, 1]$. In order to

obtain the spatially correlated residuals $\delta_{s_i}(t), s_i \in D, i \in \{1,\ldots,n\}, t \in [1,8]$, the specification of the variance-covariance matrix $\boldsymbol{\Sigma}$ is necessary. Recall that, under the isotropy condition (see Equation (4.5)), one can assume $\boldsymbol{\Sigma} = \sigma^2\rho(h)$. In our case, we assume $\sigma^2 = 100$ and a spherical model for the correlation function $\rho(h)$, i.e.,

$$\rho(h) = \begin{cases} 1 - 1.5\left(\frac{h}{\psi}\right) + 0.5\left(\frac{h}{\psi}\right)^3, & h < \psi, \\ 0, & \text{otherwise}, \end{cases} \tag{4.42}$$

where ψ is a range, i.e., the maximum distance to which the spatial observations are correlated, and $h = \|s_i - s_j\|_D$ is the distance between the points s_i and s_j. Then, to obtain the spatially correlated residuals $\delta_{s_i}(t)$, we put $\boldsymbol{\delta} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\varepsilon}$, i.e.,

$$\boldsymbol{\delta} = \left\{ \boldsymbol{\Sigma}^{1/2} \begin{pmatrix} \varepsilon_{s_1}(t) \\ \vdots \\ \varepsilon_{s_n}(t) \end{pmatrix} \right\}. \tag{4.43}$$

In the second step, the correlated residuals are added to the model for the drift (the structure of the drift is for each simulation scenario specified in the following part of this section), obtaining the spatial functional observations $\mathcal{X}_{s_i}(t)$, $s_i \in D, i \in \{1,\ldots,n\}, t \in [1,8]$. The true regression functions $\beta_0(t), \beta_1(t)$, and $\beta_2(t)$ are shown in Figure 4.5.

4.7.1. Assessing the empirical size of the T_{OLS} and T_{GLS} tests

For clarity, in the rest of this section, the notation $\mathcal{X}(t;x,y)$ and $\delta(t;x,y)$ is used instead of the short forms $\mathcal{X}_s(t)$ and $\delta_s(t)$. To evaluate the empirical size of the test, the hypothesis of the lack of effect of the covariates is tested. More precisely, the model is built as

$$\mathcal{X}(t;x,y) = \beta_0(t) + \beta_1(t)x + \beta_2(t)y + \delta(t;x,y), \tag{4.44}$$

testing the hypothesis

$$H_0 : \beta_1(t) = \beta_2(t) = 0 \; \forall t \in [1,8], \tag{4.45}$$

$$H_1 : \beta_\ell(t) \neq 0 \text{ for some } \ell \in \{1,2\} \text{ and some } t \in [1,8]. \tag{4.46}$$

Table 4.4: The results of the simulation studies performed for the model (4.44) testing the hypothesis (4.45). The empirical sizes of the omnibus T_{OLS} and T_{GLS} tests, together with the 95% confidence intervals for the empirical size are displayed, considering the nominal size of the test $\alpha = 0.05$ for the different sample sizes n and the different nominal semivariogram ranges ψ, based on 500 simulations and 1000 permutations.

		ψ			
n	Test	0.10	0.25	0.50	0.75
50	OLS	0.082 (0.061;0.109)	0.336 (0.296;0.379)	0.658 (0.615;0.698)	0.804 (0.767;0.836)
50	GLS	0.034 (0.021;0.054)	0.056 (0.039;0.080)	0.014 (0.007;0.029)	0.036 (0.023;0.056)
100	OLS	0.134 (0.107;0.167)	0.492 (0.448;0.536)	0.788 (0.750;0.822)	0.888 (0.857;0.913)
100	GLS	0.042 (0.028;0.063)	0.060 (0.042;0.084)	0.034 (0.021;0.054)	0.022 (0.012;0.039)
200	OLS	0.232 (0.197;0.271)	0.678 (0.636;0.717)	0.890 (0.860;0.915)	0.946 (0.923;0.963)
200	GLS	0.048 (0.032;0.070)	0.048 (0.032;0.070)	0.026 (0.015;0.044)	0.016 (0.008;0.031)
400	OLS	0.436 (0.393;0.480)	0.832 (0.797;0.862)	0.942 (0.918;0.959)	0.978 (0.961;0.988)
400	GLS	0.050 (0.034;0.073)	0.068 (0.049;0.094)	0.036 (0.023;0.056)	0.040 (0.026;0.061)

Considering an 100×100 equispaced grid on $D = [0, 1] \times [0, 1]$, n functional residuals are randomly drawn and their covariance structure is estimated. In the simulation study evaluating the empirical size of the test, the sample sizes $n \in \{50, 100, 200, 400\}$ and the range parameters $\psi \in \{0.10, 0.25, 0.50, 0.75\}$ are considered. For the overview of the theoretical spherical model for the semivariograms used in this study, see Figure 4.5. For each of the total 16 scenarios, 500 simulations are performed, focusing on the empirical size of the test and on the comparison to the OLS approach. The nominal level of the significance is set to $\alpha = 0.05$. The empirical sizes of the T_{OLS} and T_{GLS} tests, together with the 95% confidence intervals, are summarized in Table 4.4.

Generally speaking, the empirical size of the T_{GLS} test is close to the nominal one, slightly worsening (in a conservative direction) for the data correlated over the long distances (the range of the variogram $\psi \in \{0.50, 0.75\}$). This effect is suppressed for a sufficiently large sample size. On the other hand, the empirical size of the T_{OLS} test is extremely liberal for majority of the scenarios. Moreover, with increasing ψ, the empirical size of the T_{OLS} test approaches to 1. Neglecting the spatial correlation leads to very unreliable results regarding the significance of the regression functions being tested.

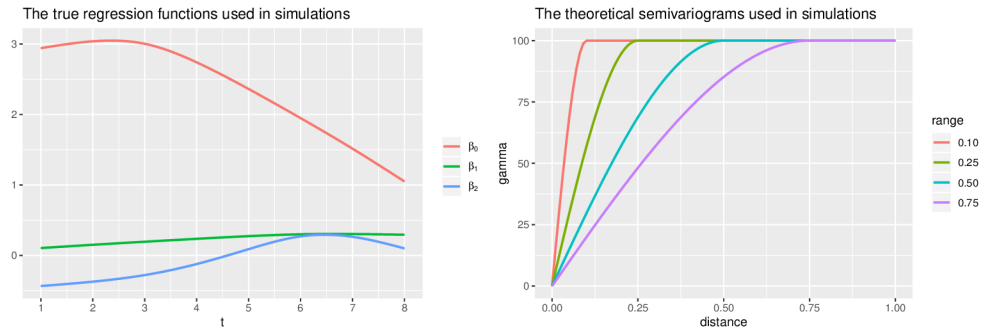


Figure 4.5: Left: The true regression functions $\beta_0(t)$, $\beta_1(t)$, and $\beta_2(t)$. Right: The theoretical spherical models for the semivariogram with different ranges used in the simulation studies.

4.7.2. Assessing the empirical size and the power of the $\mathrm{T}^{\ell}_{\mathrm{GLS}}$ test

In this section, the results of the simulation studies intended to evaluate the empirical size and the power of the $\mathrm{T}^{\ell}_{\mathrm{GLS}}$ test are presented for the sample size $n = 100$ and the range parameters $\psi \in \{0.1, 0.25\}$. For each case, 500 simulations are performed, evaluating the power of the $\mathrm{T}^{\ell}_{\mathrm{GLS}}$ test on the nominal level of significance $\alpha = 0.05$.

Specifically, we focus on three different situations; (i) a model with two spatial covariates, (ii) a model with one spatial covariate, (iii) a scenario, where the data-generative model has one spatial covariate (similarly to (ii)), but the model assumed and fitted includes two spatial covariates (as in (i)). In more detail, we consider the following models:

(i) Two-covariates model

$$\mathcal{X}(t; x, y) = \beta_0(t) + b \cdot \beta_1(t)x + \beta_2(t)y + \delta(t; x, y); \tag{4.47}$$

(ii) One-covariate model

$$\mathcal{X}(t; x, y) = \beta_0(t) + b \cdot \beta_1(t)x + \delta(t; x, y); \tag{4.48}$$

(iii) Misspecified model

$$\text{data-generative model: } \mathcal{X}(t;x,y) = \beta_0(t) + b \cdot \beta_1(t)x + \delta(t;x,y); \quad (4.49)$$

$$\text{fitted model: } \mathcal{X}(t;x,y) = \beta_0(t) + b \cdot \beta_1(t)x + \beta_2(t)y + \delta(t;x,y).$$

In general, $b \in [0,\infty)$, and a hypothesis $H_0^1 : \beta_1(t) = 0 \, \forall t \in [1,8]$, against $H_1^1 : \beta_1(t) \neq 0$ for some $t \in [1,8]$, is tested on the level of significance $\alpha = 0.05$. The power of the single-parameter test $\mathrm{T}^1_{\mathrm{GLS}}$ (4.25) is evaluated for different values of the parameter b. The power functions can be seen in Figure 4.6.

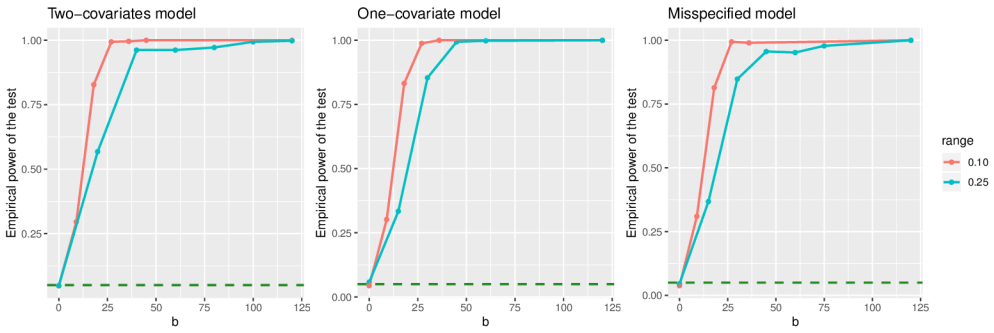


Figure 4.6: The empirical power of the $\mathrm{T}^1_{\mathrm{GLS}}$ test for models (4.47), (4.48), and (4.49) based on 500 simulations and 1000 permutations for the sample size $n = 100$. The magnitude of β_1 is modelled through parameter b as $b \cdot \beta_1$. The dashed horizontal line represents the nominal size of the test $\alpha = 0.05$.

Overall, for all the three scenarios (i), (ii), and (iii), the power functions have a similar shape and the test on the data with a less-correlated structure has more power. As expected, the power increases with the increase of b, the parameter controlling the amplitude of the regression function being tested.

4.8. Application: Analysis of the soil samples collected on the border between the field and the forest

In this section, the methodology introduced in Chapter 4 is demonstrated on a data set of the geochemical measurements introduced in Section 1.2. These

data are a result from the analysis of the soil samples collected during the growing seasons (March-October) of the years 2015 and 2016 at a site located near Křtiny, Czech Republic. The total area of the site is 9131.4 m^2. Its surface is formed by the brown earth and the cambisol. The altitude of the site ranges from 524.3 to 529.3 m with a mean altitude of 526.8 m. The slope ranges from 0.05° to 16.32° with a mean slope of 2.7°. The surface of the site is oriented to the southwest. The agricultural soil covers 60.5% of the area, and 16% is covered by the beech trees. The remaining part of the site is covered mainly by the spruces and by the fixed coniferous forest.

The analysis of the samples was conducted in the certified laboratories. The carbon samples were carried out through the method of oxidation of the sample chromium mixture and a subsequent measurement on a spectrophotometer. Following the methodology of the Central Institute for Supervising and Testing in Agriculture, Czech Republic [77], the combustion tubes were replaced by 100 ml glass flasks, heated on a heating plate instead of using a steam distiller. The measured values represent the amount of the oxidisable carbon in the weight percentage of the dry matter. The active soil reaction was determined using a pH meter in the soil water suspension with a glass ion-selective electrode after the suspension was previously shaken on a horizontal mechanical shaker for one hour followed by one-hour rest. Just before the measurement itself, the suspension was briefly stirred with a glass bar [76].

The measurements taken from each sampling location are modelled as the functions of time distributed over a one-dimensional spatial domain. Using the `R` package `FDA` [61, 63], raw monthly measurements were turned into the functional observations through a projection over a cubic B-spline basis with the knots placed at the data points (i.e., eight knots). The data were smoothed using the penalized residual sum of squares criterion with the smoothing parameter $\lambda = 10$ selected by the generalized cross-validation [62].

Note that, the data measured at the ecotone were excluded from the data set. Indeed, the ecotone is a sampling point having neither the properties of the forest

soil nor of the field soil; moreover, its influence disappears at approximately 3 m distance. Furthermore, the measurements from the growing seasons in the years 2015 and 2016 do not differ from the geochemical point of view; therefore, the data can be analysed together.

4.8.1. Modelling the KCl pH

The KCl pH data set contains 20 functional observations, that is, 10 for the field soil and 10 for the forest soil. For the KCl pH data, we consider the model

$$\mathcal{X}_s^{\mathrm{KCl}}(t) = \beta_0(t) + \beta_1(t)\mathrm{soil}(s) + \delta_s(t), \;\; s \in D, \;\; t \in [3, 10], \tag{4.50}$$

where the interval $T = [3, 10]$ denotes the months from March to October, D is a set of the sample spaces and $\mathrm{soil}(s)$ is an indicator function for the type of soil, taking values of 0 for the observations from the field soil and 1 for the observations from the forest soil, that is,

$$\mathrm{soil}(s) = \begin{cases} 0 \text{ for the field sample spaces,} \\ 1 \text{ for the forest sample spaces.} \end{cases} \tag{4.51}$$

The regression function $\beta_0(t)$ represents, $\forall t \in [3, 10]$, the mean KCl pH for the field observations. The regression function $\beta_1(t)$ can be interpreted, $\forall t \in [3, 10]$, as the difference between KCl pH from the field and the forest parts of the site.

To estimate the spatial structure of the data, the `R` package `fdagstat` [37] can be used. The fitted model (4.50) for the drift using the OLS estimation procedure and the semivariogram of the residuals, are displayed in Figure 4.7. The residuals show a pure nugget structure, and thus, they can be considered spatially uncorrelated [19].

As the next step, the homoscedasticity of the residuals from the field and the forest is verified. Let $m \in \{A, B\}$ denote the field and the forest parts of the site, respectively. The hypothesis of homoscedasticity is tested using the permutation procedure from Section 4.5. The global p-value of the test, estimated from 1000 random permutations, is equal to 0.001, and, consequently, the hypothesis of homoscedasticity among the residuals is rejected. In fact, although the residuals

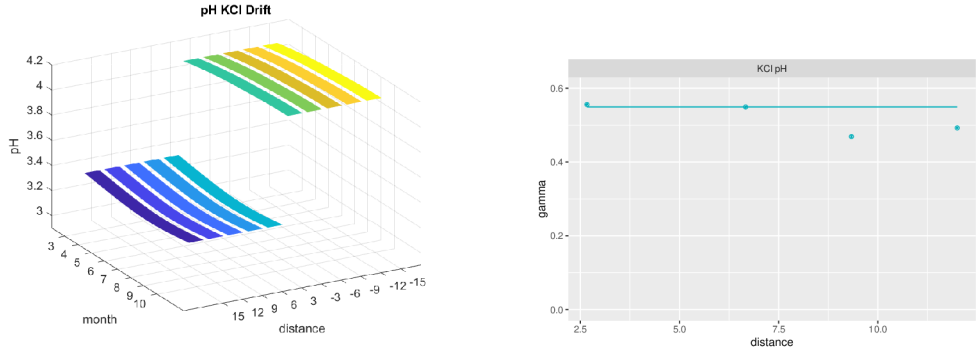


Figure 4.7: Left: The fitted model (4.50) for the drift. Right: The semivariogram of the residuals from model (4.50).

are spatially uncorrelated, the spatial position has an effect on the residuals in terms of the variance. Hence, the residuals are not exchangeable over D, and the inferential procedure based on the $\mathrm{T}_{\mathrm{OLS}}$ test would not be correct.

We thus opt for the heteroscedastic setting and decompose the residuals $\delta_{s(m)}(t)$ as $\sigma_{(m)}\varepsilon_s(t), s \in D, m \in \{A, B\}, t \in [3, 10]$, where $\sigma_{(m)}$ is the standard deviation for the m-th type of soil (Equation (4.27)), and $\varepsilon_s(t), s \in D, t \in [3, 10]$, are the standardized residuals, approximately exchangeable over D. We also assume that $\sigma_{(m)}, m \in \{A, B\}$, are constant on the corresponding part of D.

In particular, the variances $\sigma^2_{(m)}$, $m \in \{A, B\}$, are estimated as $\widehat{\sigma}^2_{(A)} = 14.36$ and $\widehat{\sigma}^2_{(B)} = 3.32$. Note that, the variance of the observations from the field soil is more than four times higher than that of the observations from the forest soil.

Having estimated the data variances under heteroscedasticity, model (4.50) is fitted to the observations through the WLS method. The fitted model (4.50) can be seen in Figure 4.8 ($\mathrm{RMSE} = 2.973$, $\mathrm{NRMSE} = 0.074$, and $\mathrm{R}^2 = 0.707$).

The effect of the type of soil on the KCl pH is tested using the permutation scheme for heteroscedastic data (Algorithm 4), based on 1000 permutations, obtaining a p-value equal to 0. In conclusion, the type of soil has a significant influence on the KCl pH, which tends to be more acidic in the forest part than in the field part of the site (the mean KCl pH ranging from 3.23 to 3.29 in the forest soil, and from 4.03 to 4.12 in the field soil; see Figure 4.8).

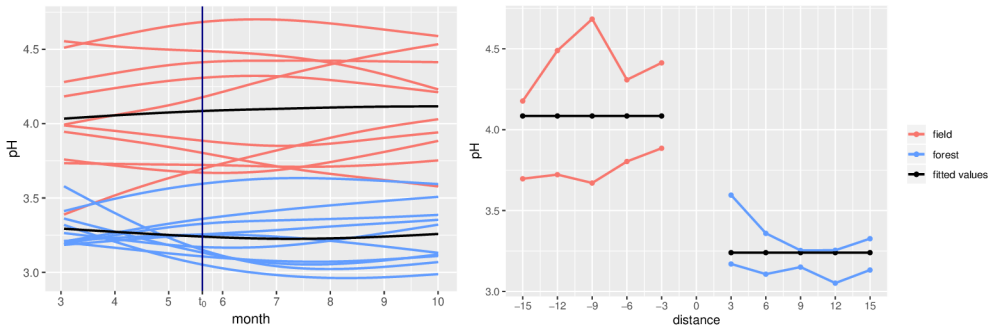


Figure 4.8: The KCl pH data, weighted least-squares model (4.50). Left: The original functional observations (red: field; blue: forest) and the fitted model (black). Right: A view of the data in the space at a fixed t_0 (specified by the vertical line in the figure on the left). The points indicate the section of the curves at t_0, and are for clarity connected by lines

4.8.2. Modelling the H_2O pH

For the H_2O pH measurements, 20 functional observations are given, evenly distributed on the field and the forest parts of the site. We consider the model

$$\mathcal{X}_s^{H_2O}(t) = \beta_0(t) + \beta_1(t)\text{soil}(s) + \delta_s(t), \ s \in D, \ t \in [3, 10], \tag{4.52}$$

for the H_2O pH observations. The fitted model (4.52) can be seen in Figure 4.9 (RMSE = 2.143, NRMSE = 0.045, and $R^2 = 0.910$). The residuals $\delta_s(t)$ show the pure nugget structure and can be considered spatially uncorrelated (not shown). A preliminary test on the variances in the two groups shows that the setting is homoscedastic in this case (p-value = 0.122). The overall variance of the data is estimated as $\widehat{\sigma}^2 = 4.59$. This result allows to test the effect of the type of soil via the T_{OLS} test, following the procedure in which directly the estimated residuals $\widehat{\delta}_s(t)$ are permuted (Algorithm 3). In total, 1000 permutations are performed, leading to the global p-value equal to 0. In conclusion, the mean H_2O pH is significantly different in the field (range: 4.93-5.00) and forest (range: 3.65-3.83) parts of the site (p-value = 0). Similarly as in case of the KCl pH, the H_2O pH is more acidic in the forest part than in the field part of the site

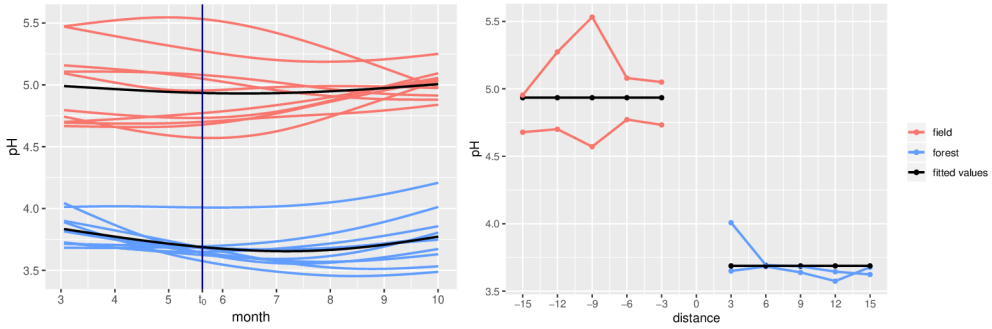


Figure 4.9: The H_2O pH data, ordinary least-squares model (4.52). Left: The original functional observations (red: field, blue: forest) and the fitted model (black). Right: A view of the data in the space at a fixed t_0 (specified by the vertical line in the figure on the left). The points indicate the section of the curves at t_0, and are for clarity connected by lines.

4.8.3. Modelling the percentage of organic carbon

Unlike in the previous cases, the carbon measurements are unbalanced. In particular, the data set contains 15 functional observations, where 10 were measured in the field and five were measured in the forest. For the percentage of carbon, consider the model

$$\mathcal{X}_s^{\mathrm{C}}(t) = \beta_0(t) + \beta_1(t)\mathrm{soil}(s) + \beta_2(t)\mathrm{dist}(s) + \delta_s(t), \;\; s \in D, \;\; t \in [3, 10], \qquad (4.53)$$

where $\mathrm{dist}(s) \in \{3, 6, 9, 12, 15\}$ denotes the distance of the sampling point from the ecotone. Unlike in the case of the pH observations, the indicator of the type of soil itself is insufficient in capturing the spatial trend in the data in the case of the carbon measurements. The results of the variography again show that the data are not spatially correlated (not shown). The test for the equality of the variances for the field and the forest parts of the site leads to the global p-value of the test (4.29) equal to 0.055; thus, we cannot reject the hypothesis of the equality of variances. However, in light of the simulation results, we opt for considering a heteroscedastic modelling setting, the latter being more robust than that based on the assumptions of homoscedasticity (see Section 4.6). We thus fit the model (4.53) by WLS (RMSE = 2.945, NRMSE = 0.083, and $R^2 = 0.857$). Furthermore, we consider a permutation test based on the standardized residuals

$\varepsilon_s(t) = \delta_{s(m)}(t)/\sigma_{(m)}, s \in D, m \in \{A, B\}, t \in [3, 10]$, where $\widehat{\sigma}^2_{(A)} = 10.85$ and $\widehat{\sigma}^2_{(B)} = 4.06$.

On the basis of the permutational test for the significance of the covariates under heteroscedasticity, we conclude that the percentage of carbon is significantly affected by the covariates (p-value = 0), again displaying a significant difference between the field and the forest parts of the site: The mean percentage of carbon is approximately 1.4 times higher in the forest than in the field, and it decreases by approximately 0.04% per 3 metres with the increasing distance from the ecotone in both parts of the site. Figure 4.10 reports the results for the estimated model (4.53).

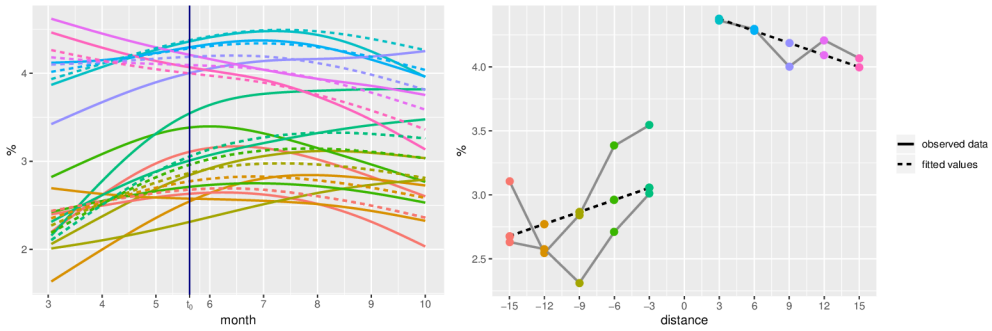


Figure 4.10: The percentage of carbon data, weighted least-squares model (4.53). Left: The original functional observations (solid lines) and the fitted model (dashed lines). Each colour represents an observation from a different sample point. Right: A view of the data in the space at a fixed t_0 (specified by the vertical line in the figure on the left). The points indicate the section of the curves at t_0, and are for clarity connected by grey lines. The fitted values are for clarity connected by a dashed black line.

4.9. Application: Analysis of the production of the municipal waste in the Venice province

In this section, the methodology introduced in Chapter 4 is demonstrated on a real data set including the yearly measurements of the per capita amount of the municipal waste from the years 1997 to 2011 in 49 cities of the Venice province.

The per capita municipal waste production from the years 1997 to 2011 was preprocessed through a projection over a cubic B-spline basis with knots placed at the sampling time (i.e., 15 knots) [62] using the `R` package `fda` [61, 63]. To further smooth the data, the penalized residual sum of squares criterion was used with the smoothing parameter $\lambda = 1$ chosen to be consistent with the previous study on this data [15], obtaining the functional observations over time, distributed over a two-dimensional geographical space.

Recall that, the same data set was used by [15], where the focus was on fitting a spatiotemporal regression model with differential regularization [14, 46]. We here focus on testing the significance of the covariates within a spatial functional regression model inspired by that of [15], but fitted by GLS (Section 4.3). We here model the production of the waste through the spatial coordinates, indicating the location of the town centre, and also through the covariate accounting for the tourism. For this purpose, we consider the proportion of the number of the beds in the accommodation facilities to the number of the residents, averaged along the years 1997-2011. It is of interest to evaluate the effect of these three covariates on the amount of the produced municipal waste. We consider the model

$$\mathcal{X}_s(t) = \beta_0(t) + \beta_1(t)\text{latitude} + \beta_2(t)\text{longitude} + \beta_3(t)\text{Nbeds} + \delta_s(t), \qquad (4.54)$$

where $s = [\text{latitude}, \text{longitude}] \in D$, and $t \in [1997, 2011]$. Following the methodology from Sections 4.1 - 4.3, firstly, an OLS model is fitted to the data. This is used to provide an initial guess of the spatial structure of the residuals, yielding to an initial estimate of the empirical semivariogram using the `R` package `fdagstat` [37]. Based on a visual inspection of the semivariogram of the residuals, a Gaussian model with a nugget is chosen for the semivariogram. As the residuals show a correlated structure, it would be inappropriate to further use the OLS to fit the model, although this method is used as a comparison to the GLS approach later in this section. Consistently with the evident spatial structure existing among the residuals, Algorithm 2 is used to fit the GLS model to the data. The overview of the estimated model (4.54) is provided in Figures 4.11 and 4.12 (RMSE = 1120.521, NRMSE = 0.194, and $R^2 = 0.869$).

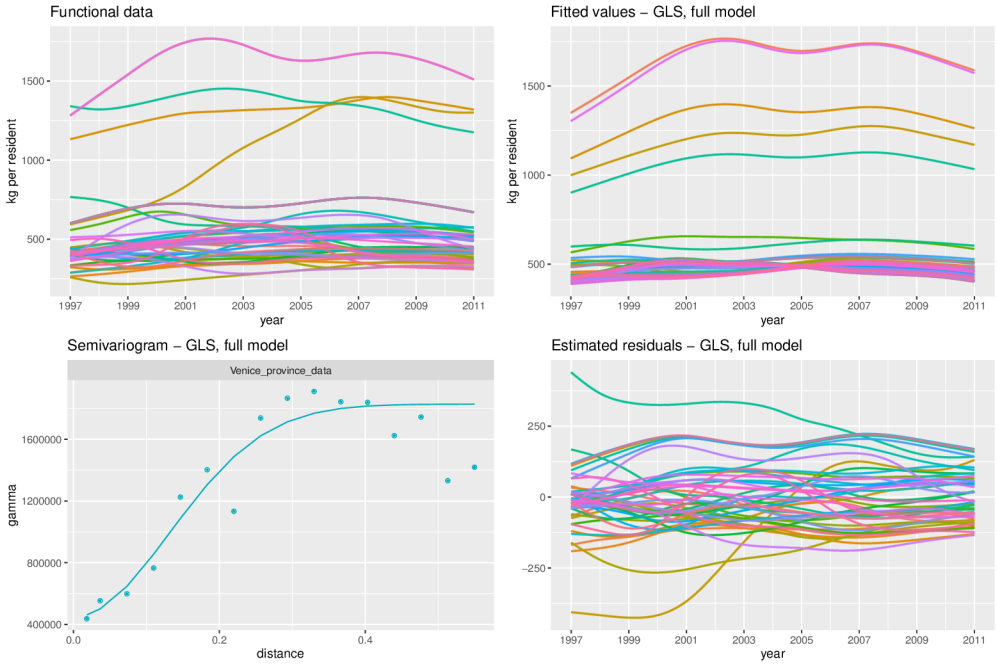


Figure 4.11: Top left: The amount of the municipal waste produced in the Venice province (kg per resident) modelled as the functional observations. Top right: The values fitted by the full model (4.54). Bottom left: The empirical semivariogram together with the estimated Gaussian model for the semivariogram for the spatial functional residuals from the full model (4.54). Bottom right: The estimated residuals of the full model (4.54).

Figure 4.12: Left: The generalized least-squares estimates of the regression functions for the full model (4.54) for the amount of the municipal waste produced in the Venice province (kg per resident). Right: A detailed view of the regression functions $\beta_1(t), \beta_2(t)$ and $\beta_3(t)$ for the full model (4.54).

The effect of the covariates in model (4.54) is tested via the spatial functional Freedman and Lane permutation scheme introduced in Section 4.4. Following the SF-FLPerm Algorithm 5, the residuals are decorrelated, resampled, and then are correlated back, obtaining the permuted responses. The focus is now on the hypotheses (4.15) and (4.18) using the T_{GLS} and T^{ℓ}_{GLS} test statistics. In particular, the tested hypotheses are the following:

(i) $H_0 : \beta_1(t) = \beta_2(t) = \beta_3(t) = 0 \; \forall t \in [1997, 2011]$, against
$H_1 : \beta_\ell(t) \neq 0$ for some $\ell \in \{1, 2, 3\}$ and some $t \in [1997, 2011]$;

(ii) $H_0^1 : \beta_1(t) = 0 \; \forall t \in [1997, 2011]$, against
$H_1^1 : \beta_1(t) \neq 0$ for some $t \in [1997, 2011]$;

(iii) $H_0^2 : \beta_2(t) = 0 \; \forall t \in [1997, 2011]$, against
$H_1^2 : \beta_2(t) \neq 0$ for some $t \in [1997, 2011]$;

(iv) $H_0^3 : \beta_3(t) = 0 \; \forall t \in [1997, 2011]$, against
$H_1^3 : \beta_3(t) \neq 0$ for some $t \in [1997, 2011]$.

For the sake of the comparison, the same hypotheses are also tested under the OLS settings using the T_{OLS} and T^{ℓ}_{OLS} test statistics, following the standard Freedman and Lane permutation scheme for homoscedastic data (Algorithm 3), where directly the spatially correlated, and thus nonexchangeable, residuals are resampled. The results for both approaches, based on 1000 permutations, are summarized in Table 4.5. Starting with the GLS approach, the omnibus hypothesis, as well as the hypothesis on the regression function $\beta_3(t)$, are rejected on the level of significance $\alpha = 0.05$. The hypotheses related to the spatial covariates latitude and longitude are not rejected, concluding that when the covariate Nbeds is included, the actual location of the city in the Venice province does not have a significant effect on the amount of the produced municipal waste. Instead, the relative number of the beds in the accommodation facilities significantly increases the municipal waste production ($\beta_3(t)$ is positive $\forall t \in [1997, 2011]$, see Figure 4.12). In the OLS case, the omnibus, $\beta_2(t)$ and $\beta_3(t)$ tests lead to the

Table 4.5: The global p-values of the $T_{OLS}, T^{\ell}_{OLS}, T_{GLS}$ and T^{ℓ}_{GLS} tests based on 1000 permutations for a data set on the amount of the municipal waste produced in the Venice province of Italy (kg per resident), model (4.54).

	omnibus	$\beta_1(t)$	$\beta_2(t)$	$\beta_3(t)$
OLS	0	0.073	0.153	0
GLS	0	0.656	0.753	0

same result, while the hypothesis on the regression function $\beta_1(t)$ is rejected on the level of significance $\alpha = 0.1$. Note that, in the case of $\beta_1(t)$ and $\beta_2(t)$, the global p-values for the OLS tests are very low, compared to the GLS results. This finding is consistent with the conclusion from the simulation studies; the OLS approach, under the assumption of a general variance-covariance matrix $\mathbf{\Sigma}$, shows a very liberal behaviour regarding the empirical size of the test (see Section 4.7). The estimates of the regression functions in model (4.54) are shown in Figure 4.12.

Based on these results, a final generalized least-squares model

$$\mathcal{X}_s(t) = \beta_0(t) + \beta_1(t)\text{Nbeds} + \delta_s(t), s \in D, t \in [1997, 2011] \tag{4.55}$$

is proposed to model the amount of the produced municipal waste (Figure 4.13), (RMSE = 1160.206, NRMSE = 0.201, and R^2 = 0.860). One can see that the relative number of the beds in the accommodation facilities is a significant factor for the increase of the municipal waste production; $\beta_1(t)$ is positive $\forall t \in [1997, 2011]$ (Figure 4.13).

Figure 4.13: Left: The generalized least-squares estimates of the regression functions for the final model (4.55) for the amount of the municipal waste produced in the Venice province (kg per resident). Right: The values fitted by the final model (4.55).

Discussion and conclusion

This thesis introduced a novel approach to the nonparametric inference for the functional data. The main focus was on the functional regression models, where the testing for the significance of the regression functions was performed by permuting the residuals of the model under the null hypothesis. In this setting, the aim was to ensure the exchangeability of the residuals being permuted.

Chapter 1, by a form of the motivational examples, provided a brief introduction of the real-world data sets concerned throughout this thesis. In Chapter 2, the basic aspects of functional data analysis were detailed, starting with turning the observed discrete data into functional observations by means of the basis expansion. Consistently with the case studies concerned in this thesis, a cubic B-spline basis, a common choice for fitting a nonperiodical functional observations, was described here in more detail. Secondly, the methods for smoothing the functional data were discussed, focusing on a spline smoothing through a roughness penalty method. Lastly, a random function was defined, as well as the basic summary statistics widely used for describing the functional data.

The aim of Chapter 3 was to define a regression model with the functional covariate and the functional response and the inferential procedure for testing the significance of the regression functions based on the Freedman and Lane permutation scheme (see [31]) adapted to the functional setting [1]. Here, to ensure the approximate exchangeability of the permuted objects, the residuals, rather than the observations, are of interest. From the application viewpoint, the objective was to analyse the data from the field of the transportation research, consisting of the functional observations of the driving speed on six expressway

ramps collected by a vehicle fleet on an interchange in Brno, Czech Republic. The functional nature of the data enables to explore the data features by studying the function's behaviour in first or higher order derivatives. In particular, the original data consist of the curves of the driving speed and the first derivative allows to directly see the acceleration/deceleration in a specific part of the functional domain (here representing the length of the ramp). Moreover, it was also of interest to examine how the driving speed differs in the different groups in the data. Such task can include detecting the differences in the driving speed of the vehicles driving in a central lane, and thus not needing to change their speed rapidly; and the vehicles driving in the merging or the exit lane, where there may be often necessary to rapidly slow down and then speed up back in order to merge into or exit the central lane. It is also of interest to detect the part of the functional domain where the mean driving speeds of both groups become indistinguishable, e.g., at which part of the functional domain do the vehicles joining the central lane from the merging lane reach the same mean speed as the vehicles driving in the central lane. The interval-wise testing procedure [60], summarized in Chapter 3, provides the framework suitable for dealing with this kind of the research questions. Indeed, a significant difference in the mean driving speed was shown for the vehicles driving in the central lane and in the auxiliary lane. Lastly, these findings were also confirmed by the testing performed in the regression models where the presence of the auxiliary lane was included as one of the explanatory variables, together with the curvature of the ramp (represented by the radius of the circle fitted to a specific part of the ramp), evaluated by the permutation-based test in a functional-on-functional regression model based on the Freedman and Lane's methodology. Similarly to the previous studies [22, 45], the relationship between the driving speed and the ramp curvature represented by the radius of the circle fitted to the specific part of the ramp by the least-square-circle fit method was confirmed: on the less complex ramps the, relationship between the speed and the radius was rather weak, which could be observed from the results of testing for the significance of the radius as a regression function.

The aim of Chapter 4 was to define a spatiotemporal regression model, together with an introduction of a novel inferential framework for testing in the spatial functional regression model, accounting for the more complex structures of the data. Firstly, concerning a testing procedure suitable for heteroscedastic functional data, we propose to fit a weighted least-squares functional regression model to the data, where the residuals are allowed to be characterized by heterogeneous variances across the groups. The permutational schemes of [1] and [31] are thus extended to the standardized residuals of the model, obtained by dividing the estimated residuals by the estimated standard deviation of the corresponding group of the data. Then, the standardized residuals obtained by this procedure are assumed to be approximately exchangeable. With the purpose of assessing the assumption of homoscedasticity, we also proposed a permutation-based test for the equality of the variances in two groups of the data. Note that, the proposed model and the testing procedures can be easily extended to any number $\mathcal{G} \geq 2$ of groups. In this case, the mean term would be associated with $\mathcal{G} - 1$ dummy variables, and the testing procedure could be performed (jointly) on the associated coefficients, similarly as detailed in this thesis. In this more general situation, the heteroscedasticity among the groups may be expected; the tests for heteroscedasticity could be developed by using an analysis of variance-like setting, or the multiple pairwise tests with appropriate level corrections. The performance of this proposed methodology was assessed by the extensive simulation studies showing that the introduced WLS test performs well concerning the empirical size and the power, especially in the case of the unbalanced designs. In contrast, if the nonconstant variance in the data is neglected and an ordinary least-squares approach is applied, the test is either too liberal or too conservative whenever the two populations in the data have unequal sample sizes. Based on the simulations, the empirical size of the OLS test could be more than 10 times smaller, or almost five times greater than the nominal one. However, for the balanced design, the OLS test performs well even under the presence of heteroscedasticity. These results are consistent with the previous studies pointing

out that the effect of heteroscedasticity is stronger for the unbalanced designs [16, 53]. In fact, one should pay close attention to the specific spatial structure of the data during the model fitting, especially in the case of the unbalanced design.

Secondly, we focused on a case of the spatial correlation among the observations. Under this setting, the drift is estimated by a generalized least-squares model and then the spatial correlation among the estimated residuals is assessed by the variography with the purpose of a correct estimation of the variance-covariance structure. The potential relationship between the spatial covariates and the functional response is then assessed by the Spatial Functional Freedman and Lane permutation scheme proposed here, based on permuting the estimated residuals, spatially filtered through the pre-multiplication by the inverse square root of the variance-covariance matrix, ensuring that the units to be permuted are asymptotically exchangeable and that the procedure is properly conducted, respecting the georeferenced nature of the data. The simulation studies intended to evaluate the performance of this inferential procedure in terms of the empirical size and the power show that the GLS test performs well whenever the range of the spatial correlation is not too high (more than half of the size of the spatial area). For the larger ranges, the procedure requires a larger sample size. This is likely to be due to the difficulty of estimating the spatial covariance structure when the range is large with respect to the study area D, which is well-known in the literature of the spatial statistics [19]. On the other hand, the ordinary least-squares approach shows very liberal behaviour whenever the data are correlated in space. With the increasing range of the spatial correlation, the test becomes even more liberal, its empirical size approaching to 1. Neglecting the underlying spatial structure of the data thus leads to an inference based on the non-exchangeable residuals and biases the results in terms of the empirical size of the test. The empirical power of the test was therefore evaluated for the GLS approach only. Three different models were considered: two-covariate, one-covariate and a misspecified one, where the data were generated as a model with one covariate, but a two-covariate model was assumed and fitted. The empirical

power is similar for all three models.

The potential of the methodology for heteroscedastic and spatially correlated functional data was shown on the real-world examples. Firstly, the introduced methodology for testing heteroscedastic data was used to study the soil samples collected in the growing seasons (March-October) of the years 2015 and 2016 at a site located near Křtiny, Czech Republic, containing 11 sampling points equally distributed on a straight line, perpendicular to the border between the field and the forest. To compare the geochemical properties of the field and the forest soil, a functional data on KCl pH, H_2O pH, and the percentage of organic carbon were considered. A balanced heteroscedastic, a balanced homoscedastic, and an unbalanced heteroscedastic model, respectively, were fit to the data. Testing for the significance of the regression functions enables to conclude that (i) KCl pH is affected by the type of the soil and the variability is different with respect to the type of the soil; (ii) values of H_2O pH depend on the type of soil, but the variability is not significantly different for the field and forest parts of the site; and (iii) the percentage of carbon is associated with a more complex spatial model, depending not only on the soil type but also on the distance from the ecotone, with increasing values when getting closer to the ecotone.

Secondly, the methodology introduced for the spatially correlated data was illustrated on the observations from the Venice province, containing the yearly per capita amounts of the produced municipal waste. The functions of the waste over the years were modelled using the scalar variable accounting for the tourism in each town, as well as for the spatial coordinates, expressed as the latitude and the longitude. As the residuals of the considered model were correlated in the space, we considered a generalized least-squares model and the proposed Spatial Functional Freedman and Lane permutation procedure to properly handle the data. Here, we found out that the effect of the tourism is significant in explaining the amount of the municipal waste, unlike the spatial covariates (longitude and latitude). In contrast, considering an ordinary least-squares approach, neglecting the spatial structure of the residuals leads to very liberal results failing to provide

an effective model selection for the production of the waste.

In the current work, we propose the OLS, WLS, and GLS classical test statistics integrated over the domain of interest, following the work of [1]. These test statistics are coherent with the classical tests of the hypotheses in the linear regression and they cover a wide variety of different tests (i.e., all tests of the linear hypotheses on the regression functions of the model), and in the absence of any information about the data-generative model, we suggest to use them. However, since the nonparametric inference based on the permutation tests is performed, this is not the only possible choice; an inference with the similar properties could be performed using the different point-wise statistics, the weighted integrals (if some information about the weighting is available), or the maximum values along the domain. Some test statistics could achieve a higher power under the specific alternatives. However, a thorough comparison between the different choices is out of the scope of this thesis, but it would be an interesting future line of research.

Bibliography

[1] K. Abramowicz, C. K. Häger, A. Pini, L. Schelin, S. Sjöstedt de Luna, and S. Vantini. Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics*, 45(4):1036–1061, 2018.

[2] C. Agostinelli and U. Lund. *R package* `circular`*: Circular Statistics (version 0.4-95).* CA: Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University, Venice, Italy. UL: Department of Statistics, California Polytechnic State University, San Luis Obispo, California, USA, 2022.

[3] J. Ambros, J. Altmann, C. Jurewicz, and A. Chevalier. Proactive assessment of road curve safety using floating car data: An exploratory study. *Archives of Transport*, 50(2):7–15, 2019.

[4] J. Ambros, J. Elgner, R. Turek, and V. Valentová. Where and when do drivers speed? A feasibility study of using probe vehicle data for speeding analysis. *Archives of transport*, 53, 2020.

[5] J. Ambros, V. Valentová, O. Gogolín, R. Andrášik, J. Kubeček, and M. Bíl. Improving the self-explaining performance of Czech national roads. *Transportation Research Record*, 2635:62–70, 2017.

[6] M. J. Anderson and J. Robinson. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2001.

[7] R. Andrášik and M. Bíl. Efficient road geometry identification from digital vector data. *Journal of Geographical Systems*, 18(3):249–264, 2016.

[8] G. Aneiros, R. Cao, R. Fraiman, C. Genest, and P. Vieu. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9, 2019.

[9] G. Aneiros, I. Horová, M. Hušková, and P. Vieu. On functional data analysis and related topics. *Journal of Multivariate Analysis*, 189:104861, 2022.

[10] J.-P. Aristizabal, R. Giraldo, and J. Mateu. Analysis of variance for spatially correlated functional data: application to brain data. *Spatial Statistics*, 32:100381, 2019.

[11] E. Arnone, L. Azzimonti, F. Nobile, and L. M. Sangalli. Modeling spatially dependent functional data via regression with differential regularization. *Journal of Multivariate Analysis*, 170:275–295, 2019.

[12] R. D. Baker. Two permutation tests of equality of variances. *Statistics and Computing*, 5(4):289–296, 1995.

[13] E. Belli. Smoothly adaptively centered ridge estimator. *Journal of Multivariate Analysis*, 189:104882, 2022.

[14] M. S. Bernardi, M. Carey, J. O. Ramsay, and L. M. Sangalli. Modeling spatial anisotropy via regression with partial differential regularization. *Journal of Multivariate Analysis*, 167:15–30, 2018.

[15] M. S. Bernardi, L. M. Sangalli, G. Mazza, and J. O. Ramsay. A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic environmental research and risk assessment*, 31(1):23–38, 2017.

[16] G. E. Box et al. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the

one-way classification. *The annals of mathematical statistics*, 25(2):290–302, 1954.

[17] J. A. Collazos, R. Dias, and A. Z. Zambom. Consistent variable selection for functional regression models. *Journal of Multivariate Analysis*, 146:63–71, 2016.

[18] I. D. Coope. Circle fitting by linear and nonlinear least squares. *Journal of Optimization Theory and Applications*, 76(2):381–388, 1993.

[19] N. A. Cressie. Statistics for spatial data. John Wiley and Sons. *Inc., New York*, 1993.

[20] C. de Boor. *A practical guide to splines, revised edition*. New York, NY: Springer-Verlag, 2001.

[21] P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):224–239, 2010.

[22] C. Dias, T. Oguchi, and K. Wimalasena. Drivers' speeding behavior on expressway curves: exploring the effect of curve radius and desired speed. *Transportation research record*, 2672:48–60, 2018.

[23] J. Diquigiovanni, M. Fontana, and S. Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

[24] M. A. Elliott, V. McColl, and J. V. Kennedy. Road design measures to reduce drivers' speed via 'psychological' processes: a literature review, 2003.

[25] B. Erbas, M. Akram, D. M. Gertig, D. English, J. L. Hopper, A. M. Kavanagh, and R. Hyndman. Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the United States and England–Wales. *Journal of epidemiology*, 20(2):159–165, 2010.

[26] B. Erbas, R. J. Hyndman, and D. M. Gertig. Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine*, 26(2):458–470, 2007.

[27] M. Escabias, A. Aguilera, and M. Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics: The official journal of the International Environmetrics Society*, 16(1):95–107, 2005.

[28] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[29] B. N. Fildes and S. J. Lee. The Speed Review: Road Environment, Behaviour, Speed Limits, Enforcement and Crashes. Report CR 127, 1993.

[30] M.-J. Fortin and G. M. Jacquez. Randomization tests and spatially autocorrelated data. *Bulletin of the Ecological Society of America*, 81(3):201–205, 2000.

[31] D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

[32] L. Garach, J. de Oña, and M. Pasadas. Mathematical formulation and preliminary testing of a spline approximation algorithm for the extraction of road alignments. *Automation in construction*, 47:1–9, 2014.

[33] R. Giraldo, W. Caballero, and J. Camacho-Tamayo. Mantel test for spatial functional data. *AStA Advances in Statistical Analysis*, 102(1):21–39, 2018.

[34] R. Giraldo, P. Delicado, and J. Mateu. Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, 18(3):411–426, 2011.

[35] A. Goia and P. Vieu. An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*, 146:1–6, 2016.

[36] P. Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

[37] O. Grujic and A. Menafoglio. *fdagstat, an R package*, 2017. R package version 1.0.

[38] P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002.

[39] L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer Science & Business Media, 2012.

[40] Y. Huang, H. Xu, V. Calian, and J. C. Hsu. To permute or not to permute. *Bioinformatics*, 22(18):2244–2248, 2006.

[41] R. Ignaccolo, J. Mateu, and R. Giraldo. Kriging with external drift for functional data for air quality monitoring. *Stochastic environmental research and risk assessment*, 28(5):1171–1186, 2014.

[42] H.-J. Kim, M. P. Fay, E. J. Feuer, and D. N. Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19(3):335–351, 2000.

[43] P. Kokoszka and M. Reimherr. *Introduction to functional data analysis*. Chapman and Hall/CRC, 2017.

[44] P. Kokoszka and M. Reimherr. Some recent developments in inference for geostatistical functional data. *Revista Colombiana de Estadística*, 42(1):101–122, 2019.

[45] E. D. Liapis, B. Psarianos, and E. Kasapi. Speed behavior analysis at curved ramp sections of minor interchanges. *Transportation Research Record*, 1751:35–43, 2001.

[46] E. Lila, J. A. Aston, L. M. Sangalli, et al. Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging. *The Annals of Applied Statistics*, 10(4):1854–1879, 2016.

[47] K.-O. Lindgren. Dyadic regression in the presence of heteroscedasticity - An assessment of alternative approaches. *Social networks*, 32(4):279–289, 2010.

[48] B. F. Manly. *Randomization, bootstrap and Monte Carlo methods in biology.* Chapman and Hall/CRC, 2018.

[49] M. Marozzi. Some notes on nonparametric inferences and permutation tests. *Metron International Journal of Statistics*, 60(3-4):139–151, 2002.

[50] A. Menafoglio, O. Grujic, and J. Caers. Universal Kriging of functional data: Trace-variography vs cross-variography? Application to gas forecasting in unconventional shales. *Spatial Statistics*, 15:39–55, 2016.

[51] A. Menafoglio and P. Secchi. Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European journal of operational research*, 258(2):401–410, 2017.

[52] A. Menafoglio, P. Secchi, and M. Dalla Rosa. A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics*, 7:2209–2240, 2013.

[53] K. Moder. Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling*, 52(4):343–353, 2010.

[54] H. Oja. On permutation tests in multiple regression and analysis of covariance problems. *Australian Journal of Statistics*, 29(1):91–100, 1987.

[55] K. A. Passetti and D. B. Fambro. Operating speeds on curves with and without spiral transitions. *Transportation research record*, 1658(1):9–16, 1999.

[56] E. Pebesma. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446, 2018.

[57] A. M. Pérez-Zuriaga, F. J. Camacho-Torregrosa, and A. García. Tangent-to-curve transition on two-lane rural roads based on continuous speed profiles. *Journal of transportation engineering*, 139(11):1048–1057, 2013.

[58] F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, 2010.

[59] A. Pini and S. Vantini. *fdatest: Interval Wise Testing for Functional Data*, 2017. R package version 2.1.0.

[60] A. Pini and S. Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2):407–424, 2017.

[61] J. Ramsay, G. Hooker, and S. Graves. *Functional data analysis with R and MATLAB*. Springer-Verlag New York, 2009.

[62] J. Ramsay and B. W. Silverman. *Functional data analysis*. Springer New York, 2005.

[63] J. O. Ramsay, H. Wickham, S. Graves, and G. Hooker. *fda: Functional Data Analysis*, 2018. R package version 2.4.8/r771.

[64] V. Římalová, J. Elgner, J. Ambros, and E. Fišerová. Modelling the driving speed on expressway ramps based on floating car data. *Measurement*, 195:110995, 2022.

[65] V. Římalová, E. Fišerová, A. Menafoglio, and A. Pini. Inference for spatial regression models with functional response using a permutational approach. *Journal of Multivariate Analysis*, 189:104893, 2022.

[66] V. Římalová, A. Menafoglio, A. Pini, V. Pechanec, and E. Fišerová. A permutation approach to the analysis of spatiotemporal geochemical data in the presence of heteroscedasticity. *Environmetrics*, 31(4):e2611, 2020.

[67] E. Romano, J. Mateu, and R. Giraldo. On the performance of two clustering methods for spatial functional data. *AStA Advances in Statistical Analysis*, 99(4):467–492, 2015.

[68] E. Romano, R. Verde, and V. Cozza. Clustering Spatial Functional Data: A Method Based on a Nonparametric Variogram Estimation. In *New Perspectives in Statistical Modeling and Data Analysis*, pages 339–346. Springer, 2011.

[69] K. Sulovská, E. Fišerová, M. Chvosteková, and M. Adámek. Appropriateness of gait analysis for biometrics: Initial study using FDA method. *Measurement*, 105:1–10, 2017.

[70] SWOV Institute for Road Safety Research. Speed choice: the influence of man, vehicle, and road. https://www.swov.nl/sites/default/files/publicaties/gearchiveerde-factsheet/uk/fs_speed_choice_archived.pdf, 2012. [Online; accessed 23-January-2022].

[71] C. J. ter Braak. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and related techniques*, pages 79–85. Springer, 1992.

[72] TRB. Modeling operating speed: Synthesis report. Transportation Research Board (TRB), Washington, DC, 2011.

[73] ÚNMZ. ČSN 73 6101 Projektování silnic a dálnic. Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Prague., 2018.

[74] A. van Beinum, H. Farah, F. Wegman, and S. Hoogendoorn. Driving behaviour at motorway ramps and weaving segments based on empirical trajectory data. *Transportation Research Part C: Emerging Technologies*, 92:426–441, 2018.

[75] A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.

[76] J. Zbíral and I. Honsa. *Unified working procedures. Soil analysis I.* Central Institute for Supervising and Testing in Agriculture, 2010.

[77] J. Zbíral, P. Tieffová, Š. Plhalová, E. Urbánková, E. Niedobová, J. Srnková, and I. Střížová. *Unified working procedures. Soil analysis II.* Central Institute for Supervising and Testing in Agriculture, 2011.

PALACKÝ UNIVERSITY IN OLOMOUC
FACULTY OF SCIENCE

DISSERTATION THESIS SUMMARY

Functional regression models

Supervisor: **doc. RNDr. Eva Fišerová, Ph.D.**
Author: **Mgr. Veronika Římalová**
Study program: P1104 Applied Mathematics
Field of study: Applied Mathematics
Form of study: Full-time
The year of submission: 2023

The dissertation thesis was carried out under the full-time postgradual programme Applied Mathematics, field Applied Mathematics, in the Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc.

Applicant: **Mgr. Veronika Římalová**
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Supervisor: **doc. RNDr. Eva Fišerová, Ph.D.**
Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science
Palacký University Olomouc

Reviewers: **doc. RNDr. Zdeněk Hlávka, Ph.D.**
Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics
Charles University in Prague

doc. Mgr. Jan Koláček, Ph.D.
Department of Mathematics and Statistics
Faculty of Science
Masaryk University Brno

Dissertation thesis summary was sent to distribution on .....

Oral defence of dissertation thesis will be performed on June 14 2023 at Department of Mathematical Analysis and Applications of Mathematics in front of the committee for Ph.D. study programme Applied Mathematics, Faculty of Science, Palacký University Olomouc, room ...., 17. listopadu 12, Olomouc.

Full text of the dissertation thesis is available at Study Department of Faculty of Science, Palacký University Olomouc.

Contents

1. Abstract

This thesis proposes a novel nonparametric approach to the significance testing of the null hypothesis in a functional linear model for data with a heterogeneous spatial structure. A permutation approach is introduced to test for the effect of covariates in a spatial functional regression model with heteroscedastic or spatially correlated residuals. The proposed methods account for the heterogeneous spatial structure of the data by grounding on the Freedman and Lane permutation scheme for the estimated residuals of the functional regression model. Indeed, due to the spatial dependence among the data, the residuals of the regression model are not exchangeable, and the basic assumption of the Freedman and Lane permutation scheme is violated. To overcome this issue in the case of heteroscedasticity, we propose to fit a weighted least squares model to the observations, and then to divide the estimated residuals by their corresponding standard deviation, leading to asymptotically exchangeable residuals. In the case of spatially correlated observations it is proposed here to estimate the variance-covariance structure of the residuals by variography, remove this correlation by spatial filtering of the residuals and base the permutation test on these approximately exchangeable residuals. To evaluate the performance of the proposed methods in terms of the empirical size and power, simulation studies are conducted, examining the behaviour of the tests under different covariance settings. We show that neglecting the spatial structure of the residuals in the permutation scheme, yields very liberal or conservative results, whereas the proposed procedures are close to the nominal size of the test. The results of modelling and testing on the case studies are shown and discussed on the data from geochemistry and tourism. Moreover, the potential of FDA methodology is shown in the field of transportation research, focusing on the permutation-based inference in a regression model with functional covariates and functional response.

Key words: Functional data analysis, functional regression models, permutation tests, nonparametric inference, heteroscedasticity, spatial correlation, functional geostatistics, spatially filtered residuals, exchangeability

2. Abstrakt v českém jazyce

Cílem této disertační práce je představit nový neparametrický přístup k testování významnosti parametrů ve funkcionálním lineárním regresním modelu pro prostorová data pro případ heteroskedasticity nebo prostorové korelace. Pro testování významnosti regresních parametrů v prostorovém funkcionálním regresním modelu navrhujeme využít metodiku založenou na permutačních testech. Tyto metody jsou navrhnuty tak, aby bylo možno s jejich pomocí pracovat s daty s heterogenní prostorovou strukturou a to na základě Freedmanova a Laneova permutačního schématu. Přirozeně, vlivem prostorové závislosti mezi daty nejsou residua regresního modelu permutovatelná, čímž je porušen základní předpoklad Freedmanova a Laneova permutačního schématu. Abychom tento problém vyřešili, navrhujeme v případě heteroskedasticity modelovat data pomocí vážené metody nejmenších čtverců a poté vydělit odhadnutá residua jejich směrodatnou odchylkou, čímž získáme asymptoticky permutovatelná residua. V případě prostorově závislých dat navrhujeme odhadnout varianční strukturu pomocí variogramu a poté dekorelovat odhadnutá residua a založit permutační test na těchto přibližně permutovatelných residuích. Za účelem ohodnocení navrhovaných testů z hlediska empirické velikosti a síly testu byly provedeny simulační studie zkoumající chování testů pro různé varianční struktury dat. Ukážeme, že zanedbání varianční struktury residuí (tedy permutování přímo heteroskedastických či prostorově korelovaných dat) vede v permutačním testování k buďto velmi liberálním, či velmi konzervativním výsledkům testů, zatímco empirická velikost námi navrhovaných testů je v případě heteroskedasticity či prostorové korelace blízká té nominální. Navrhované metody jsou též prezentovány na reálných datech z oblastí geochemie a turismu. Mimo to je potenciál analýzy funkcionálních dat ukázán na datech z oblasti dopravního výzkumu s důrazem na regresní model s funkcionálním regresorem a funkcionální závisle proměnnou.

Klíčová slova: Analýza funkcionálních dat, funkcionální regresní modely, permutační testy, neparametrická inference, heteroskedasticita, prostorová korelace, funkcionální geostatistika, prostorově filtrovaná residua, permutovatelnost

3. Introduction

In many practical tasks involving high-dimensional measurements, such as curves or surfaces, it is convenient to treat the data not as a sequence of single measurements taken one after another, but as the whole entities. Such approach has developed into functional data analysis (FDA), where the basic elements of the statistical analysis are the functional observations as a whole [19, 41]. In these days, functional data occur in many scientific fields, such as an analysis of human gait by means of linear models [1], and analysis of variance [47], modelling the time trends and forecasting the future rates of age-specific breast cancer mortality using time series models [16, 17], or predicting the risk of drought by principal component logistic regression [18]. The overview of the recent trends in FDA can be found in summary surveys [3, 4, 23].

Most FDA methods strongly rely on the assumption of independence among the observations. Under the presence of spatial dependence, applying these methods is inappropriate, and the analysis could fail because of the consistency problems [25]. Spatial dependence needs to be properly treated; for this purpose, the classical geostatistical methods can be extended to the functional framework [9, 34, 35].

In the geostatistical framework, the functional observations can occur, e.g., as soil or weather measurements taken from the same site with possibly daily, weekly or monthly frequency (see, e.g., [26, 41, 45]). In general, the extension of the geostatistical methods to the functional setting is well-developed these days, including ordinary or universal kriging [22, 33, 34, 35], or other aspects related to spatial statistics, such as the change point analysis [29], clustering [46], tests for detecting the spatial autocorrelation [21], or the regression with differential regularization to model dependent functional data in space [6].

In functional data analysis, regression models play an important part of the methods used. A thorough overview of the methodology with application can be found in vast amount of the literature, see, e.g., [40, 41], focusing on situations where at least the covariates, or the response, or both is functional. The cases

of the concurrent model, as well as the functional-on-functional model, where the functional domains of the covariate and the response differ, i.e., $\beta(t_1, t_2)$ is a function of these two domains, are detailed here. In the work of [25], the methodology for models for independent, dependent, and spatially correlated data is introduced. In the work of [19], the focus is on the analysis of nonparametric functional data, including the models for such data. In the work of [28], functional regression models, including the generalized models and models for spatially correlated data, are concerned. An overview of recent trends in FDA can be found in summary surveys [3, 4, 23], including also topics related to functional regression models, see, e.g., the work on variables selection, such as [11], or [8], where the variables selection is based on smoothed centred ridge regression approach. In the work of [15], a nonparametric approach to the construction of prediction bands in the multi-functional regression framework is concerned.

Focusing on the inference in the context of functional data, testing for the significance can be performed by means of either parametric or nonparametric methods. The nonparametric approach appears particularly promising, as it allows for minimal assumptions on the data-generative model. In this context, permutation tests have been recently successfully developed, leading to flexible approaches for testing in FDA [1, 39]. Indeed, the permutation tests are only based on the assumption of exchangeability between the units under the null hypothesis [27, 32], meaning that, under H_0, the data distribution does not change if the units are randomly resampled [24, 38]. In practice, this is evaluated by comparing the test statistic evaluated on the original data to the distribution of the test statistics obtained from the permuted data. If the observed statistic lies in the tail of this distribution, the null hypothesis can be rejected [7, 40].

In the framework of linear models, it is not possible to directly observe the exchangeable quantities under the null hypothesis unless the model only contains a single covariate. The permutation scheme can be adapted in different ways to obtain the approximate exchangeability. In the work of [20], permuting the estimated residuals from the model under the null hypothesis (also referred to as

the reduced model) was proposed; in the work of [48], permuting the residuals of the full model was proposed; other authors, see [31, 37], proposed to permute other quantities, such as the covariates' values, or the responses. In the work of [2], Freedman and Lane's methodology (see [20]) and other approaches were compared by a simulation study, concluding that the permutation scheme proposed by Freedman and Lane [20], relying on the permutation of the estimated residuals from the reduced model, gives the best empirical results in terms of the power and the size of the test. From a theoretical point of view, since the estimated residuals of the null model are asymptotically exchangeable, the obtained test is asymptotically exact. In the work of [49], this approach was applied to the multivariate case of testing for the differences between two sets of images; in the work of [1], Freedman and Lane permutation scheme was adapted to testing in a functional regression model for the knee movement.

4. Recent state summary

4.1. Overview of the functional data

In the framework of functional data analysis, we assume that the observed discrete data are generated by some unknown function, i.e., [41]

$$x_g = y(t_g) + \xi_g, g \in \{1, \ldots, G\}, t \in T \subset [0, \infty), \tag{1}$$

where G is the number of discrete observations of the function $y(t)$ and ξ_g is the random error. Moreover, it is assumed that the underlying function is smooth, therefore the two adjacent discrete data points are likely to be similar to each other.

Assume now that $i \in \{1, \ldots, n\}$ functions were observed. Then, the i-th function $y_i(t)$ can be represented by a linear combination of the known basis functions and the unknown basis coefficients as [41]

$$y_i(t) = \sum_{k=1}^{K_{y,i}} c_{ki}\phi_{ki}(t) = \boldsymbol{\phi}_i'(t)\mathbf{c}_i, i \in \{1, \ldots, n\}, \tag{2}$$

where $K_{y,i}$ is the number of the basis coefficients of the i-th functional observation, $\phi_{ki}(t)$ are known basis functions and c_{ki} are unknown basis coefficients, which can be estimated, e.g., by the least-squares method. The symbol $'$ denotes the transposition.

A better control over the smoothness can be provided by the roughness penalties, where the roughness of a function can be quantified as its curvature, i.e., the square of the second derivative: $[D^2 y_i(t)]^2$. Then, the roughness can be measured by an integrated squared second derivative [41]

$$\mathrm{PEN}_2(y_i) = \int_T [D^2 y_i(t)]^2 dt. \tag{3}$$

Note that, in general, one can assume the ν-th power of the ν-th derivative, that is, $[D^\nu y_i(t)]^\nu, \nu \in \{1, 2, 3, \ldots\}$. Following Equation (2), Equation (3) can be rewritten as $\mathbf{c}_i'\mathbf{R}_{y,i}\mathbf{c}_i$, where $\mathbf{R}_{y,i}$ represents the roughness penalty matrix for the

i-th function $y_i(t)$. The (j,k)-th entry of $\mathbf{R}_{y,i}$ is given as $\mathbf{R}_{y,i_{(j,k)}} = \int_T D^2\phi_{ji}(t)D^2\phi_{ki}(t)dt$. An optimal fit to the observed discrete data is then found by the minimization of the penalized residual sum of squares criterion [41]

$$\text{PENSSE}_{\lambda_i}(y_i|\mathbf{x_i}) = [\mathbf{x_i} - y_i(\mathbf{t})]'[\mathbf{x_i} - y_i(\mathbf{t})] + \lambda_{y,i}\mathbf{c}_i'\mathbf{R}_{y,i}\mathbf{c}_i, \tag{4}$$

where $\mathbf{x_i}$ is a set of discrete observations of $y_i(t)$ and $\lambda_{y,i} \geq 0$ is a smoothing parameter controlling the smoothness of $y_i(t)$, which can be selected, e.g., by the generalized cross-validation. Note that, in general, $\lambda_{y,i}$ can vary from function to function.

To take into account the whole data set, define the diagonal basis functions matrix $\mathbf{\Phi} = \text{diag}(\boldsymbol{\phi}_1'(t), \ldots, \boldsymbol{\phi}_n'(t))$, the set of all discrete observations $\mathbf{x} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$, and a symmetric block-diagonal matrix $\mathbf{R}_y = \text{diag}(\lambda_{y,1}\mathbf{R}_{y,1}, \ldots, \lambda_{y,n}\mathbf{R}_{y,n})$. Then, the vector of the basis coefficients $\mathbf{c} = (\mathbf{c}_1', \ldots, \mathbf{c}_n')'$ can be estimated as [40]

$$\widehat{\mathbf{c}} = (\mathbf{\Phi}'\mathbf{\Phi} + \mathbf{R}_y)^{-1}\mathbf{\Phi}'\mathbf{x}. \tag{5}$$

Define further the total number of the basis coefficients $K_y = \sum_{i=1}^n K_{y,i}$, the K_y-dimensional column vector $\boldsymbol{\phi} = (\boldsymbol{\phi}_1'(t), \ldots, \boldsymbol{\phi}_n'(t))'$ and the $n \times K_y$ matrix of the basis coefficients $\widehat{\mathbf{C}} = \text{diag}(\widehat{\mathbf{c}}_1', \ldots, \widehat{\mathbf{c}}_n')$. Then, one can write $\widehat{\mathbf{y}} = \widehat{\mathbf{C}}\boldsymbol{\phi}$.

4.2. Functional-on-functional model specification

Assume that a sample of functions $y_i(t), i \in \{1, \ldots, n\}$, has been observed. In the following, $y_i(\cdot)$ represents an element of the Hilbert space of the square-integrable functions $L^2(T)$. We further assume that the functional observations can be represented through the linear model [1]

$$\mathbf{y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\mathbf{y} = (y_1(t), \ldots, y_n(t))', t \in T$, are the functional observations, $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_L(t))', t \in T$, are the unknown regression functions, and

$\boldsymbol{\varepsilon} = (\varepsilon_1(t), \ldots, \varepsilon_n(t))', t \in T$, are the random error functions. A known $n \times (L+1)$ functional design matrix $\mathbf{F}$ can be symbolically expressed through a common notation with the i-th row $(1, f_{i1}(t), \ldots, f_{iL}(t)), t \in T$, where $f_{i\ell}(t), \ell \in \{1, \ldots, L\}$, are the known covariates related to the i-th observation. As the covariates can in general be functions, the element $f_{i\ell}(t)$ of the design matrix would in that case represent a functional covariate.

4.3. Estimation of the regression functions

Express now the unknown regression functions through the basis expansion [40, 41]

$$\beta_\ell(t) = \sum_{k=1}^{K_{\beta,\ell}} b_{k\ell}\theta_{k\ell}(t) = \boldsymbol{\theta}'_\ell(t)\mathbf{b}_\ell, \ell \in \{0, 1, \ldots, L\}, \tag{7}$$

where $K_{\beta,\ell}$ is the number of the basis coefficients of the ℓ-th regression function, $\theta_{k\ell}(t)$ are the known basis functions and $b_{k\ell}$ are the unknown coefficients. In this way, the problem of estimating the functions $\beta_\ell(t)$ reduces to estimating the coefficients $b_{k\ell}$ by, e.g., the least-squares method.

In order to express the model (6) with respect to the basis expansion (7), define the total number of the basis coefficients of the regression functions, i.e., $K_\beta = \sum_{\ell=0}^{L} K_{\beta,\ell}$, and stack the vectors $\mathbf{b}_\ell$ vertically to obtain a K_β-dimensional column vector $\mathbf{b} = (\mathbf{b}'_0, \mathbf{b}'_1, \ldots, \mathbf{b}'_L)'$. By defining an $(L+1) \times K_\beta$ matrix $\boldsymbol{\Theta} = \text{diag}(\boldsymbol{\theta}'_1(t), \ldots, \boldsymbol{\theta}'_L(t))$ and an $n \times K_\beta$ matrix $\tilde{\mathbf{F}} = \mathbf{F}\boldsymbol{\Theta}$, model (6) can be formally transformed to a *constant coefficient linear model* as [41]

$$\mathbf{y} = \tilde{\mathbf{F}}\mathbf{b} + \boldsymbol{\varepsilon}. \tag{8}$$

Following the general assumption that the roughness penalty term $\text{PEN}_2(\beta_\ell) = \int_T [D^2\beta_\ell(t)]^2 dt$ can be different for each regression function $\beta_\ell, \ell \in \{0, 1, \ldots, L\}$, one can, following Equation (7), write $\text{PEN}_2(\beta_\ell) = \mathbf{b}'_\ell \mathbf{R}_{\beta,\ell}\mathbf{b}_\ell$, where $\mathbf{R}_{\beta,\ell} = \int_T D^2\boldsymbol{\theta}_\ell(t) D^2\boldsymbol{\theta}'_\ell(t) dt$ represents the roughness penalty matrix. Then, a symmetric block-diagonal $K_\beta \times K_\beta$ matrix

$\mathbf{R}_\beta = \text{diag}(\lambda_{\beta,0}\mathbf{R}_0, \ldots, \lambda_{\beta,L}\mathbf{R}_L)$, where $\lambda_{\beta,\ell}, \ell \in \{0, 1, \ldots, L\}$, are the smoothing parameters, can be created. The penalized least-squares estimator of the regression coefficients $\mathbf{b}$ is then defined as [41]

$$\widehat{\mathbf{b}} = \left[\int_T \tilde{\mathbf{F}}'(t)\tilde{\mathbf{F}}(t)dt + \mathbf{R}_\beta \right]^{-1} \int_T \tilde{\mathbf{F}}(t)'\mathbf{C}\boldsymbol{\phi}(t)dt. \tag{9}$$

Then, $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_L(t))', t \in T$, can be estimated as $\widehat{\boldsymbol{\beta}} = \boldsymbol{\Theta}\widehat{\mathbf{b}}$.

4.4. Spatial functional model specification

Let the spatial domain of interest be denoted as $D \subset \mathbb{R}^d$. Let s be a point in D and denote by $X_s(t)$ an observation at location $s \in D$ at time $t \in T = [t_1, t_2]$. From this point, we assume that for the fixed location s, the random function $\mathcal{X}_s(\cdot)$ is an element of the Hilbert space $L^2(T)$ of the square-integrable functions. The set

$$\{\mathcal{X}_s(t), t \in T, s \in D \subset \mathbb{R}^d\} \tag{10}$$

is called the functional random field of the time-varying entity, defined on L^2.

In the following, we assume that the functional random field (10) is second-order stationary and isotropic in the sense of the work of [35], i.e.,

1. Second-order stationarity
   - $\text{E}(\mathcal{X}_s(t)) = \mu_s(t), \forall s \in D$,
   - $\text{Cov}(\mathcal{X}_{s_i}, \mathcal{X}_{s_j}) = \text{E}(\langle \mathcal{X}_{s_i} - m_{s_i}, \mathcal{X}_{s_j} - m_{s_j} \rangle) = C(h), \forall s_i, s_j \in D$, $h = s_i - s_j, h \in D$, and s_i, s_j, h are d-dimensional vectors,
2. Isotropy
   - $\text{Cov}(\mathcal{X}_{s_i}, \mathcal{X}_{s_j}) = \text{E}(\langle \mathcal{X}_{s_i} - \mu_{s_i}, \mathcal{X}_{s_j} - \mu_{s_j} \rangle) = C(\|h\|_D), \forall s_i, s_j \in D$, $h = s_i - s_j, h \in D$, and s_i, s_j, h are d-dimensional vectors,

where $\|\cdot\|_D$ is a norm over D and $\langle \cdot, \cdot \rangle$ denotes the inner product in L^2. This means that, only the mean $\text{E}(\mathcal{X}_s(t))$ is a function of the location s, whereas the

global covariances $\mathrm{Cov}(\mathcal{X}_{s_i}, \mathcal{X}_{s_j})$ do not depend on the locations s_i, s_j, but only on $\|h\|_D$, the distance between these locations [13, 25, 35].

Following the work of [35], we assume that the random field can be modelled as

$$\mathcal{X}_s(t) = m_s(t) + \delta_s(t), s \in D, t \in T, \tag{11}$$

where $m_s(t)$ is the drift and $\delta_s(t)$ are the residuals, the realizations of a zero-mean, second-order stationary and isotropic random process. The drift captures a non-constant mean variation in space D and can be represented by a linear model

$$m_s(t) = \beta_0(t) + \sum_{\ell=1}^{L} \beta_\ell(t) f_\ell(s), s \in D, t \in T, \tag{12}$$

where $\beta_\ell(t), \ell \in \{0, \ldots, L\}$, are the unknown regression functions, independent on the location in space D, and $f_\ell(s)$, $\ell \in \{1, \ldots, L\}, s \in D$, are the known functions of the spatial variable $s \in D$.

Model (11) can be rewritten into the matrix form as

$$\boldsymbol{\mathcal{X}} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\delta}, \tag{13}$$

where $\boldsymbol{\mathcal{X}} = (\mathcal{X}_{s_1}(t), \ldots, \mathcal{X}_{s_n}(t))', t \in T$, is a matrix of the functional observations at the spatial points $s_1, \ldots, s_n \in D, \mathbf{F}_i = (1, f_1(s_i), \ldots, f_L(s_i)), s_i \in D$, $i \in \{1, \ldots, n\}$, is an i-th row of a known $n \times (L+1)$ design matrix $\mathbf{F}$, $\boldsymbol{\beta} = (\beta_0(t), \ldots, \beta_L(t))', t \in T$, are the unknown regression functions, and $\boldsymbol{\delta} = (\delta_{s_1}(t), \ldots, \delta_{s_n}(t))', s_1, \ldots, s_n \in D, t \in T$, are the spatially correlated residuals with an unknown variance-covariance structure represented by the matrix $\boldsymbol{\Sigma}$. Note that, $\boldsymbol{\Sigma}$ is constant over T.

4.5. Analysis of the spatial correlation

From here, to avoid a heavy notation, we directly set $h = \|s_i - s_j\|_D$. A covariance function of a second-order stationary and isotropic random field (10) can be typically parametrized as

$$C(h) = \sigma^2 \rho(h), h \geq 0, \rho(0) = 1, \tag{14}$$

where $\rho(h)$ is a correlation function measuring the linear relationship between the observations at the distance h. Under these assumptions, the (i,j)-th entry of the variance-covariance matrix $\mathbf{\Sigma}$ is given as

$$\Sigma_{ij} = \mathrm{Cov}(\delta_{s_i}, \delta_{s_j}) = C(h), s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}. \quad (15)$$

The spatial dependence among the residuals can be also specified by means of the (functional) semivariogram

$$\gamma(h) = \frac{1}{2}\,\mathrm{Var}\left(\delta_{s_i} - \delta_{s_j}\right) = \frac{1}{2}\,\mathrm{E}[\|\delta_{s_i} - \delta_{s_j}\|^2], s_i, s_j \in D, h = \|s_i - s_j\|_D\,, \quad (16)$$

where $\|\cdot\|$ is the L^2-norm. Note that, one can define the variogram as $2\gamma(h)$. The relationship between the semivariogram (16) and the covariance function (14) can be expressed as

$$\gamma(h) = C(0) - C(h), s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}. \quad (17)$$

The functional semivariogram is estimated by the empirical functional semivariogram [34, 35]

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} \|\delta_{s_i} - \delta_{s_j}\|^2, s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}, \quad (18)$$

where $N(h)$ is the set of all pairs of the observations at a distance approximately h and $|N(h)|$ is the cardinality of this set.

Similarly to the scalar case [12], the shape of the trace-variogram can be used to determine if $\mathbf{\Sigma}$ is diagonal. Indeed, the uncorrelated residuals δ_{s_i} are associated with a pure nugget model for a variogram, that is, a constant trace-variogram function.

4.6. Estimation of the regression functions in the spatial functional model

Let the residuals be either homoscedastic, heteroscedastic, or spatially correlated, and let their spatial variance-covariance matrix be invertible.

When the conditions $\text{rank}(\mathbf{F}) = L + 1 \leq n$ and $\text{rank}(\boldsymbol{\Sigma}) = n$ hold, the regression functions $\boldsymbol{\beta}$ in the model (13) can be estimated as follows.

For the case of homoscedastic and uncorrelated residuals, the variance-covariance matrix is of the form $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is an identity matrix, and the estimator of the regression functions $\boldsymbol{\beta}$ simplifies to the ordinary least-squares (OLS) one [35] as

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\mathcal{X}}. \tag{19}$$

The OLS estimator is the best linear unbiased estimator and is associated with the variance-covariance matrix $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{OLS}}) = \sigma^2(\mathbf{F}'\mathbf{F})^{-1}$. The parameter σ^2 is usually unknown and is simply estimated from the estimated residuals as

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \|\widehat{\delta}_i(t)\|^2, \tag{20}$$

where $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{\text{OLS}}$ is a vector of the estimated functional residuals.

For the case of heteroscedastic and uncorrelated residuals, the variance-covariance matrix $\boldsymbol{\Sigma}$ is a diagonal matrix. Let $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$ be a diagonal matrix of the weights. In general, the weight $w_{ii}, i \in \{1, \ldots, n\}$, can differ for each observation. However, in the following, we shall assume that the weights are constant within the groups and possibly different between the groups in the data. The regression functions $\boldsymbol{\beta}$ can be estimated through the weighted least-squares (WLS) estimator as

$$\widehat{\boldsymbol{\beta}}_{\text{WLS}} = (\mathbf{F}'\mathbf{W}\mathbf{F})^{-1}\mathbf{F}'\mathbf{W}\boldsymbol{\mathcal{X}}. \tag{21}$$

Assuming $\mathbf{W}$ is a diagonal matrix, the WLS estimator is the best linear unbiased estimator and is associated with the variance-covariance matrix $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{WLS}}) = (\mathbf{F}'\mathbf{W}\mathbf{F})^{-1}$.

For the case of spatially correlated residuals, a general variance-covariance matrix $\boldsymbol{\Sigma}$ is associated with the residuals, and the regression functions $\boldsymbol{\beta}$ can be estimated through the generalized least-squares (GLS)

estimator [35] as

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{X}}, \tag{22}$$

The GLS estimator is the best linear unbiased estimator and is associated with the variance-covariance matrix $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{GLS}}) = (\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F})^{-1}$.

In practice, the variance-covariance matrix $\boldsymbol{\Sigma}$ describing the spatial dependence is usually unknown, so that the GLS estimator is not available. The general variance-covariance matrix $\boldsymbol{\Sigma}$ can be estimated, e.g., from the estimated residuals by the semivariogram through an iterative two-step procedure for the GLS estimates of the regression functions. Particularly, with respect to Equations (14) and (17), the (i, j)-th entry of $\boldsymbol{\Sigma}$ is estimated as

$$\widehat{\boldsymbol{\Sigma}}_{i,j} = \widehat{\sigma}^2 - \widehat{\gamma}(h), s_i, s_j \in D, h = \|s_i - s_j\|_D, i, j \in \{1, \ldots, n\}, \tag{23}$$

where $\widehat{\sigma}^2$ and $\widehat{\gamma}(h)$ are given by Equations (20) and (18), respectively.

In the case of the diagonal variance-covariance matrix $\boldsymbol{\Sigma} = \mathbf{W}^{-1}$, under the presence of groups in the data where the variances are assumed to be constant within each group and possibly different between the groups, the estimation of the diagonal matrix $\mathbf{W}$ is as follows. Assume $\delta_{s_i(m)}(t), s_i \in D, i \in \{1, \ldots, n_m\}$, $m \in \{A, B\}$, are the two groups of the functional residuals from the model specified in Equations (11) and (12), where i is the unit index and m is the population index. Let $\sigma^2_{(A)}$ and $\sigma^2_{(B)}$ be the global variances (constant over T) of these two populations. The variances are estimated separately for each group by Equation (20), and their estimates $\widehat{\sigma}^2_{(A)}$ and $\widehat{\sigma}^2_{(B)}$ are indeed the estimates of the reciprocal diagonal elements $w_{ii(A)}$ and $w_{ii(B)}$ of the matrix of weights $\mathbf{W}$, i.e., $\widehat{w}_{ii(m)} = 1/\widehat{\sigma}^2_{(m)}, i \in \{1, \ldots, n\}, m \in \{A, B\}$. The rest of the procedure is analogous to the GLS case.

The two-step iterative procedure for the estimation of $\boldsymbol{\beta}$ by the GLS or the WLS is as follows. Firstly, the estimated residuals $\widehat{\boldsymbol{\delta}}$, are obtained by the OLS (Equation (19)). Then, the variance-covariance matrix $\boldsymbol{\Sigma}$ (or $\mathbf{W}$) is estimated from the estimated residuals $\widehat{\boldsymbol{\delta}}$ and the regression functions $\boldsymbol{\beta}$ are estimated by the GLS (or the WLS) by plugging in $\widehat{\boldsymbol{\Sigma}}$ (or $\widehat{\mathbf{W}}$) into Equation (22) (or Equation

(21)). As the next step, the estimated residuals $\widehat{\boldsymbol{\delta}}$ are obtained as $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ (or as $\widehat{\boldsymbol{\delta}} = \boldsymbol{\mathcal{X}} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$). In both cases, this estimation procedure of $\boldsymbol{\beta}$ should be repeated few times until the convergence is reached [35]. The iterative procedure is stopped when the change of the estimates of the residuals in two consecutive steps is sufficiently small, i.e., if

$$\|\widehat{\delta}_{s_i,l} - \widehat{\delta}_{s_i,l-1}\| < \text{given value}, s_i, s_j \in D, i \in \{1, \ldots, n\}. \tag{24}$$

Here, $\widehat{\delta}_{s_i,l}$ denotes the estimate of the residual δ_{s_i} in the l-th step.

5. Thesis objectives

The aim of this thesis is to introduce a novel approach to the inference in functional regression models, aiming also at more complex data structures. In particular, a testing procedure for a regression model with functional covariate and functional response, based on a Freedman and Lane permutation scheme is outlined, being a novel approach in the field of transportation research, grounding on a real-world task from this research area.

Extending this methodology from homoscedastic [1, 42] to the space-time setting, both the observed data and the residuals are likely to be heterogeneous by virtue of their geographical closeness and thus treating them as exchangeable may bias the results. Firstly, a case of heteroscedasticity among the spatial observations is concerned [44]. Under the presence of more groups in the data, an interesting research question is whether these groups have different properties. Assuming the same variance within the groups and the different variance between the groups, a permutation testing procedure based on a spatial regression model is here proposed. To deal with the nature of spatial observations, which depend on their spatial coordinates, we propose to permute the residuals of the spatial regression model and not the observations themselves. For this purpose, the permutation scheme by Freedman and Lane (see [20]) is adapted here. An emphasis shall be given to the exchangeability of the residuals, since this assumption can be easily violated whenever the data are heteroscedastic.

Secondly, a case of spatially correlated observations is concerned [43]. To ensure the approximate exchangeability of the units being permuted, a spatial filtering of the residuals of the functional linear model is proposed. More precisely, the spatial covariance of the residuals is estimated through the functional variography [25] and then the residuals are de-correlated by using the inverse-square root of the estimated variance-covariance matrix, eventually obtaining approximately exchangeable residuals, suitable for using the Freedman and Lane permutation scheme. A similar approach can be found in [5], where the spatial covariance is estimated through the trace variogram and used to fit the functional

analysis of variance model with GLS. Then, the inference is based on an approximate chi-squared distribution of the test statistic, that is derived based on the assumption of normality of the residuals, whereas a nonparametric permutational approach is proposed in this work.

To summarize the previous paragraphs, the aims of this thesis are to

1. model functional data and to set up the methodology for spatiotemporal observations, including the estimation of the spatial relationship among the observations through variography,

2. introduce a novel approach to the inference in spatial functional regression models based on a permutation scheme by Freedman and Lane [20], aiming at the inference in case of (i) heteroscedastic and (ii) spatially correlated functional observations,

3. apply the proposed methodology to real-world spatiotemporal data,

4. show a potential of FDA framework in the field of transportation research, focusing on the inference in functional-on-functional regression models.

6. Theoretical framework and applied methods

6.1. The global permutation tests for the effect of the covariates in the spatial functional regression model

This section focuses on the problem of testing for the significance of the regression functions in model (12). Recall a compact matrix form of the linear hypotheses about the effects of covariates as

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}, \text{ against } H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}, \tag{25}$$

where $\mathbf{C}$ is an $h \times (L+1)$ matrix of constants and $1 \leq h \leq L+1$ is a number of linearly independent hypotheses. An important special case is the omnibus test considering

$$H_0 : \beta_1(t) = \ldots = \beta_L(t) = 0 \, \forall t \in T, \text{ against} \tag{26}$$

$$H_1 : \beta_\ell(t) \neq 0 \text{ for some } \ell \in \{1, \ldots, L\} \text{ and some } t \in T, \tag{27}$$

to compare the full and the null models. In this case, $\mathbf{C}$ is an $L \times (L+1)$ block-matrix partitioned as $\mathbf{C} = (\mathbf{0}|\mathbf{I})$, where $\mathbf{0}$ is a vector of zeros of length L and $\mathbf{I}$ is an $(L \times L)$ identity matrix. In the case of testing the effect of a single regression function $\beta_\ell(t), \ell \in \{1, \ldots, L\}$, through the hypothesis

$$H_0^\ell : \beta_\ell(t) = 0 \, \forall t \in T, \text{ against } H_1^\ell : \beta_\ell(t) \neq 0 \text{ for some } t \in T, \tag{28}$$

the matrix $\mathbf{C}$ is reduced to a unit row vector with 1 on the ℓ-th entry and zeros otherwise, i.e, $\mathbf{C}_\ell = (0, \ldots, 0, 1, 0, \ldots, 0)$.

If the variance-covariance matrix has the form $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, one can use the test statistic [44]

$$\mathrm{T}_{\mathrm{OLS}} = \int_T \frac{1}{\widehat{\sigma}^2} \big(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\big)' \left[\mathbf{C}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \big(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\big) dt \tag{29}$$

for testing the hypothesis (25). The test of a single regression function $\beta_\ell(t)$, $\ell \in \{1, \ldots, L\}$, expressed through the hypothesis (28), can be run using the test

statistic

$$\mathrm{T}^{\ell}_{\mathrm{OLS}} = \int_T \frac{\left(\mathbf{C}_\ell \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)^2}{\widehat{\sigma}^2 \mathbf{C}_\ell (\mathbf{F}'\mathbf{F})^{-1} \mathbf{C}'_\ell} dt. \tag{30}$$

Without the presence of spatial correlation and under homoscedasticity, a decision about the hypotheses (26) and (28) can be made by the Freedman and Lane permutation scheme (see [20]) suitably adapted to the functional data as in [1]. This procedure allows one to estimate the distribution of the test statistics under the random permutations by Monte Carlo techniques. In particular, the residuals $\widehat{\boldsymbol{\varepsilon}}$ of model (6) are estimated from the reduced model (i.e., the model fitted under the null hypothesis) and they are permuted at each iteration, each time creating a permuted data set over which the test statistic is evaluated. Under H_0, the distribution of the test statistic is asymptotically invariant to permutations. The asymptotical invariance is a consequence of performing the test on the estimated and therefore asymptotically exchangeable residuals. Finally, one can compute the global p-value of the test statistic as the proportion of permutations leading to a value of the test statistic that is higher than or equal to the statistic from the observed data.

Under the presence of heteroscedasticity or spatial correlation, a modification of the testing procedure is necessary to account for the heterogeneous structure of the estimated residuals $\widehat{\boldsymbol{\delta}}$ which are no longer asymptotically exchangeable if characterized by a heteroscedastic or spatial structure.

Without the presence of spatial correlation and under heteroscedasticity, the hypothesis (25) can be tested using a test statistic accounting for the covariance structure of the residuals,

$$\mathrm{T}_{\mathrm{WLS}} = \int_T \left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}\right)' \left[\mathbf{C}(\mathbf{F}'\widehat{\mathbf{W}}\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}\right) dt, \tag{31}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}$ is the weighted least-squares estimator of $\boldsymbol{\beta}$, obtained using the two-step iteration procedure for the estimation of the regression functions. One can proceed analogously in the case of the test of a single regression function (see

hypothesis (28)), using the test statistic

$$\mathrm{T}^{\ell}_{\mathrm{WLS}} = \int_T \frac{\left(\mathbf{C}_\ell \widehat{\boldsymbol{\beta}}_{\mathrm{WLS}}\right)^2}{\mathbf{C}_\ell (\mathbf{F}' \widehat{\mathbf{W}} \mathbf{F})^{-1} \mathbf{C}'_\ell} dt. \tag{32}$$

Define the vector $\boldsymbol{\varepsilon} = \mathbf{W}^{1/2}\boldsymbol{\delta}$, as

$$\boldsymbol{\varepsilon}_{s_i} = w_{ii}^{1/2}\, \delta_{s_i}, s_i \in D, i \in \{1, \ldots, n\}. \tag{33}$$

Clearly, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathbf{I}$, and $\{\varepsilon_{s_i}\}, s_i \in D, i \in \{1, \ldots, n\}$, are exchangeable. In principle, one may specify any model for the weights $w_{ii}, i \in \{1, \ldots, n\}$, without substantial modifications of the method here proposed.

In this case, we propose to consider for the permutation scheme the standardized residuals $\boldsymbol{\varepsilon}$, which can be obtained from the estimated residuals $\widehat{\boldsymbol{\delta}}$ as $\widehat{\boldsymbol{\varepsilon}} = \widehat{\mathbf{W}}^{1/2}\widehat{\boldsymbol{\delta}}$. Assuming the homoscedasticity within the groups and the heteroscedasticity between the groups, the variance $\sigma^2_{(m)}$ in the m-th group can be estimated via the sample variance within this group (Equation (20)). The estimated standardized residuals $\widehat{\boldsymbol{\varepsilon}}$ are then randomly permuted, and the corresponding permuted responses are used similarly as in the case of homoscedasticity. This permutation scheme is a new proposal for the case of heteroscedasticity among the observations.

Similarly to the case of heteroscedasticity, the residuals $\widehat{\boldsymbol{\delta}}$ are no longer asymptotically exchangeable if characterized by a spatial correlation. As an approach to dealing with this issue, we propose a modification of the testing procedure which will be here referred to as the spatial functional Freedman and Lane Permutation scheme (SF-FLPerm). In the spatial functional setting, assuming a global variance-covariance matrix $\boldsymbol{\Sigma}$, the test statistic can be formulated as

$$\mathrm{T}_{\mathrm{GLS}} = \int_T \left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)' \left[\mathbf{C}(\mathbf{F}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right) dt, \tag{34}$$

where $\widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ is the generalized least-squares estimator of $\boldsymbol{\beta}$, obtained using the two-step iteration procedure for the estimation of the regression functions. One

can analogously proceed when testing the hypothesis (28), using the test statistic

$$\mathrm{T}^{\ell}_{\mathrm{GLS}} = \int_T \frac{\left(\mathbf{C}_\ell \widehat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)^2}{\mathbf{C}_\ell (\mathbf{F}' \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{F})^{-1} \mathbf{C}'_\ell} dt. \tag{35}$$

Define the vector of spatially filtered residuals $\boldsymbol{\varepsilon}$ as

$$\boldsymbol{\varepsilon} = \boldsymbol{\Sigma}^{-1/2}\, \boldsymbol{\delta}, \tag{36}$$

where $\boldsymbol{\Sigma}^{-1/2}$ is a symmetric square root of $\boldsymbol{\Sigma}^{-1}$ [27]. Clearly, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathbf{I}$, and $\{\varepsilon_{s_i}\}, s_i \in D, i \in \{1, \ldots, n\}$, are exchangeable.

The key difference of the SF-FLPerm scheme, compared to the Freedman and Lane scheme (see [20]) is that the inference is based on the permutations of the spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}$ instead of on the permutations of the correlated residuals $\widehat{\boldsymbol{\delta}}$. In more detail, at each iteration, the residuals $\widehat{\boldsymbol{\delta}}$ are de-correlated by using the inverse-square root of the residual spatial covariance $\widehat{\boldsymbol{\Sigma}}$, i.e., $\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\delta}}$. The spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}$ are permuted and then re-correlated by using the square root of $\widehat{\boldsymbol{\Sigma}}$, i.e., $\widehat{\boldsymbol{\delta}} = \widehat{\boldsymbol{\Sigma}}^{1/2}\widehat{\boldsymbol{\varepsilon}}$, to create the permuted data set. That is, the distribution of the $\mathrm{T}_{\mathrm{GLS}}$ ($\mathrm{T}^{\ell}_{\mathrm{GLS}}$) test statistic under the permutations is estimated in a similar manner as in the homoscedasticity or heteroscedasticity cases, but through the estimated approximately exchangeable spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}$. The SF-FLPerm scheme is a generalization of the Freedman and Lane scheme and can properly perform the inference procedure under the presence of spatial correlation. The detailed overview of the spatial functional Freedman and Lane Permutation scheme is provided by Algorithm 1 which is designed for testing the general linear hypothesis about the effect of the covariates, extending the scheme for heteroscedastic observations proposed in [44] to the more complex structures of the data. Note that, to avoid repetition, an algorithm is not shown for OLS and WLS cases and for the single-hypothesis testing, but Algorithm 1 can be adapted to these cases as well.

Algorithm 1 Spatial functional Freedman and Lane permutation scheme for hypothesis (25).

1. Compute the test statistic $\mathrm{T}_{\mathrm{GLS}}^{\mathrm{data}}$ from the original data.
2. During the e-th permutation, $e \in \{1, \ldots, \mathcal{E}\}$,
   (a) Determine the estimated residuals $\widehat{\delta}_{s_i,r}$ of the reduced model $\mathcal{X}_{s_i,r} = \beta_0 + \delta_{s_i,r}$ using the iterative two-step procedure. Obtain $\widehat{\beta}_0, \widehat{\boldsymbol{\Sigma}}_r, \widehat{\delta}_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (b) Determine the exchangeable estimated spatially filtered residuals $\widehat{\boldsymbol{\varepsilon}}_r$ from $\widehat{\boldsymbol{\delta}}_r$ as $\widehat{\boldsymbol{\varepsilon}}_r = \widehat{\boldsymbol{\Sigma}}_r^{-1/2}\widehat{\boldsymbol{\delta}}_r$.
   (c) Permute the estimated exchangeable spatially filtered residuals $\widehat{\varepsilon}_{s_1,r}, \ldots, \widehat{\varepsilon}_{s_n,r}$, obtaining $\widehat{\varepsilon}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (d) Compute the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ through the fitted reduced model and the permuted estimated spatially correlated residuals $\widehat{\boldsymbol{\delta}}^*_r = \widehat{\boldsymbol{\Sigma}}_r^{1/2}\widehat{\boldsymbol{\varepsilon}}^*_r$ as $\mathcal{X}^*_{s_i,r} = \widehat{\beta}_0 + \widehat{\delta}^*_{s_i,r}, s_i \in D, i \in \{1, \ldots, n\}$.
   (e) Estimate the regression functions $\boldsymbol{\beta}$ of the full model from the permuted responses $\boldsymbol{\mathcal{X}}^*_r$ by the GLS (Equation (22)) using the iterative two-step procedure. Obtain $\widehat{\boldsymbol{\beta}}^*_{\mathrm{GLS}}, \widehat{\boldsymbol{\delta}}^*_p(t)$, $\widehat{\boldsymbol{\Sigma}}_p$.
   (f) Compute the test statistic $\mathrm{T}^{*e}_{\mathrm{GLS}}$ as $\mathrm{T}^{*e}_{\mathrm{GLS}} = \int_T \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{GLS}}\right)' \left[\mathbf{C}(\mathbf{F}'\widehat{\boldsymbol{\Sigma}}_p^{-1}\mathbf{F})^{-1}\mathbf{C}'\right]^{-1} \left(\mathbf{C}\widehat{\boldsymbol{\beta}}^*_{\mathrm{GLS}}\right) dt$.
3. Compute the global p-value of the $\mathrm{T}_{\mathrm{GLS}}$ test as the proportion of all permutations for which $\mathrm{T}^{*e}_{\mathrm{GLS}} \geq \mathrm{T}^{\mathrm{data}}_{\mathrm{GLS}}, e \in \{1, \ldots, \mathcal{E}\}$.

6.2. Assessment of the model assumptions under heteroscedasticity

As was outlined earlier, three possible scenarios may occur under the spatial setting. Which approach to choose can be decided by the shape of the semivariogram. Firstly, the concave shape, where the semivariogram increases in the beginning part and then becomes constant or non-decreasing, is associated with the spatial correlation and a general variance-covariance matrix $\boldsymbol{\Sigma}$. Secondly, the constant shape of the semivariogram (i.e., the pure nugget structure) is associated with a diagonal variance-covariance matrix ($\boldsymbol{\Sigma}$ equals either $\sigma^2\mathbf{I}$, or $\mathbf{W}$).

However, even for the case of spatially uncorrelated residuals, one still needs to verify the assumption of homoscedasticity, as this determines the exchangeability of the residuals. For this reason, we employ a test based on the same permutations as described in Sections 3.4, 3.5, and 4.4, but with a test statistic specifically aiming at detecting the differences in variances. Note that, using the multiple comparison principles, the test can be extended to the case of more populations.

Let $\delta_{s_i(m)}(t), s_i \in D, i \in \{1, \ldots, n_m\}, m \in \{A, B\}$, be two groups of the functional residuals from model in Equations (11) and (12), where i is the unit index and m is the population index. Let $\sigma^2_{(A)}$ and $\sigma^2_{(B)}$ be the global variances (constant over T) of these two populations, that is, $\sigma^2_{(m)} = \mathrm{E}[\|\delta_{s_i(m)}\|^2]$. The latter can be estimated by Equation (20), restricted to the m-th group in the data. Note that, by construction, the overall residual sample mean is zero; furthermore, the residual sample mean within the groups is zero if the dummy variables are included among the covariates (i.e., in a functional analysis of covariance setting). The goal is now to test the hypothesis

$$H_0 : \sigma^2_{(A)} = \sigma^2_{(B)}, \text{ against } H_1 : \sigma^2_{(A)} \neq \sigma^2_{(B)}. \tag{37}$$

For this purpose, we propose as a test statistic the absolute value of the log-proportion of variances, as follows:

$$\mathrm{T}_{\mathrm{Var}} = \left| \log\left(\frac{\widehat{\sigma}^2_{(A)}}{\widehat{\sigma}^2_{(B)}} \right) \right|. \tag{38}$$

To perform the test, we consider a permutation scheme similar to those discussed previously. Under H_0, the residuals are approximately exchangeable, and the permutation procedure from [1] can be applied. The testing procedure is as follows. During each permutation, the function's indices are ordered as $1, 2, \ldots, n_A$ (group A), and $n_A+1, n_A+2, \ldots, n_A+n_B$ (group B). Then, the estimated residuals $\widehat{\delta}_{s_i(m)}$, $m \in \{A, B\}$, $i \in \{1, \ldots, n\}$, from both groups are permuted together, obtaining the permuted estimated residuals $\widehat{\delta}^*_{s_1}, \ldots, \widehat{\delta}^*_{s_{n_A}}$, belonging to group A, and the permuted estimated residuals $\widehat{\delta}^*_{s_{n_A+1}}, \ldots, \widehat{\delta}^*_{s_{n_A+n_B}}$, belonging to group B.

Using the permuted estimated residuals, the group variances $\widehat{\sigma}^{*2}_{(A)}$ and $\widehat{\sigma}^{*2}_{(B)}$ are estimated (Equation (20)). Then, the test statistic $\mathrm{T}_{\mathrm{Var}}$ is calculated (Equation (38)). Finally, the global p-value of the $\mathrm{T}_{\mathrm{Var}}$ test (37) is computed as the proportion of the permutations leading to a value of $\mathrm{T}_{\mathrm{Var}}$ higher than or equal to the one observed in the data. The results can be used to establish which permutation strategy should be used to test the significance of the regression functions. Note that, the proposed statistical inference concerning homoscedasticity is only approximate due to its definition on the estimated residuals.

6.3. The global permutation tests for the effect of the covariates in functional-on-functional regression model

The omnibus hypothesis (26) can be equivalently written as

$$H_0 : \mathbf{b}_1 = \mathbf{b}_2 = \ldots = \mathbf{b}_L = 0, \text{ against } H_1 : \mathbf{b}_\ell \neq 0 \text{ for some } \ell \in \{1, \ldots, L\}, \tag{39}$$

or in a matrix form as

$$H_0 : \boldsymbol{\mathcal{C}}\mathbf{b} = \mathbf{0}, \text{ against } H_1 : \boldsymbol{\mathcal{C}}\mathbf{b} \neq \mathbf{0}, \tag{40}$$

where $\boldsymbol{\mathcal{C}}$ is a $(\sum_{\ell=1}^{L} K_{\beta,\ell}) \times K_\beta$ block matrix of the form $\boldsymbol{\mathcal{C}} = (\mathbf{0}|\mathbf{I})$, where $\mathbf{0}$ is a $\sum_{\ell=1}^{L} K_{\beta,\ell} \times K_0$-dimensional matrix of zeros and $\mathbf{I}$ is an $(\sum_{\ell=1}^{L} K_{\beta,\ell} \times \sum_{\ell=1}^{L} K_{\beta,\ell})$ identity matrix. To verify the omnibus hypothesis, one can use the test statistic

$$\mathcal{T} = \frac{1}{\widehat{\sigma}^2} \widehat{\mathbf{b}}' \boldsymbol{\mathcal{C}}' \left[\boldsymbol{\mathcal{C}} (\tilde{\mathbf{F}}' \tilde{\mathbf{F}} + \mathbf{R}_\beta)^{-1} \boldsymbol{\mathcal{C}}' \right]^{-1} \boldsymbol{\mathcal{C}} \widehat{\mathbf{b}}. \tag{41}$$

The test for the significance of a single regression function $\beta_\ell(t), \ell \in \{1, \ldots, L\}$, through the hypothesis

$$H_0^\ell : \mathbf{b}_\ell = 0, \text{ against } H_1^\ell : \mathbf{b}_\ell \neq 0, \tag{42}$$

can be run using the test statistic

$$\mathcal{T}_\ell = \frac{1}{\widehat{\sigma}^2} \widehat{\mathbf{b}}_\ell' (\tilde{\mathbf{F}}_\ell' \tilde{\mathbf{F}}_\ell + \mathbf{R}_l) \widehat{\mathbf{b}}_\ell, \tag{43}$$

where $\tilde{\mathbf{F}}_\ell$ denotes the block-diagonal part of $\tilde{\mathbf{F}}$ corresponding to $\mathbf{b}_\ell$.

7. Original results and summary

This thesis introduced a novel approach to the nonparametric inference for the functional data. The main focus was on the functional regression models, where the testing for the significance of the regression functions was performed by permuting the residuals of the model under the null hypothesis. In this setting, the aim was to ensure the exchangeability of the residuals being permuted.

The first aim was to define a regression model with the functional covariate and the functional response and the inferential procedure for testing the significance of the regression functions based on the Freedman and Lane permutation scheme (see [20]) adapted to the functional setting [1]. Here, to ensure the approximate exchangeability of the permuted objects, the residuals, rather than the observations, are of interest. From the application viewpoint, the objective was to analyse the data from the field of the transportation research, consisting of the functional observations of the driving speed on six expressway ramps collected by a vehicle fleet on an interchange in Brno, Czech Republic. Moreover, it was also of interest to examine how the driving speed differs in the different groups in the data, such as between the main and the auxiliary (merging or exit) lane. This assumption of the difference was confirmed by the testing performed in the regression models where the presence of the auxiliary lane was included as one of the explanatory variables, together with the curvature of the ramp (represented by the radius of the circle fitted to a specific part of the ramp), evaluated by the permutation-based test in a functional-on-functional regression model based on the Freedman and Lane's methodology. Similarly to the previous studies [14, 30], the relationship between the driving speed and the ramp curvature represented by the radius of the circle fitted to the specific part of the ramp by the least-square-circle fit method was confirmed: on the less complex ramps the, relationship between the speed and the radius was rather weak, which could be observed from the results of testing for the significance of the radius as a regression function.

The second aim was to define a spatiotemporal regression model, together with an introduction of a novel inferential framework for testing in the spatial

functional regression model, accounting for the more complex structures of the data. Firstly, concerning a testing procedure suitable for heteroscedastic functional data, we propose to fit a weighted least-squares functional regression model to the data, where the residuals are allowed to be characterized by heterogeneous variances across the groups. The permutational schemes of [1] and [20] are thus extended to the standardized residuals of the model, obtained by dividing the estimated residuals by the estimated standard deviation of the corresponding group of the data. Then, the standardized residuals obtained by this procedure are assumed to be approximately exchangeable. With the purpose of assessing the assumption of homoscedasticity, we also proposed a permutation-based test for the equality of the variances in two groups of the data. Note that, the proposed model and the testing procedures can be easily extended to any number $\mathcal{G} \geq 2$ of groups. In this case, the mean term would be associated with $\mathcal{G} - 1$ dummy variables, and the testing procedure could be performed (jointly) on the associated coefficients, similarly as detailed in this thesis. In this more general situation, the heteroscedasticity among the groups may be expected; the tests for heteroscedasticity could be developed by using an analysis of variance-like setting, or the multiple pairwise tests with appropriate level corrections. The performance of this proposed methodology was assessed by the extensive simulation studies showing that the introduced WLS test performs well concerning the empirical size and the power, especially in the case of the unbalanced designs. In contrast, if the nonconstant variance in the data is neglected and an ordinary least-squares approach is applied, the test is either too liberal or too conservative whenever the two populations in the data have unequal sample sizes. Based on the simulations, the empirical size of the OLS test could be more than 10 times smaller, or almost five times greater than the nominal one. However, for the balanced design, the OLS test works well even under the presence of heteroscedasticity. These results are consistent with the previous studies pointing out that the effect of heteroscedasticity is stronger for the unbalanced designs [10, 36]. In fact, one should pay close attention to the specific spatial structure of

the data during the model fitting, especially in the case of the unbalanced design.

Secondly, we focused on a case of the spatial correlation among the observations. Under this setting, the drift is estimated by a generalized least-squares model and then the spatial correlation among the estimated residuals is assessed by the variography with the purpose of a correct estimation of the variance-covariance structure. The potential relationship between the spatial covariates and the functional response is then assessed by the Spatial Functional Freedman and Lane permutation scheme proposed here, based on permuting the estimated residuals, spatially filtered through the pre-multiplication by the inverse square root of the variance-covariance matrix, ensuring that the units to be permuted are asymptotically exchangeable and that the procedure is properly conducted, respecting the georeferenced nature of the data. The simulation studies intended to evaluate the performance of this inferential procedure in terms of the empirical size and the power show that the GLS test performs well whenever the range of the spatial correlation is not too high (more than half of the size of the spatial area). For the larger ranges, the procedure requires a larger sample size. This is likely to be due to the difficulty of estimating the spatial covariance structure when the range is large with respect to the study area D, which is well-known in the literature of the spatial statistics [12]. On the other hand, the ordinary least-squares approach shows very liberal behaviour whenever the data are correlated in space. With the increasing range of the spatial correlation, the test becomes even more liberal, its empirical size approaching to 1. Neglecting the underlying spatial structure of the data thus leads to an inference based on the non-exchangeable residuals and biases the results in terms of the empirical size of the test. The empirical power of the test was therefore evaluated for the generalized least-squares approach only. Three different models were considered: two-covariate, one-covariate and a misspecified one, where the data were generated as a model with one covariate, but a two-covariate model was assumed and fitted. The empirical power is similar for all three models.

In the current work, we propose the OLS, WLS, and GLS classical test statis-

tics integrated over the domain of interest, following the work of [1]. These test statistics are coherent with the classical tests of the hypotheses in the linear regression and they cover a wide variety of different tests (i.e., all tests of the linear hypotheses on the regression functions of the model), and in the absence of any information about the data-generative model, we suggest to use them. However, since the nonparametric inference based on the permutation tests is performed, this is not the only possible choice; an inference with the similar properties could be performed using the different point-wise statistics, the weighted integrals (if some information about the weighting is available), or the maximum values along the domain. Some test statistics could achieve a higher power under the specific alternatives. However, a thorough comparison between the different choices is out of the scope of this thesis, but it would be an interesting future line of research.

8. List of publications

Related to a dissertation thesis

- **V. Římalová**, A. Menafoglio, A. Pini, V. Pechanec, and E. Fišerová. A permutation approach to the analysis of spatiotemporal geochemical data in the presence of heteroscedasticity. *Environmetrics*, 31(4):e2611, 2020.
- **V. Římalová**, E. Fišerová, A. Menafoglio, and A. Pini. Inference for spatial regression models with functional response using a permutational approach. *Journal of Multivariate Analysis*, 189:104893, 2022.
- **V. Římalová**, J. Elgner, J. Ambros, and E. Fišerová. Modelling the driving speed on expressway ramps based on floating car data. *Measurement*, 195:110995, 2022.

Other publications

- V. Pechanec, T. Pohanka, H. Kilianová, E. Fišerová, and V. Římalová. Variability of soil carbon on land borderlines. *Listy Cukrovarnické a Řepařské*, 135(4), 2019.
- R. Marciniak, R. Šumec, M. Vyhnálek, K. Bendíčková, P. Lázničková, G. Forte, A. Jeleník, V. Římalová, J. Frič, J. Hort, and K. Sheardová. The effect of mindfulness-based stress reduction (MBSR) on depression, cognition, and immunity in mild cognitive impairment: A pilot feasibility study. *Clinical interventions in aging*, 1365-1381, 2020.
- M. Zeidler, E. Fišerová, V. Římalová, M. Bednář, M. Banaš. Sex-based differences in vigor and site preferences of *Juniperus communis* subsp. *nana*. *Nordic Journal of Botany*, 38(11), 2020.
- V. Biel, J. Novak, V. Římalová, M. Tomandlova, J. Tomandl, J. Spaca, and M. Soucek. Levels of endothelial substances in patients with newly identified hypertension compared with healthy controls. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub*, 165(4), 395-401, 2021.

- E. Odehnalová, L. Valíková, G. Caluori, T. Kulík, V. Římalová, T. Jadczyk, E. Dražanová, I. Pavlova, M. Pešl, V. Kubeš, and Z. Stárek. Comparison of gross pathology inspection and 9.4 T magnetic resonance imaging in the evaluation of radiofrequency ablation lesions in the left ventricle of the swine heart. *Frontiers in Physiology*, 2204, 2022.

9. List of conferences

- ODAM, 31. 5. - 1. 6. 2017, Olomouc, Czech Republic: On functional data analysis with application to geology (presentation)
- ROBUST, 21. - 26. 1. 2018, Rybník, Czech Republic: Analýza prostorově závislých funkcionálních dat (poster + presentation, in Czech; awarded in Students Works Competition)
- Workshop on Functional Data Analysis 12. 7. 2018, Prague, Czech Republic: An inferential framework for the analysis of spatio-temporal geochemical data (presentation)
- LinStat, 20. - 24. 8. 2018, Będlewo, Poland: Spatio-temporal geochemical data: a novel inferential framework for the analysis (presentation; awarded by 3rd place in Young Scientists Awards)
- IAMG, 2. - 8. 9. 2018, Olomouc, Czech Republic: A novel inferential framework for the analysis of spatio-temporal geochemical data (presentation)
- CRoNoS & MDA, 14. - 16. 4. 2019, Limassol, Cyprus: A permutation approach to the analysis of spatio-temporal geochemical data under heteroscedasticity (invited presentation)
- ODAM, 29. - 31. 5. 2019, Olomouc, Czech Republic: A permutation approach to the analysis of spatio-temporal geochemical data in the presence of heteroscedasticity (presentation)
- AMISTAT, 16. - 19. 9. 2019, Liberec, Czech Republic, A permutation approach to the analysis of spatio-temporal geochemical data (presentation)
- LinStat, 30. 8. - 3. 9. 2021, Będlewo, Poland: Inference for spatial regression models with functional response using a permutational approach (invited presentation)

References

[1] K. Abramowicz, C. K. Häger, A. Pini, L. Schelin, S. Sjöstedt de Luna, and S. Vantini. Nonparametric inference for functional-on-scalar linear models applied to knee kinematic hop data after injury of the anterior cruciate ligament. *Scandinavian Journal of Statistics*, 45(4):1036–1061, 2018.

[2] M. J. Anderson and J. Robinson. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2001.

[3] G. Aneiros, R. Cao, R. Fraiman, C. Genest, and P. Vieu. Recent advances in functional data analysis and high-dimensional statistics. *Journal of Multivariate Analysis*, 170:3–9, 2019.

[4] G. Aneiros, I. Horová, M. Hušková, and P. Vieu. On functional data analysis and related topics. *Journal of Multivariate Analysis*, 189:104861, 2022.

[5] J.-P. Aristizabal, R. Giraldo, and J. Mateu. Analysis of variance for spatially correlated functional data: application to brain data. *Spatial Statistics*, 32:100381, 2019.

[6] E. Arnone, L. Azzimonti, F. Nobile, and L. M. Sangalli. Modeling spatially dependent functional data via regression with differential regularization. *Journal of Multivariate Analysis*, 170:275–295, 2019.

[7] R. D. Baker. Two permutation tests of equality of variances. *Statistics and Computing*, 5(4):289–296, 1995.

[8] E. Belli. Smoothly adaptively centered ridge estimator. *Journal of Multivariate Analysis*, 189:104882, 2022.

[9] M. S. Bernardi, L. M. Sangalli, G. Mazza, and J. O. Ramsay. A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stochastic environmental research and risk assessment*, 31(1):23–38, 2017.

[10] G. E. Box et al. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the

one-way classification. *The annals of mathematical statistics*, 25(2):290–302, 1954.

[11] J. A. Collazos, R. Dias, and A. Z. Zambom. Consistent variable selection for functional regression models. *Journal of Multivariate Analysis*, 146:63–71, 2016.

[12] N. A. Cressie. Statistics for spatial data. John Wiley and Sons. *Inc., New York*, 1993.

[13] P. Delicado, R. Giraldo, C. Comas, and J. Mateu. Statistics for spatial functional data: some recent contributions. *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):224–239, 2010.

[14] C. Dias, T. Oguchi, and K. Wimalasena. Drivers' speeding behavior on expressway curves: exploring the effect of curve radius and desired speed. *Transportation research record*, 2672:48–60, 2018.

[15] J. Diquigiovanni, M. Fontana, and S. Vantini. Conformal prediction bands for multivariate functional data. *Journal of Multivariate Analysis*, 189:104879, 2022.

[16] B. Erbas, M. Akram, D. M. Gertig, D. English, J. L. Hopper, A. M. Kavanagh, and R. Hyndman. Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the United States and England–Wales. *Journal of epidemiology*, 20(2):159–165, 2010.

[17] B. Erbas, R. J. Hyndman, and D. M. Gertig. Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine*, 26(2):458–470, 2007.

[18] M. Escabias, A. Aguilera, and M. Valderrama. Modeling environmental data by functional principal component logistic regression. *Environmetrics: The official journal of the International Environmetrics Society*, 16(1):95–107, 2005.

[19] F. Ferraty and P. Vieu. *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[20] D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298, 1983.

[21] R. Giraldo, W. Caballero, and J. Camacho-Tamayo. Mantel test for spatial functional data. *AStA Advances in Statistical Analysis*, 102(1):21–39, 2018.

[22] R. Giraldo, P. Delicado, and J. Mateu. Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, 18(3):411–426, 2011.

[23] A. Goia and P. Vieu. An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis*, 146:1–6, 2016.

[24] P. Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

[25] L. Horváth and P. Kokoszka. *Inference for functional data with applications*. Springer Science & Business Media, 2012.

[26] R. Ignaccolo, J. Mateu, and R. Giraldo. Kriging with external drift for functional data for air quality monitoring. *Stochastic environmental research and risk assessment*, 28(5):1171–1186, 2014.

[27] H.-J. Kim, M. P. Fay, E. J. Feuer, and D. N. Midthune. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*, 19(3):335–351, 2000.

[28] P. Kokoszka and M. Reimherr. *Introduction to functional data analysis*. Chapman and Hall/CRC, 2017.

[29] P. Kokoszka and M. Reimherr. Some recent developments in inference for geostatistical functional data. *Revista Colombiana de Estadística*, 42(1):101–122, 2019.

[30] E. D. Liapis, B. Psarianos, and E. Kasapi. Speed behavior analysis at curved ramp sections of minor interchanges. *Transportation Research Record*, 1751:35–43, 2001.

[31] B. F. Manly. *Randomization, bootstrap and Monte Carlo methods in biology.* Chapman and Hall/CRC, 2018.

[32] M. Marozzi. Some notes on nonparametric inferences and permutation tests. *Metron International Journal of Statistics*, 60(3-4):139–151, 2002.

[33] A. Menafoglio, O. Grujic, and J. Caers. Universal Kriging of functional data: Trace-variography vs cross-variography? Application to gas forecasting in unconventional shales. *Spatial Statistics*, 15:39–55, 2016.

[34] A. Menafoglio and P. Secchi. Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European journal of operational research*, 258(2):401–410, 2017.

[35] A. Menafoglio, P. Secchi, and M. Dalla Rosa. A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space. *Electronic Journal of Statistics*, 7:2209–2240, 2013.

[36] K. Moder. Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling*, 52(4):343–353, 2010.

[37] H. Oja. On permutation tests in multiple regression and analysis of covariance problems. *Australian Journal of Statistics*, 29(1):91–100, 1987.

[38] F. Pesarin and L. Salmaso. *Permutation tests for complex data: theory, applications and software.* John Wiley & Sons, 2010.

[39] A. Pini and S. Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2):407–424, 2017.

[40] J. Ramsay, G. Hooker, and S. Graves. *Functional data analysis with R and MATLAB.* Springer-Verlag New York, 2009.

[41] J. Ramsay and B. W. Silverman. *Functional data analysis.* Springer New York, 2005.

[42] V. Římalová, J. Elgner, J. Ambros, and E. Fišerová. Modelling the driving speed on expressway ramps based on floating car data. *Measurement*, 195:110995, 2022.

[43] V. Římalová, E. Fišerová, A. Menafoglio, and A. Pini. Inference for spatial regression models with functional response using a permutational approach. *Journal of Multivariate Analysis*, 189:104893, 2022.

[44] V. Římalová, A. Menafoglio, A. Pini, V. Pechanec, and E. Fišerová. A permutation approach to the analysis of spatiotemporal geochemical data in the presence of heteroscedasticity. *Environmetrics*, 31(4):e2611, 2020.

[45] E. Romano, J. Mateu, and R. Giraldo. On the performance of two clustering methods for spatial functional data. *AStA Advances in Statistical Analysis*, 99(4):467–492, 2015.

[46] E. Romano, R. Verde, and V. Cozza. Clustering Spatial Functional Data: A Method Based on a Nonparametric Variogram Estimation. In *New Perspectives in Statistical Modeling and Data Analysis*, pages 339–346. Springer, 2011.

[47] K. Sulovská, E. Fišerová, M. Chvosteková, and M. Adámek. Appropriateness of gait analysis for biometrics: Initial study using FDA method. *Measurement*, 105:1–10, 2017.

[48] C. J. ter Braak. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and related techniques*, pages 79–85. Springer, 1992.

[49] A. M. Winkler, G. R. Ridgway, M. A. Webster, S. M. Smith, and T. E. Nichols. Permutation inference for the general linear model. *Neuroimage*, 92:381–397, 2014.