Improving The Applicability of Functional Regression in Econometrics

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

eingereicht an der

Fakultät für Wirtschaftswissenschaften der Universität Regensburg

vorgelegt von Christoph Rust

Berichterstatter:

Prof. Dr. Rolf Tschernig, Universität Regensburg
Prof. Dr. Dominik Liebl, Universität Bonn

Tag der Disputation: 21. Februar 2022

Acknowledgments

At the first place, I want to express my gratitude to Prof. Rolf Tschernig for providing me the opportunity to work on my doctoral thesis. I really have enjoyed my time at the Chair of Econometrics, which not only was a stimulating working atmosphere and a place for inspiring discussions but also provided a living environment which was about much more than science. I am profoundly grateful to Prof. Dominik Liebl from whom I have learned a lot and who also helped me substantially in developing my ideas. For arousing my interest in the research topic, I want to thank Stefan Rameseder.

Many colleagues and fellow Ph.D. students at the University of Regensburg and the IAB in Nuremberg have substantially supported my work, not only by providing invaluable and insightful comments but also by being a source of motivation. I deeply appreciate to have worked with Johann Eppelsheimer, Philipp Gersing, Peter Haller, Tobias Hartl, and Stephan Huber.

Finally, I am profoundly grateful to my family, in particular Lena and our children. Without their understanding and unwavering support all this would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and overview

1.1 Introduction

During the last two decades, a branch of statistics called "functional data analysis" (FDA) has become quite popular. Numerous applications, especially related to environmental statistics, biostatistics, but also sports statistics, to name a few, have proven its usefulness. FDA's growing importance is due not least to the increased availability of high-resolution data, which, in many cases, is even automatically recorded. Very often, such data is generated by a continuous process, which, in principle, can be sampled at arbitrary discretization points. Observations of such random functions are called *functional data* and FDA refers to the statistical analysis of such data. Typical examples of such random functions in the time domain are growth curves or temperature curves.

Although much economic data, by their nature, can be regarded as random curves, applications of FDA have been relatively rare in the context of economic data. The author of this thesis believes that the following two reasons could have played a major role in this. First, some of FDA's important methods do not provide the same level of structural interpretability as their classical counterparts. This is the case in particular for the functional regression model with a scalar response, which, potentially, could be an adequate framework for a variety of different applications in the context of economic data. Second, applying methods of FDA involves some theoretical but also technical obstacles. The theoretical foundations of FDA are taken from functional analysis (in particular the theory related to infinite dimensional Hilbert spaces and linear operators). These concepts are typically not covered by the curriculum taken by empiricists, at least this is what the author of this thesis experienced. Usually, courses in econometrics only cover some linear algebra and calculus. In addition, it is not very

straightforward to apply the methods of FDA because it is often necessary to implement (at least some parts of) the algorithms from scratch. Empiricists tend not to use such methods.

This thesis intends to contribute to the literature by developing novel methods, which will be shown to be useful in practice. Aside from these methodological contributions, this thesis also aims to demonstrate that functional regression can be successfully applied to economic data and reveal important insights. Moreover, the implementations are made publicly available as packages for the open-source programming language GNU R with the hope that other empiricists will find them useful.

The remainder of this chapter contains a brief introduction to the methodological framework by giving a formal definition of functional data and by introducing the functional linear regression model with a scalar response variable. The second part of this chapter provides an overview of the individual papers contained in this dissertation.

Functional data analysis is concerned with the statistical analysis of so called "functional data". Functional data refers to (discretized) observations of functional random variables, which are typically modelled as random variables taking values in the Hilbert space $L^2(\mathcal{I})$, where $\mathcal{I}$ is a continuum on which the random functions are defined. In many typical applications $\mathcal{I}$ is a closed interval on the real line and the random functions are curves. This, however, is not a limitation, and in the literature on FDA there are also applications to random surfaces and, more recently, random functions defined on manifolds.

In econometrics, one is often interested in how economic variables influence each other and the empiricist's workhorse is regression analysis. FDA extends regression analysis to functional variables and differentiates between three different types of functional regression models. The first type is a regression model with a functional predictor and a scalar outcome variable, often referred to as the *scalar-on-function regression*. The model where a functional outcome variable is explained by scalar predictors is the second type (*function-on-scalar regression*), and in the third type, a function is regressed on another function (*function-on-function regression*). This dissertation is concerned with the first type of model and addresses methodological extensions and questions related to its applicability in the context of economic data, one of which is interpretability of the estimated quantities. Let us therefore briefly introduce the model.

To this end, we assume that we observe an i.i.d. sample of tuples (y_i, X_i), $i = 1, \ldots, n$, where X_i are realizations of a functional random variable taking values

in the Hilbert space $L^2([0,1])$ and the y_i are generated from the model

$$y_i = \alpha + \int_{[0,1]} \beta(t) X_i(t)\,\mathrm{d}t + \varepsilon_i, \tag{1.1}$$

where ε_i is an scalar-valued i.i.d. error term which has mean zero, finite variance, and is independently distributed from the X_i. For the function-valued parameter β, which determines the dependency between X_i and y_i and often is the quantity of interest, several estimation approaches are available in the literature. Classically, estimation builds on functional principal components (FPCA) which is the cornerstone of many FDA methods. In this setting, both the coefficient function β and the functional observations are approximated in terms of the leading eigenfunctions of the random curves' empirical covariance operator as a basis for a finite dimensional function space. On the one hand, this finite dimensional subspace of L^2 is optimal to approximate the functional predictors for a given number of eigenfunctions but, on the other hand, may not be the best fit to approximate β. Alternatively, the estimation of β can also be achieved by using other basis systems, such as splines.

A fundamental problem, however, arises in the scalar-on-function model if an estimate $\widehat{\beta} \in L^2[0,1]$ has to be interpreted. It has been shown by Cardot et al. (2007) that the quantity $\widehat{\beta} - \beta$ does not converge in distribution to a non-degenerate random element of $L^2[0,1]$. This negative result is a consequence of the infinite-dimensional nature of functional data and shows the limitations of FDA, where classical multivariate results are not generalizable to the functional setup. With this result, it is also not possible to construct confidence bands for the estimate $\widehat{\beta}$ and draw pointwise (with respect to the domain of β) inference about the coefficient function. Without pointwise inference, empiricists do not know whether an observed pattern of a realized estimate $\widehat{\beta}$ in fact reveals characteristics of the underlying data generating mechanism or whether this pattern is random. Consequently, estimation results of a functional regression (with a functional predictor) lack interpretability and are therefore less useful when the purpose of modeling is to understand more about the underlying mechanism. The individual contributions to this field address this problem in different ways and will be briefly summarized in the following.

1.2 Overview and contribution

1.2.1 Functional linear regression with points of impact

The first paper has been published as Liebl et al. (2020) and is coauthored with Dominik Liebl, University of Bonn, and Stefan Rameseder. It extends the literature on an augmented scalar-on-function regression model, namely the functional linear regression model with points of impact. The paper builds on the work of Kneip et al. (2016) and introduces several adjustments to the estimation procedure improving the finite sample performance. The model suggested in Kneip et al. (2016) is defined by

$$Y_i = \int \beta(t) X_i(t)\,\mathrm{d}t + \sum_{s=1}^{S} \beta_s X_i(\tau_s) + \epsilon_i,\ i = 1, \ldots, n. \tag{1.2}$$

Here, the functional predictors not only influence the scalar-valued response variable via the integral, but the functional predictors X_i are also directly related to the outcome variable Y_i at specific influential points $\tau_s \in [0,1], s = 1, \ldots, S$. These so-called *points of impact* are not known a priori and have to be estimated together with the other model parameters $\beta, \beta_1, \ldots, \beta_S$. To identify points of impact, the random functions X_i must not be too smooth, and must possess what Kneip et al. (2016) call *specific local variation*. The principal idea for identifying points of impact is to search for peaks over discretization points t of the correlation between Y_i and $Z_i(t, \delta)$, where $Z_i(t, \delta)$ is the central second-order difference quotient of $X_i(t)$ with differencing parameter $\delta > 0$. Given a set of potential points of impact, the model can be estimated using standard FDA approaches. Kneip et al. (2016) use a finite-dimensional representation of the curves (functional principal components) and select the model complexity (number of principal components, number of points of impact, differencing parameter δ) by minimizing the Bayes Information Criterion (BIC) *simultaneously* over these three parameters. They also suggest theoretically motivated choices for the number of points of impact and the differencing parameter δ, which, however, are not really suited for practical applications.

In our paper, we suggest several adjustments to the original procedure. First, we use a different basis system and expand β using B-splines. As we argue, the originally proposed FPCA approach is optimal to approximate the functional covariates but may be less well-suited to approximating β, in particular if it is a relatively smooth function. Additionally, as we show in a simulation exercise, selecting the number of (FPCA) basis functions and number of points of impact simultaneously may result in severe misspecifications of the model. Second, we

decouple the estimation of the points of impact and the estimation (and choice of complexity) of β by using an alternating procedure which reduces the risk of choosing a misspecified model.

In an extensive simulation study, we show that our procedure outperforms the original approach in an MSE sense. This is in particular true for smaller sample sizes. Besides that, the paper also provides theoretical arguments for these adjustments and treats a case study with real data using data from Google AdWords. This case study also shows the usefulness of the model framework for decomposing the dependency between the functional predictor and the scalar outcome variable into time-global effects (via β) and time-local effects (via points of impact), which facilitates the interpretability of the estimation results at the point-of-impact locations. To make this methodology accessible to practitioners, an R package implementing both the original estimation procedure and the proposed one is available from the author's Github account.

1.2.2 The spatial decay of human capital externalities

The second paper, which is joint work with Johann Eppelsheimer, is mainly an empirical contribution, where functional regression is applied to very precise labour market data to address a significant question in regional economics. An earlier version of the paper included in this thesis is available as an IAB discussion paper (Eppelsheimer and Rust, 2020).

The paper contributes to the extensive literature on the external effects of the accumulation of human capital. In this literature, there is a broad consensus that there are such external effects, which is supported by many empirical studies, showing the existence of geographically bounded externalities (e.g. Moretti, 2004; Cornelissen et al., 2017). One potential channel for such externalities is the occurrence of knowledge transfers increasing a worker's productivity. Such knowledge transfers are more likely to occur in regions with a higher share of the high-skilled population. However, to the best of our knowledge, there is only a limited number of empirical studies addressing the question how large the spatial extent of these effects is. We are aware of two papers, one by Rosenthal and Strange (2008), using cross-sectional data from the US and measuring the effect on wages of the concentration of human capital within concentric rings around workplaces. Using the same strategy, Fu (2007) address the same question using precisely geocoded data of the Boston metropolitan area. The findings in both cases are that knowledge spillovers attenuate with distance and vanish completely beyond a certain distance.

Using data on precisely geocoded workplace locations of almost all German

employees[1], our paper contributes to the existing literature by measuring the spatial extent of human capital spillovers in a much more detailed way. To this end, we compute for every workplace a function measuring the concentration of high-skilled workers with respect to the distance to the given workplace. These curves can be used as functional predictors in a functional wage regression to measure the effect on an individual's wages of the concentration of human capital around the workplace. An estimate of the corresponding functional coefficient then quantifies the magnitude of productivity gains induced by local human capital allocation with respect to the distance. To control for potential confounding market mechanisms, our specification includes fixed effects at various levels.

For this purpose, we augment the classical scalar-on-function regression model to incorporate further scalar-valued explanatory variables and use an estimation procedure based on smoothing splines, suggested by Crambes et al. (2009), which is also used in chapter 2. Of course, the same limitations discussed above regarding inference about (and interpretation of) the functional coefficient also apply, as the model used in this paper also involves a functional predictor. However, to be able to interpret the results, it is necessary to have some notion about the estimator's second moment. Similar to the multivariate regression case, it is possible to compute the sampling variability of an estimate $\widehat{\beta}$ given discretized observations of the functional covariates and tuning parameters of the model. We use this sampling variability of $\widehat{\beta}$ to conduct an inference about β and decide where the functional predictors significantly influence wages. To validate this strategy, we run a simulation study showing that the sampling variability indeed can be used to construct confidence bands around $\widehat{\beta}$. Such confidence bands, although not constituting a formal testing procedure, are helpful for deciding whether the estimation result reflects a genuine signal or whether it is merely driven by noise.

Our results show that spillover effects from human capital concentration decrease with distance and are no longer measurable beyond 20 kilometers. These findings are buttressed using several robustness checks.

1.2.3 Directed local testing in the functional linear model

The third paper is single authored, and inspired by the lack of a formal testing framework that would be appropriate for the research question of the second paper.

Quite often, it is reasonable to assume that β has certain structural properties. In the second paper of this thesis, for instance, it is reasonable to assume that β is

[1] The data contains all employees eligible for social security contributions, hence only civil servants, family workers and the self-employed are not included in the data.

a monotonically decreasing function of distance (to the workplace), approaching zero when the distance gets larger. If such a structural assumption is appropriate, it is possible to get a more 'local' inference about β compared to the global test of association in the functional linear model. While the latter is only capable of testing the null $\beta = \beta_0$ for a given β_0, the approach suggested in the third paper can be used to detect in a data-driven way subparts of the domain where β significantly differs from a given β_0.

To achieve that, the suggested approach uses a sequential procedure where in each step the domain is split into two parts and the global test is performed on one of the two parts. In case the null is rejected in this step, the procedure continues with the next step where the new restricted domain is a subset of the previous one. The first time the test does not reject the null, the procedure stops.

If one can assume, for instance, that β is positive for $t = 0$ and monotonically decreasing towards zero (as in the second paper), the strategy would be to start testing $\beta = 0 \, \forall t \in [0, 1]$, where $[0, 1]$ is the full domain of β. If the test rejects this, then one can continue with the test $\beta = 0 \, \forall t \in [\delta_1, 1]$, for some $\delta_1 > 0$. If this is rejected too, the next test is $\beta = 0 \, \forall t \in [\delta_2, 1]$, where $\delta_2 > \delta_1$, and one repeats the procedure with $\delta_3 < \delta_4 < \ldots$ until the test does not reject the hypothesis the first time, say for δ_{n+1}. Because of the assumption about the parameter function β, it is then possible to conclude that the β is significantly different from zero on the interval $[0, \delta_n]$. This is a much stronger result than can be achieved with the global test.

As the procedure is based on a sequence of tests, it must be ensured that the procedure does not suffer from the multiple comparison problem and that the overall significance level is maintained. As it turns out, the proposed testing procedure already controls the family-wise error rate in the strong sense by design. This result follows from the closed testing principle of Marcus et al. (1976). Besides establishing theoretical results, the paper also contains a simulation study assessing the finite sample performance of the suggested procedure.

The test is applied to data from a sports biomechanics experiment where in a functional regression model the strike index (measuring the centre of pressure when a foot of a runner hits the ground) is explained by torque curves at the ankle joint during the stance phase. Using the test, it can be shown that the strike index is significantly associated to the torque curves on the first 14% of the stance phase and it is argued that the functions on the remaining part of the domain are very unlikely to be associated to the strike index. A package for the statistical language R implementing the testing procedure, making the method accessible to others, is available from the author's Github account.

Chapter 2

Improving the estimation of functional linear regression with points of impact

2.1 Introduction

In many practical applications, one is interested in the relationship between a real-valued outcome variable Y_i and a function-valued predictor $\{X_i(t);\ a \leq t \leq b\}$. In our motivating Google AdWords case study, for instance, we aim to explain the numbers of clicks Y_i using impression trajectories $X_i(t)$, where t denotes a certain day within the considered time interval $[a, b]$ of one year and $i = 1, \ldots, n$ indexes the cross section of keywords associated with the considered Google AdWords ad campaign.[1] The economic success of any ad campaign depends on product specific (time-global) seasonalities as well as on (time-local) events. The slowly varying seasonal component could be estimated using the function-valued slope parameter of the classical functional linear regression model (see, e.g., Hall and Horowitz, 2007). The presence of time-local effects, however, harms such a simple estimation approach (see the right plot in Figure 2.2 for notable examples). Therefore, we use the recent functional linear regression models with so-called Points of Impact (PoI) that allow us to identify and to control for time-local effects.

Point of impact models are originally introduced by McKeague and Sen (2010), who argue that these models are better to interpret than the classical functional linear regression models. Indeed, several convincing real data applications are

[1] Online ad campaigns use text corpora populations of relevant search keywords (for instance, `outdoor jacket`, `mountain boots`, etc., in the case of an outdoor equipment campaign) to identify potential customers by their Google searches (see Section 2.4 for more details).

presented in the related work of Lindquist and McKeague (2009). The method of Kneip et al. (2016) generalizes the original point of impact model by adding a classical functional linear regression component. While the original point of impact model captures only time-local effects, the augmented point of impact model of Kneip et al. (2016) allows also for time-global effects. In our paper we present a new and relevant case study where time-local as well as time-global effects are important for modeling the outcome.

As demonstrated in our simulation study, however, the finite sample performance of the estimation procedure proposed by Kneip et al. (2016) is very sensitive to the performance of the involved model selection. Therefore, we propose an adjusted sequential estimation algorithm that leads to significantly improved and more robust estimation results by using a refined model selection procedure.

The functional linear regression model with PoIs of Kneip et al. (2016) is related to several other works in the literature. Identifiability and estimation of points of impact was originally studied by McKeague and Sen (2010). The authors focus on a one-point of impact model without functional linear model component; however, the possibility of a partial model misspecification by an additional functional linear model component is also discussed theoretically. Ferraty et al. (2010) and Poß et al. (2020) allow for multiple PoIs within a nonparametric model, but both do not consider a functional linear model component. Matsui and Konishi (2011) consider the extraction of local information within functional linear regressions using a LASSO-type approach, but do not estimate global components. Torrecilla et al. (2016) focus on a classification context, and Fraiman et al. (2016) consider feature selection for functional data at a more general level. Our estimation algorithm uses the penalized smoothing splines estimator for functional linear regression models proposed by Crambes et al. (2009). The related literature is extensive and the following examples are by no means exhaustive. Cardot et al. (2007) consider functional linear regression with errors-in-variables, Crambes et al. (2009) address optimality issues, Goldsmith et al. (2010) focus on penalized smoothing splines within a mixed model framework, and Maronna and Yohai (2013) propose a robust version of the penalized smoothing splines estimator. Scalar-on-function regression models are successfully applied to solve important practical problems. Chiou (2012) proposes a functional regression model for predicting traffic flows. Goldsmith et al. (2012) introduce a penalized functional regression model to explore the relationship between cerebral white matter tracts in multiple-sclerosis patients. Koeppe et al. (2014) consider regularized functional linear regression for brain image data. Gellar et al. (2014) and Gromenko et al. (2017) propose functional regression models for incomplete curves. An overview article on methods for scalar-on-function regression is found in Reiss et al. (2017).

Readers with a general interest in Functional Data Analysis (FDA) are referred to the textbooks of Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), and Hsing and Eubank (2015). To the best of our knowledge, we are the first to use methods from FDA to analyze data from an online ad campaign; however, there are several contributions in FDA on related applications. Reddy and Dass (2006) use a classical functional linear regression model to analyze online art auctions, Liu and Müller (2008) analyze eBay auction prices using methods for sparse functional data, Wang et al. (2008) forecast eBay auction prices, Wang et al. (2008) develop a model for the price dynamics at eBay using differential equation models, and Zhang et al. (2010) consider real-time forecasting of eBay auctions using functional K-nearest neighbors.

The rest of the paper and our contributions are structured as follows. The next section (Section 2.2) contains our methodological part. In Section 2.2.1, we begin with a short presentation of the original procedure of Kneip et al. (2016). In Section 2.2.2, we introduce our three main proposals (1. Sequential model selection and estimation, 2. Smoothing splines estimator, and 3. Standardizations) which we use to stabilize and improve the estimation procedure of Kneip et al. (2016). The implementation of our estimation algorithm is presented in Section 2.2.3. Section 2.3 contains our simulation results, our case study on analyzing Google AdWords data is found in Section 2.4 and Section 2.5 concludes. Appendix 2.A presents further simulation results. The online supplement supporting this article contains the R-package `FunRegPoI` and the R-codes to reproduce our simulation study and the real data application.

2.2 Methodology

We formally consider the following functional linear regression model with PoIs introduced by Kneip et al. (2016):

$$Y_i = \int_a^b \beta(t) X_i(t)\,\mathrm{d}t + \sum_{s=1}^{S} \beta_s X_i(\tau_s) + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.1}$$

Here, $(Y_1, X_1), \ldots, (Y_n, X_n)$ denote an i.i.d. sample of scalar response variables $Y_i \in \mathbb{R}$ and random predictor functions $X_i \in L^2([a,b])$, where $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[X_i(t)] = 0$ for all $t \in [a,b]$. Without loss of generality, we set $[a,b] = [0,1]$. The i.i.d. error term ϵ_i has mean zero, variance $\mathbb{E}[\epsilon_i^2] = \sigma_\epsilon^2 < \infty$, and is independent of X_i. The assumption that Y_i and X_i have mean zero is only for notational simplicity; for the estimation, however, we will explicitly denote the centering of the data.

The function-valued slope parameter $\beta \in L^2([0,1])$ in Model (2.1) describes the time-global influences of X_i on Y_i. The scalar-valued slope parameters $\beta_s \in \mathbb{R}$ take into account the time-local influences where the corresponding (unknown) time-points τ_s denote the locations of the PoIs. The estimation algorithm described below addresses the estimation of all unknown model parameters, namely, the global slope coefficient β, the local influences of the PoIs $\beta_1, \ldots, \beta_S$, and the set of PoI locations $\mathcal{T} = \{\tau_1, \ldots, \tau_S\}$.

In the following, we introduce our basic notation. The functions $X_i(t)$ are observed at p equidistant grid points $t_1, \ldots, t_p$ with $t_j = (j-1)/(p-1)$. For non-equidistant designs, this can always be achieved by pre-smoothing the data. In $\mathbf{Y} = (Y_1, \ldots, Y_n)' \in \mathbb{R}^n$, we collect all observations of the response variable Y_i, and in $\mathbf{X} = (X_i(t_j))_{ij} \in \mathbb{R}^{n \times p}$, we collect all discretizations $X_i(t_j), i = 1, \ldots, n, j = 1, \ldots, p$. Furthermore, let $\mathbf{Y}^c$ and $\mathbf{X}^c$ define the centered versions of $\mathbf{Y}$ and $\mathbf{X}$, i.e., $\mathbf{Y}^c = (Y_1^c, \ldots, Y_n^c)'$, $\mathbf{X}^c = (X_i^c(t_j))_{ij}$, where $Y_i^c = Y_i - \overline{\mathbf{Y}}$, $X_i^c(t_j) = X_i(t_j) - \overline{\mathbf{X}}_j$, $\overline{\mathbf{Y}} = n^{-1} \sum_{i=1}^n Y_i$, $\overline{\mathbf{X}}_j = n^{-1} \sum_{i=1}^n X_i(t_j)$.

2.2.1 The original procedure of Kneip et al. (2016)

In this section, we briefly describe the estimation and model selection procedures proposed in Kneip et al. (2016). Afterwards, in Section 2.2.2, we describe our adjustments to improve the original procedure and explain why these adjustments result in superior estimation performances.

To estimate the potential PoIs $\widetilde{\tau}_s, s = 1, \ldots, \widetilde{S}$, Kneip et al. (2016) propose a local maxima search (over t_j) based on the sample version $|n^{-1} \sum_{i=1}^n Z_{X_i}(t_j; \delta) Y_i|$ of the cross-moment $|\,\mathbb{E}[Z_{X_i}(t;\delta) Y_i]|$, where $Z_{X_i}(t;\delta) = X_i(t) - (X_i(t-\delta) + X_i(t+$

Algorithm 1 Search Potential Points of Impact Algorithm

procedure SEARCHPOTPOI($\delta \in \mathcal{D} = (0, \delta_{\max}]$, $\mathbf{X} = \mathbf{X}^c$, $\mathbf{Y} = \mathbf{Y}^c$)
  Given δ, define the index $k_\delta \in \mathbb{N}$ such that $1 \leq k_\delta < (p-1)/2 \iff \delta \approx k_\delta/(p-1)$.
  Restrict the set of possible grid indices, i.e., define $\mathcal{J}_{0,\delta} = \{k_\delta + 1, \ldots, p - k_\delta\}$.
  For each index $j \in \mathcal{J}_{0,\delta}$, calculate $Z_{X_i}(t_j; \delta) = X_i(t_j) - \frac{1}{2}(X_i(t_j - \delta) + X_i(t_j + \delta))$.
  while $\mathcal{J}_{s,\delta} \neq \emptyset$, iterate over $s = 1, 2, 3, \ldots,$ and **do**
    Determine the index $j_s \in \mathcal{J}_{s-1,\delta}$ of the empirical maximum of $Z_X(t; \delta)Y$, i.e.,
$$j_s = \underset{j \in \mathcal{J}_{s-1,\delta}}{\operatorname{argmax}} \left| \frac{1}{n} \sum_{i=1}^{n} Z_{X_i}(t_j; \delta) Y_i \right|.$$
    Define the s-th potential impact point $\widetilde{\tau}_s = t_{j_s}$ as grid point at index j_s.
    Eliminate all points in an environment of size $\sqrt{\delta}$ around $\widetilde{\tau}_s$, i.e., define
$$\mathcal{J}_{s,\delta} = \{j \in \mathcal{J}_{s-1,\delta} \mid |t_j - \widetilde{\tau}_s| \geq \sqrt{\delta}/2\}.$$
  end while
  return $\widetilde{\mathcal{T}} = \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{S}}\}$
end procedure

$\delta))/2$ is the central second-order difference quotient of $X_i(t)$ with $\delta > 0$. The statistic $Z_{X_i}(t; \delta)$ acts as a filter on $X_i(t)$ that uncovers the local-specific variance component of the process $X_i(t)$; see the left plot in Figure 2.1.

The existence of a local-specific variance component in X_i is crucial for the estimation procedure of Kneip et al. (2016) and allows to show the identifiability of the points of impact and of the model parameters model parameters (see Kneip et al., 2016, Theorem 1). Processes that have a local-specific variance component are typically rough stochastic processes (for instance, Brownian motions, Ornstein-Uhlenbeck processes, etc.), i.e., processes with covariance functions that are sufficiently non-smooth at the diagonal (see Kneip et al., 2016, Theorem 3). Kneip et al. (2016) use a parameter $0 < \kappa < 2$ to quantify the smoothness of the covariance function at the diagonal and propose an estimator $\widehat{\kappa}$ to decide in practice, whether the covariance function is sufficiently non-smooth at the diagonal. The reader is referred to Section 2.4 for an application of this procedure.

The estimation procedure proposed by Kneip et al. (2016) to detect potential PoIs is formally described in Algorithm 1. In each iteration, $s = 1, 2, \ldots,$ one PoI is selected by the global maximum of the trajectory of $|n^{-1} \sum_{i=1}^{n} Z_{X_i}(t_j; \delta) Y_i|$ over $j \in \mathcal{J}_{s-1,\delta}$, where $\mathcal{J}_{s-1,\delta} \subset \{1, \ldots, p\}$ denotes an index set defined in Algorithm 1

(see lines 3 and 8). Once a PoI is selected, the algorithm eliminates the grid points within a $\sqrt{\delta}/2$-neighborhood around the selected PoI (see line 8 of Algorithm 1). The algorithm terminates when $\mathcal{J}_{s,\delta}$ is the empty set. The elimination step in line 8 is necessary for providing a consistent estimation procedure.

The selection of the first PoI is shown in the middle plot of Figure 2.1. The first elimination step is shown in the right plot of Figure 2.1, where the second PoI, $\widetilde{\tau}_2$, is determined by the global maximum of the remaining parts of the trajectory of $|n^{-1}\sum_{i=1}^{n} Z_{X_i}(t_j;\delta)Y_i|$ over $j \in \mathcal{J}_{1,\delta}$.

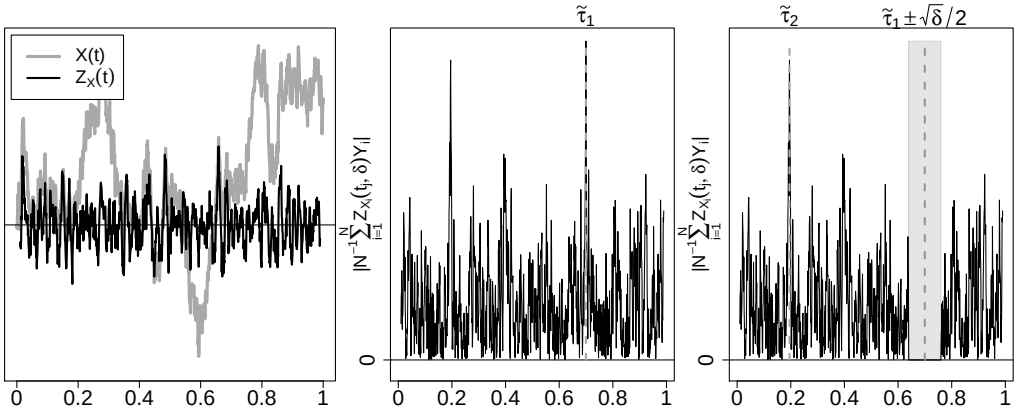


Figure 2.1: LEFT: Trajectories of $X_i(t_j)$ and $Z_{X_i}(t_j;\delta)$, with $\delta = 0.01$. MIDDLE: Trajectory of $|n^{-1}\sum_{i=1}^{n} Z_{X_i}(t_j;\delta)Y_i|$, with first choice $\widetilde{\tau}_1$. RIGHT: Visualization of the second iteration of the `searchPotPoI`-Algorithm.

To estimate the model coefficients for given PoIs $\widehat{\tau}_s$, Kneip et al. (2016) propose an FPCA-based estimation procedure using the approximate model $Y_i \approx \int_0^1 \beta_K(t)X_{i,K}(t)\,\mathrm{d}t + \sum_{s=1}^{\widehat{S}} \beta_s X_i(\widehat{\tau}_s) + \epsilon_i$, where $\beta_K(t) \approx \beta(t)$ and $X_{i,K}(t) \approx X_i(t)$ are K-dimensional approximations based on the first K eigenfunctions of the empirical covariance operator of X_i (see Kneip et al., 2016, Eq. (6.1)). Besides the smoothing parameter, K, one needs to choose a good value of the tuning parameter δ and a subset $\widehat{\mathcal{T}} \subseteq \widetilde{\mathcal{T}}$ of the set of potential PoIs $\widetilde{\mathcal{T}}$ from Algorithm 1. For selecting $\widehat{\mathcal{T}} \subseteq \widetilde{\mathcal{T}}$, Kneip et al. (2016) propose an asymptotic cut-off approach and data-driven Bayesian Information Criterion (BIC)-based approach. In this paper, we focus on the data-driven BIC-based approach as this approach performs clearly better than the asymptotic approach (see Table 1 in Kneip et al., 2016). Moreover, the asymptotic cut-off parameter is hardly applicable in practice as it depends on a generally unknown constant $A > \sqrt{2}$ (see Table 4 in Kneip et al., 2016).

Kneip et al. (2016) propose an infeasible version and a more general feasible version of their data-driven BIC-based procedure to select K, δ, and $\widehat{\mathcal{T}} \subseteq \widetilde{\mathcal{T}}$. The

infeasible strategy is used in their simulation study where the authors perform a BIC-based selection of K and $\widehat{\mathcal{T}}$, and set $\delta = 1/\sqrt{n}$. This naive parametrization of δ is appropriate in their simulation study, but can be arbitrarily bad in practice. The more general and strategy is used in the application section of Kneip et al. (2016) where the authors optimize the $\mathrm{BIC}(K, \widehat{\mathcal{T}}, \delta)$ *simultaneously* over K, subsets $\widehat{\mathcal{T}} \subset \widetilde{\mathcal{T}}$, and a fine grid of $\delta \in (0, \delta_{\max}]$. In this paper, we only focus on the latter general model selection strategy since this is the practically most relevant strategy proposed in Kneip et al. (2016) which is not based on unknown constants or naive choices of tuning parameters.

2.2.2 Improving the procedure of Kneip et al. (2016)

In this section, we explain our three main proposals to improve the estimation procedure of Kneip et al. (2016): 1. Sequential model selection and estimation, 2. Smoothing splines estimator, and 3. Standardizations. Afterwards, in Section 2.2.3, we describe the implementation our estimation algorithm which builds upon these proposals.

1. Sequential model selection and estimation. Estimating the model parameters in Model (2.1) bears the substantial risk of an omitted-variable-bias since not incorporating the (unknown) true PoI locations τ_s can result in a heavily biased estimator $\widehat{\beta}(t)$ (see the right plot in Figure 2.2 for noteworthy examples). This is a critical issue in practice, and our simulation results show that the original estimation procedure of Kneip et al. (2016) may suffer severely from such biases.

The underlying problem is that the selection of the number $\widehat{S}$ of PoIs and their locations $\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widehat{S}}$ and the selection of the smoothing parameter, K, for estimating $\beta(\cdot)$ are two ambiguous selection problems. It is easy to trade model complexities between the empirical PoI model component and the empirical functional model component without affecting the model fit. This results in a quite delicate model selection problem which generally leads to instable estimates when trying to solve both selection problems simultaneously as suggested in Kneip et al. (2016).

Let us explain the reason for this instability by considering the following two extreme situations—both approximating the regression Model (2.1):

- Let $K \gg 0$ and $\widehat{S} = 0$. For very large K the estimator $\widehat{\beta}_K(t)$ is flexible enough, such that

$$\int_0^1 \widehat{\beta}_K(t) X_{i,K}(t)\,\mathrm{d}t \approx \int_0^1 \beta(t) X_i(t)\,\mathrm{d}t + \sum_{s=1}^{S} \beta_s X_i(\tau_s).$$

In this case, $\widehat{\beta}_K(t)$ approximates $\beta(t)$, except at the points of impact locations $t = \tau_s$, where $\widehat{\beta}_K(t)$ approximates $\beta_s X_i(\tau_s)$, i.e., where $\int_{\tau_s-h}^{\tau_s+h} \widehat{\beta}_K(t) X_{i,K}(t)\,\mathrm{d}t \approx \beta_s X_i(\tau_s)$ with, e.g., $h = 0.01$ (see the right plot in Figure 2.2 for examples of such estimates $\widehat{\beta}_K(t)$).

- Let $K = 0$ and $\widehat{S} \gg 0$. A large set of PoI candidates $X_i(\widehat{\tau}_1), \ldots, X_i(\widehat{\tau}_{\widehat{S}})$ leads to a very flexible linear model, such that

$$\sum_{s=1}^{\widehat{S}} \widehat{\beta}_s X_i(\widehat{\tau}_s) \approx \int_0^1 \beta(t) X_i(t)\,\mathrm{d}t + \sum_{s=1}^{S} \beta_s X_i(\tau_s).$$

  In this case, $\sum_{s=1}^{\widehat{S}} \widehat{\beta}_s X_i(\widehat{\tau}_s)$ acts like a Riemann sum for approximating the integral $\int_0^1 \beta(t) X_i(t)\,\mathrm{d}t$, except for the $\widehat{\beta}_s$-values at $\widehat{\tau}_s \approx \tau_s$, where $\widehat{\beta}_s X_i(\widehat{\tau}_s) \approx \beta_s X_i(\tau_s)$.

These two extreme situations demonstrate that there is a certain ambiguity between the model selection parameters K and $\widehat{S} = |\widehat{\mathcal{T}}(\delta)|$ that allows for shifting the model-complexities between the integral-part and the PoI-part of the empirical model. This ambiguity generally leads to unstable model selections when optimizing $\mathrm{BIC}(K, \widehat{\mathcal{T}}, \delta)$ *simultaneously* over K, subsets $\widehat{\mathcal{T}} \subset \widetilde{\mathcal{T}}$, and δ—as proposed in Kneip et al. (2016). As a consequence, one gets instable estimates of $\beta(\cdot)$ caused by omitted-variable biases in $\widehat{\beta}(\cdot)$, as shown in the right plot in Figure 2.2.

To stabilize the model selection procedure we propose a sequential selection and estimation procedure (see Section 2.2.3). In the first ("Pre-select") step, our procedure pre-selects all potential points of impact $\widetilde{\mathcal{T}} = \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{S}}\}$ while ignoring the estimation of the functional parameter $\beta(\cdot)$. In the second ("Estimate") step, our procedure estimates the model parameters, $\beta(\cdot)$ and β_s, given the pre-selected points of impact.

The theoretical justification for this sequential approach is given by the following result which holds under the assumptions of Kneip et al. (2016) and implies that the points of impact can be estimated consistently without knowledge (or pre-estimation) of the slope function $\beta(\cdot)$ (see Lemmas 3 and 4 in the supplemenary paper of Kneip et al., 2016):

$$\begin{array}{ll} |\mathbb{E}(Z_{\delta,i}(t) Y_i)| = \beta_r c(\tau_s) \delta^\kappa + o(\delta^\kappa) & \text{if } \; t_j = \tau_s \text{ for some } s = 1, \ldots, S \\ |\mathbb{E}(Z_{\delta,i}(t) Y_i)| = O(\delta^2) & \text{if } \; t \notin \{\tau_1, \ldots, \tau_S\} \end{array} \tag{2.2}$$

as $\delta \to 0$, where $0 < \kappa < 2$ and $0 < c(\tau_r) < \infty$ are constants specific to the considered process X_i.

That is, the trajectory of $|\mathbb{E}(Z_{\delta,i}(t_j))Y_i)|$, $j = 1, \ldots, p$, will have peaks at grid points $t_j \approx \tau_r$, even without knowledge (or pre-estimation) of the slope function $\beta(\cdot)$. Consequently, Step 1 ("Pre-select") of our algorithm (Section 2.2.3) leads to a consistent point of impact selection if, for instance, $\delta^\kappa \sim n^{-1}$, since $|\mathbb{E}(Z_{\delta,i}(t_j))Y_i)|$ can be consistently estimated using its empirical counterpart $|n^{-1}\sum_{i=1}^n Z_{\delta,i}(t_j))Y_i|$ for all $j = 1, \ldots, p$ as $n \to \infty$.

Using the consistently pre-selected points of impact in Step 2 ("Estimate") leads to a more stable estimation of the model parameters, $\beta(\cdot)$ and β_s, as it avoids a simultaneous selection of the PoIs and the smoothing parameters. The further Steps 3-5 (see the overview in Section 2.2.3) of our selection and estimation algorithm in Section contain repetitions of the selection and estimation steps (Step 1 and Step 2). These repetitions are asymptotically irrelevant, but further improve the estimation results in practice (see Section 2.3).

2. Smoothing splines estimator. Deviating from Kneip et al. (2016), we propose using a penalized smoothing splines estimator. The FPCA-based estimator, proposed by Kneip et al. (2016), is optimal only under the restrictive assumption of a structural link between the functional regression parameter, $\beta(\cdot)$, and the functional regressor, X (see Assumptions (3.1-3.3) in Hall and Horowitz, 2007). However, this link does not necessarily hold in applications and also cannot be tested in practice.[2] Therefore, we propose using the penalized smoothing splines estimator of Crambes et al. (2009). While this estimator achieves minimax optimal rates under similar structural link assumptions (see Crambes et al. , 2009), it is also known to perform well if these structural assumptions do not hold since the spline basis system has some very general approximation properties (see, for instance, De Boor, 1978, and Crambes et al., 2009).

3. Standardizations. The standardization of the curves in Step 1 ("Pre-select") and Step 3 ("Sub-select") of our algorithm scales the trajectories of the process Z_δ by the inverse of the pointwise standard deviation of the process X. From an asymptotic perspective, this is irrelevant, since this scaling only leads to different constants $c(\tau_r)$ in (2.2). However, standardization of the data is a typical pre-processing step in model selection problems leading to more homogenous signals which further improves the selection results in practice (see Section 2.3).

2.2.3 The PES-ES estimation algorithm

Our estimation algorithm is built up from the following three Pre-select, Estimate, and Sub-select (PES) steps:

[2] Remember that the FPCA-basis, based on the eigendecomposition of the covariance operator of X, is the optimal empirical basis to approximate X, but generally not the optimal basis to approximate $\beta(\cdot)$.

1. **Pre-select:** Pre-select potential PoIs $\widetilde{\mathcal{T}} = \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{S}}\}$. (See Section 2.2.3)
2. **Estimate:** Estimate the function- and scalar-valued slope parameters $\beta, \beta_1, \ldots, \beta_{\widetilde{S}}$ given the set of potential PoIs $\widetilde{\mathcal{T}}$. (See Section 2.2.3)
3. **Sub-select:** Sub-select PoIs from the set of potential PoIs $\widetilde{\mathcal{T}}$. (See Section 2.2.3)

Typically, the estimation step (Step 2) leads to inefficient estimators $\widehat{\beta}(\cdot)$, but avoids omitted-variable biases. Inefficient, because $\widetilde{\mathcal{T}}$ tends to contain many redundant PoI locations ($\widetilde{S} > S$), which reduces the number of degrees of freedom. We reduce the risk of omitted-variable biases, because the large set of potential PoIs $\widetilde{\mathcal{T}}$ has a high likelihood of containing the true PoI locations (as explained in more detail in Section 2.2.2). Our final PES-ES algorithm, described in Section 2.2.3, uses a repetition of the latter two Estimate-Sub-select (ES) steps, which can result in a further improvement of the estimation results (see Section 2.3).

Pre-Select PoIs

To select potential PoIs, we use Algorithm 1 with the difference that instead of using centered observations of the functions $\mathbf{X}^c$, we use the pointwise standardized curves $\mathbf{X}^{st}$ as input of the algorithm, where $X_i^{st}(t_j) = X_i^c(t_j)/\mathrm{sd}(\mathbf{X}_j)$ and $\mathrm{sd}(\mathbf{X}_j) = (n^{-1}\sum_{i=1}^n (X_i(t_j) - \overline{\mathbf{X}}_j)^2)^{1/2}$. As described in Section 2.2.2, this is irrelevant from an asymptotic point of view, but typically stabilizes and improves the PoI selection in practice.

Estimate Slope Parameters

To estimate the slope parameters—given the pre-selected PoIs $\widetilde{\mathcal{T}}$—we adapt the penalized smoothing splines estimator proposed by Crambes et al. (2009) in order to incorporate PoIs. Let us initially recap the situation of Model (2.1) without PoIs ($S = 0$, $\mathcal{T} = \emptyset$), as considered by Crambes et al. (2009). Their estimator of $\beta(\cdot)$, evaluated at the grid points $t_1, \ldots, t_p$, is given by

$$\left(\widehat{\beta}^{\rho}(t_1), \ldots, \widehat{\beta}^{\rho}(t_p)\right) = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}^{c\prime}\mathbf{X}^c + \rho\mathbf{A}\right)^{-1}\mathbf{X}^{c\prime}\mathbf{Y}^c, \tag{2.3}$$

where the penalty matrix $\mathbf{A} = \mathbf{P} + p\mathbf{A}^\star$ is composed of a non-classical projection matrix $\mathbf{P}$ and a classical regularization matrix $\mathbf{A}^\star$. The non-classical $p \times p$ projection matrix $\mathbf{P} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$, with $\mathbf{W} = (t_j^l)_{j,l} \in \mathbb{R}^{p\times m}$ is introduced by Crambes et al. (2009) in order to guarantee uniqueness of their estimator, where t_j^l denotes the lth power of the grid point t_j with $j = 1, \ldots, p$ and $l = 0, \ldots, m-1$.

Following the usual convention, we set $m = 2$, which results in the classical choice of *cubic* splines. The classical $p \times p$ regularization matrix $\mathbf{A}^\star$ is defined as

$$\mathbf{A}^\star = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\left(\int_0^1 \mathbf{b}^{(2)}(t)\mathbf{b}^{(2)}(t)'\,\mathrm{d}t\right)(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}',$$

where $\mathbf{b}(t) = (b_1(t), \ldots, b_p(t))'$ are natural cubic spline basis functions, $\mathbf{b}^{(2)}(t)$ denotes their second derivatives, and $\mathbf{B}$ is a $p \times p$ matrix with elements $b_i(t_j), i, j = 1, \ldots, p$. For the implementation of the natural cubic spline basis functions, we use the `ns`-function contained in the R-package `splines`.

In order to incorporate the pre-selected PoIs, we need to extend the matrices $\mathbf{X}^c$ and $\mathbf{A}$. The extended data matrix is given by $\mathbf{X}^c_{\widetilde{\mathcal{T}}} = (\mathbf{X}^c, p\mathbf{X}^c(\widetilde{\tau}_1), \ldots, p\mathbf{X}^c(\widetilde{\tau}_{\widetilde{S}})) \in \mathbb{R}^{n \times (p+\widetilde{S})}$, where $\mathbf{X}^c(\widetilde{\tau}_s) = (X_1^c(\widetilde{\tau}_s), \ldots, X_n^c(\widetilde{\tau}_s))' \in \mathbb{R}^n$. The extended penalty matrix is given by

$$\mathbf{A}_{\widetilde{\mathcal{T}}} = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{(p+\widetilde{S}) \times (p+\widetilde{S})},$$

where all entries with respect to the PoIs are zero (see Goldsmith et al., 2010, for an equivalent extension of the penalty matrix). The augmented estimator of $\beta(t_1), \ldots, \beta(t_p)$ and $\beta_1, \ldots, \beta_S$,

$$\widehat{\beta}^\rho_{\widetilde{\mathcal{T}}} = \left(\widehat{\beta}^\rho_{\widetilde{\mathcal{T}}}(t_1), \ldots, \widehat{\beta}^\rho_{\widetilde{\mathcal{T}}}(t_p), \widehat{\beta}^\rho_{\widetilde{\mathcal{T}},1}, \ldots, \widehat{\beta}^\rho_{\widetilde{\mathcal{T}},\widetilde{S}}\right) = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}^{c\prime}_{\widetilde{\mathcal{T}}}\mathbf{X}^c_{\widetilde{\mathcal{T}}} + \rho\mathbf{A}_{\widetilde{\mathcal{T}}}\right)^{-1}\mathbf{X}^{c\prime}_{\widetilde{\mathcal{T}}}\mathbf{Y}^c, \tag{2.4}$$

depends on the included set of PoIs $\widetilde{\mathcal{T}}$ and on the smoothing parameter ρ. In order to determine an optimal smoothing parameter, we use the following Generalized Cross-Validation (GCV) criterion, as proposed by Crambes et al. (2009):

$$\mathrm{GCV}(\rho) = \frac{\frac{1}{n}\mathrm{RSS}(\widehat{\beta}^\rho_{\widetilde{\mathcal{T}}})}{\left(1 - \frac{1}{n}\mathrm{Tr}(\mathbf{H}^c_{\rho,\widetilde{\mathcal{T}}})\right)^2}. \tag{2.5}$$

Here, the Residual Sum of Squares (RSS) is defined as $\mathrm{RSS}(\widehat{\beta}^\rho_{\widetilde{\mathcal{T}}}) = ||\mathbf{Y}^c - \mathbf{H}^c_{\rho,\widetilde{\mathcal{T}}}\mathbf{Y}^c||^2$, where $||.||$ denotes the Euclidean norm, and the smoother matrix $\mathbf{H}^c_{\rho,\widetilde{\mathcal{T}}}$ is defined as $\mathbf{H}^c_{\rho,\widetilde{\mathcal{T}}} = (np)^{-1}\mathbf{X}^c_{\widetilde{\mathcal{T}}}((np)^{-1}\mathbf{X}^{c\prime}_{\widetilde{\mathcal{T}}}\mathbf{X}^c_{\widetilde{\mathcal{T}}} + \rho\mathbf{A}_{\widetilde{\mathcal{T}}})^{-1}\mathbf{X}^{c\prime}_{\widetilde{\mathcal{T}}}$. Our final estimator for the slope parameters is given by the GCV-optimized version of (2.4),

$$\widehat{\beta}_{\widetilde{\mathcal{T}}} = \left(\widehat{\beta}_{\widetilde{\mathcal{T}}}(t), \widehat{\beta}_{\widetilde{\mathcal{T}},1}, \ldots, \widehat{\beta}_{\widetilde{\mathcal{T}},\widetilde{S}}\right) = \left(\widehat{\beta}^{\rho_{\mathrm{GCV}}}_{\widetilde{\mathcal{T}}}(t), \widehat{\beta}^{\rho_{\mathrm{GCV}}}_{\widetilde{\mathcal{T}},1}, \ldots, \widehat{\beta}^{\rho_{\mathrm{GCV}}}_{\widetilde{\mathcal{T}},\widetilde{S}}\right), \quad t \in \{t_1, \ldots, t_p\}, \tag{2.6}$$

where $\rho_{\mathrm{GCV}} = \mathrm{argmin}_{\rho\in(0,\rho_{\max}]}\mathrm{GCV}(\rho)$.

Sub-Select PoIs

This part of our estimation algorithm is aimed at selecting the true PoIs from the pre-selected set of potential PoIs $\widetilde{\mathcal{T}} = \widetilde{\mathcal{T}}(\delta)$ given the estimate $\widehat{\beta}_{\widetilde{\mathcal{T}}}$ in (2.6). This sub-selection is performed by minimizing the following BIC over subsets $\mathcal{R} \subseteq \widetilde{\mathcal{T}}(\delta)$:

$$\widehat{\mathcal{T}} = \underset{\mathcal{R}\subseteq\widetilde{\mathcal{T}}(\delta)}{\mathrm{argmin}}\,\mathrm{BIC}(\mathcal{R}), \quad \text{where}$$

$$\mathrm{BIC}(\mathcal{R}) = n\log\left(\frac{\mathrm{RSS}(\mathcal{R})}{n}\right) + \log(n)\cdot S_{\mathcal{R}}, \quad \text{with} \quad S_{\mathcal{R}} = |\mathcal{R}|. \tag{2.7}$$

Here, $\mathrm{RSS}(\mathcal{R})$ is made up of the residuals from regressing the $\widehat{\beta}_{\widetilde{\mathcal{T}}}$-neutralized $Y_{i,\widehat{\beta}_{\widetilde{\mathcal{T}}}} = Y_i^c - \int_0^1 \widehat{\beta}_{\widetilde{\mathcal{T}}}(t)X_i^c(t)\,\mathrm{d}t$ onto $X_i^{st}(\widetilde{\tau}_s), \ldots, X_i^{st}(\widetilde{\tau}_{S_{\mathcal{R}}})$, with $\{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{S_{\mathcal{R}}}\} = \mathcal{R}$, where $\widehat{\beta}_{\widetilde{\mathcal{T}}}(t)$ is the estimate of $\beta(\cdot)$ defined as the first element in the vector of estimates (2.6).

For optimizing $\mathrm{BIC}(\mathcal{R})$ over $\mathcal{R} \subseteq \widetilde{\mathcal{T}}(\delta)$, we use a directed search strategy taking into account the information content in $\widetilde{\mathcal{T}} = \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{S}}\}$. By construction, the order of the PoI locations $\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{S}}$ reflects a decreasing signal-to-noise ratio and, therefore, a decreasing quality of the estimates. This suggests minimizing $\mathrm{BIC}(\mathcal{R})$ using a directed search strategy where $\mathrm{BIC}(\mathcal{R})$ is evaluated consecutively at the sets $\mathcal{R} = \{\widetilde{\tau}_1\}$, $\mathcal{R} = \{\widetilde{\tau}_1, \widetilde{\tau}_2\}, \ldots, \mathcal{R} = \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widetilde{S}}\}$.

The full PES-ES estimator

Our estimation algorithm, PES-ES, consists of the above described Pre-select-Estimate-Sub-select (PES) steps and uses a repetition of the latter two Estimate-Sub-select (ES) steps:

1. **P**re-Select	$\widetilde{\mathcal{T}} = \widetilde{\mathcal{T}}(\delta)$	(Section 2.2.3)
2. **E**stimate	$\widehat{\beta}_{\widetilde{\mathcal{T}}}$	(Section 2.2.3)
3. **S**ub-Select	$\widehat{\mathcal{T}} \subseteq \widetilde{\mathcal{T}}$	(Section 2.2.3)
4. re**E**stimate	$\widehat{\beta}_{\widehat{\mathcal{T}}}$	(Section 2.2.3, with $\widetilde{\mathcal{T}}$ replaced by $\widehat{\mathcal{T}}$)
5. re**S**ub-Select	$\widehat{\mathcal{T}}_{\mathrm{re}} \subseteq \widehat{\mathcal{T}}$	(Section 2.2.3, with $\widetilde{\mathcal{T}}$ replaced by $\widehat{\mathcal{T}}$)

Note that the entire PES-ES algorithm depends on the initially pre-selected set of potential PoIs $\widetilde{\mathcal{T}}(\delta)$ and, therefore, on the choice of δ. In the following, we write $\widehat{\mathcal{T}}_{\mathrm{re}}(\delta)$ in order to emphasize this entire dependency on δ. We follow Kneip et al. (2016) and determine an optimal δ by minimizing the BIC. For each δ-value on a fine grid in $(0, \delta_{\max}]$, we run the entire PES-ES algorithm and select

the optimal δ by,

$$\delta_{\text{BIC}} = \underset{\delta \in (0, \delta_{\max}]}{\operatorname{argmin}} \ \text{BIC}(\delta), \quad \text{with}$$

$$\text{BIC}(\delta) = n \log \left(\frac{\text{RSS}(\widehat{\mathcal{T}}_{\text{re}}(\delta))}{n} \right) + \log(n) \cdot \text{edf}(\widehat{\mathcal{T}}_{\text{re}}(\delta)), \tag{2.8}$$

where $\text{RSS}(\widehat{\mathcal{T}}_{\text{re}}(\delta)) = ||\mathbf{Y}^c - \mathbf{H}^c_{\rho_{\text{GCV}}, \widehat{\mathcal{T}}_{\text{re}}(\delta)} \mathbf{Y}^c||^2$ with smoother matrix $\mathbf{H}^c_{\rho_{\text{GCV}}, \widehat{\mathcal{T}}_{\text{re}}(\delta)}$ defined as $\mathbf{H}^c_{\rho_{\text{GCV}}, \widehat{\mathcal{T}}_{\text{re}}(\delta)} = (np)^{-1} \mathbf{X}^c_{\widehat{\mathcal{T}}_{\text{re}}(\delta)} ((np)^{-1} \mathbf{X}^{c\prime}_{\widehat{\mathcal{T}}_{\text{re}}(\delta)} \mathbf{X}^c_{\widehat{\mathcal{T}}_{\text{re}}(\delta)} + \rho_{\text{GCV}} \mathbf{A})^{-1} \mathbf{X}^{c\prime}_{\widehat{\mathcal{T}}_{\text{re}}(\delta)}$ and effective degrees of freedom $\text{edf}(\widehat{\mathcal{T}}_{\text{re}}(\delta)) = \text{Tr}(\mathbf{H}^{c\prime}_{\rho_{\text{GCV}}, \widehat{\mathcal{T}}_{\text{re}}(\delta)} \mathbf{H}^c_{\rho_{\text{GCV}}, \widehat{\mathcal{T}}_{\text{re}}(\delta)})$; see Hastie and Tibshirani (1990), Ch. 3.5 for an overview of possible definitions of edf.

2.3 Simulations

In the following simulation study, we assess the finite sample properties of our PES-ES algorithm. The original estimation procedure proposed by Kneip, Poss, and Sarda (2016), abbreviated as KPS, serves as our main benchmark, and its implementation is described in Section 2.2.1. The smoothing splines estimator (2.3) by Crambes, Kneip, and Sarda (2009), abbreviated hereafter as CKS, serves as a challenging benchmark for our NoPoI data generating process (i.e., a functional linear regression model *without* points of impact). Section 2.3.1 introduces the considered data generating processes and presents our simulation results.

We aim to provide an in-depth assessment of our PES-ES estimation algorithm. Therefore, in order to assess the improvements that are due to the final ES (Estimation and Subselection) step, we compare the PES-ES results with those of the reduced PES estimation algorithm without the final ES step. We also show that an additional second repetition of the ES step (PES-2ES) does not lead to a significant improvement of our estimation algorithm.

Kneip et al. (2016) arbitrarily set $K_{\max} = 6$, which is, however, too small for our simulation study where $K_{\max} = 6$ often becomes a binding upper optimization threshold. The choice of $K_{\max}$ is crucial since it constrains the magnitude of possible omitted-variable biases in $\widehat{\beta}_K(t)$. The same issue applies to $\rho_{\min}$ when optimizing the GCV in (2.5) over $\rho \in [\rho_{\min}, \rho_{\max}]$ with $\rho_{\min} \approx 0$. Therefore, we choose very conservative optimization intervals $[K_{\min}, K_{\max}] = [1, 150]$ and $[\rho_{\min}, \rho_{\max}] = [10^{-6}, 200]$.

2.3.1 Data Generating Processes and Simulation Results

We consider five different Data Generating Processes (DGPs), as described in Table 2.1. The DGPs Easy and Complicated represent a simple and a more complex version of Model (2.1). The Complicated DGP is challenging due to the closeness of the PoI locations τ_1 and τ_2, which may trigger omitted-variable biases in $\widehat{\beta}(t)$ when omitting either τ_1 or τ_2. The two further DGPs NoPoI ($\mathcal{T} = \emptyset$) and OnlyPoI ($\beta(t) \equiv 0$) are used to check the robustness of our PES-ES algorithm against model-misspecifications.

Table 2.1: Data Generating Processes.

DGP	$\beta(t)$	S	$\mathcal{T} = \{\tau_1, \ldots, \tau_S\}$	$\{\beta_1, \ldots, \beta_S\}$
Easy	$\beta(t) = -(t-1)^2 + 2$	2	$\{0.3, 0.6\}$	$\{-3, 3\}$
Complicated	$\beta(t) = -5(t-0.5)^3 - t + 1$	3	$\{0.3, 0.4, 0.6\}$	$\{-3, 3, 3\}$
OnlyPoI	$\beta(t) \equiv 0$	2	$\{0.3, 0.6\}$	$\{-3, 3\}$
NoPoI	$\beta(t) = -(t-1)^2 + 2$	0	$\emptyset$	$\emptyset$

For each DGP and two sample sizes ($n = 250$ and $n = 500$), we generate 1000 replications of n functions $X_i(t)$ observed at $p = 300$ equidistant points $t_1, \ldots, t_p$ in $[0, 1]$. In Appendix 2.A we additionally present simulation results for $p = 500$. The functions $X_i(t)$ are standard Brownian Motions, and the dependent variables Y_i are generated according to Model (2.1) with $\epsilon_i \sim N(0, 0.125^2)$. Our simulation is implemented using the statistical language R (R Core Team, 2017), and the R-codes for reproducing the simulation results are part of the online supplement supporting this article.

The upper panel of Table 2.2 reports the integrated squared bias and the integrated variance for the estimator $\widehat{\beta}(t)$ of $\beta(t)$. The integrated squared bias is computed as $\int_0^1 (\bar{\beta}(t) - \beta(t))^2 \, \mathrm{d}t$, where $\bar{\beta}(t) = 1000^{-1} \sum_{r=1}^{1000} \widehat{\beta}_r(t)$ is the mean of the estimates over all replications. The integrated variance is computed as $1000^{-1} \int_0^1 \sum_{r=1}^{1000} (\widehat{\beta}_r(t) - \bar{\beta}(t))^2 \, \mathrm{d}t$. The lower panel of Table 2.2 reports the average squared bias $S^{-1} \sum_{s=1}^{S} (\bar{\beta}_s - \beta_s)^2$, with $\bar{\beta}_s = 1000^{-1} \sum_{r=1}^{1000} \widehat{\beta}_{r,s}$, and the average variance $S^{-1} \sum_{s=1}^{S} 1000^{-1} \sum_{r=1}^{1000} (\widehat{\beta}_{r,s} - \bar{\beta}_s)^2$ for the PoI coefficient estimators $\widehat{\beta}_s$, conditionally on the event that τ_s was correctly found[3], where a single τ_s is considered to be found correctly if $|\widehat{\tau}_s - \tau_s| < 0.01$. The latter requirement corresponds to an estimation precision of only ± 3 grid points, which is substantially more challenging than the matching requirement originally used in Kneip et al. (2016). The shades of gray in Table 2.2 show the ranking of the mean squared error (MSE); the lowest/highest MSE (=Bias^2 + Var.) has the

[3]Note that it is impossible to compute estimation errors for non-found τ_s.

Table 2.2: Squared bias and variance of the estimators. Shades of gray show the ranking of the Mean Squared Error (MSE): lowest/highest MSE has the darkest/lightest gray-scale.

		Easy		Complicated		NoPoI		OnlyPoI	
$\int \widehat{\beta}(t)$		Bias^2	Var.	Bias^2	Var.	Bias^2	Var.	Bias^2	Var.
$n = 250$	PES	0.02	0.22	0.21	1.98	0.00	0.02	0.00	0.08
	PES-ES	0.02	0.24	0.16	1.68	0.00	0.01	0.00	0.06
	PES-2ES	0.02	0.25	0.16	1.66	0.00	0.01	0.00	0.06
	KPS	2.81	51.17	155.17	303.03	0.01	0.02	0.05	6.65
	CKS	-	-	-	-	0.00	0.00	-	-
$n = 500$	PES	0.01	0.06	0.05	0.55	0.00	0.01	0.00	0.01
	PES-ES	0.00	0.05	0.04	0.38	0.00	0.01	0.00	0.01
	PES-2ES	0.00	0.05	0.04	0.38	0.00	0.01	0.00	0.01
	KPS	0.35	16.69	91.32	245.88	0.01	0.01	0.00	0.5
	CKS	-	-	-	-	0.00	0.00	-	-
$\frac{1}{S}\sum \widehat{\beta}_s$									
$n = 250$	PES	0.01	0.1	0.01	0.09	-	-	0.00	0.02
	PES-ES	0.00	0.08	0.01	0.06	-	-	0.00	0.02
	PES-2ES	0.00	0.08	0.01	0.06	-	-	0.00	0.02
	KPS	0.03	0.54	1.59	4	-	-	0.00	0.06
$n = 500$	PES	0.00	0.02	0.00	0.01	-	-	0.00	0.02
	PES-ES	0.00	0.01	0.00	0.01	-	-	0.00	0.00
	PES-2ES	0.00	0.01	0.00	0.01	-	-	0.00	0.00
	KPS	0.01	0.2	0.78	2.92	-	-	0.00	0.01

darkest/lightest gray-scale.

The simulation results for the slope parameters $\beta(t)$ and $\beta_1, \ldots, \beta_S$ in the upper and lower panel of Table 2.2 show that the smoothing-spline-based estimation algorithms PES and PES-ES clearly outperform the FPCA-based KPS estimator. The final ES-step in the PES-ES algorithm aims to remove further falsely selected point of impact candidates. This is advantageous in all DGPs, except for the Easy DGP with $n = 250$ and the NoPoI DGP, where PES-ES and PES achieve essentially equivalent results. Note that the final ES-step is particularly advantageous for the Complicated DGP and the smaller sample size

Table 2.3: Percentage of replications with correct detection of all points of impact $\tau_1, \ldots, \tau_S$.

		300 grid points			500 grid points		
		Easy	Compl.	OnlyPoI	Easy	Compl.	OnlyPoI
$n = 250$	PES	97.5	77.4	99	97	83.8	99.1
	PES-ES	97.6	79.3	99.2	97.3	85.3	99.2
	PES-2ES	97.6	79.3	99.2	97.3	85.8	99.2
	KPS	89.7	19.3	98.7	89.5	24	98.5
$n = 500$	PES	99.3	94.6	99.9	99.3	94	99.9
	PES-ES	99.4	95.7	99.9	99.2	95.3	99.9
	PES-2ES	99.4	95.8	99.9	99.2	95.3	99.9
	KPS	96.9	37.2	100	97	41.9	99.5

$n = 250$, where KPS shows a very poor performance. Only in this particular case, one additional second repetition of the ES-step (PES-2ES) further reduces the variance. This improvement, however, is not substantial and does not justify the additionally involved computational burden of PES-2ES. PES-ES also performs very well in the NoPoI and the OnlyPoI DGPs, where PES-ES is actually a misspecified estimation procedure. In the case of NoPoI, PES-ES performs almost as well as the corresponding (minimax-optimal) benchmark-estimator CKS, and in the case of OnlyPoI, PES-ES is the best performing method.

Table 2.3 reports for each estimator and sample size the percentage of replications where all PoI locations $\tau_1, \ldots, \tau_S$ are found correctly. The left part of the table contains the results for functions observed on $p = 300$ grid points and the right part for $p = 500$ grid points. PES-ES and PES-2ES outperform all competitors, except in the case of OnlyPoI with $n = 500$, where all estimation procedures show essentially the same performance. Again, the difference between PES(-(2)ES) and KPS is particularly large for the smaller sample size $n = 250$ and the Complicated DGP. Increasing the resolution of the grid from $p = 300$ to $p = 500$ does not change the results. Similarly, the increased resolution also does not affect the precision of the estimate for the slope parameter $\beta(\cdot)$ and $\beta_1, \ldots, \beta_s$, see Table 2.A.2 in Appendix 2.A.

To show the performance boost of using standardized data for locating the potential PoIs (as described at the end of Section 2.2.3), we report the simulation results without standardizing the data (see Tables 2.A.1 and 2.A.3 in Appendix 2.A). The results show that the standardization of the data is beneficial for the Complicated DGP. Table 2.A.4 in Appendix 2.A shows the simulation results

for the Complicated DGP, but with different noise-to-signal ratios, that is, with different values for the error variance in model (2.1). PES(-ES) still outperforms KPS significantly; however, it turns out that the difference becomes less pronounced as the noise-to-signal ratio increases.

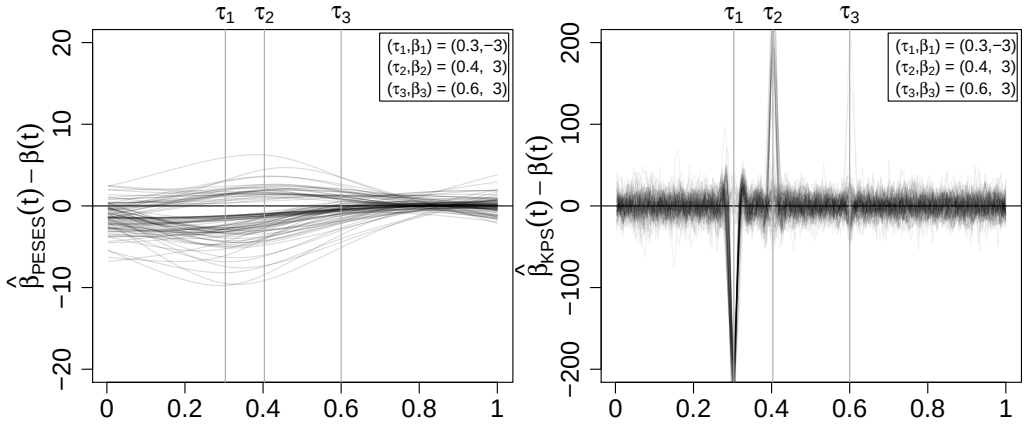


Figure 2.2: Pointwise deviations $\widehat{\beta}(t) - \beta(t)$ of the 10% largest L^2 distances $\int_0^1 (\widehat{\beta}(t) - \beta(t))^2 \, \mathrm{d}t$ for the Complicated DGP. Note that the scales of the two y-axes differ by a factor of 10.

2.4 Application

To illustrate the practical importance of the functional linear regression model with points of impact, we present an application to data from Google AdWords, which is the most popular online advertising platform and of fundamental importance for Alphabet's (Google's parent company) economic success (in 2014, 90 percent of Alphabet's sales came from AdWords). Online advertising, in turn, is the most important branch of today's advertising industry, with an expected U.S. revenue of 60 billion USD in 2016 (Doty et al., 2016). The case study described below is motivated by the needs of *Crealytics*, the company that generously provided the data. Today this company uses the described method—with some further confidential enhancements—to support their daily business.

The main pricing mechanism at Google AdWords is the so-called Pay-Per-Click (PPC) mechanism. Here, advertisers (e.g., an online outdoor shop in our application) can bid for a sponsored "impression" to be displayed along with Google's search results when a user conducts a search query related to a specific keyword (e.g., `outdoor jacket`)[4]. The basic building block of an online ad

[4]Sponsored impressions link to the advertised homepage—they are similar to, but distin-

campaign is a text corpus of (hundreds, thousands, or ten-thousands of, etc.) keywords related to the advertised products.

The limited number of sponsored impressions is allocated by an auction. Advertisers whose impression appears on the display are chosen according to their ad-rank, which is basically their original bid, i.e., the maximum "costs-per-click" an advertiser is willing to pay times the quality score, a discrete metric (from 1, the lowest, to 10, the best) determining the relevance of an advertiser's impression. Google AdWords auctions are time continuous and an advertiser only pays if a user clicks on the displayed impression. (See Geddes, 2014, for an in-depth introduction to Google AdWords.)

The bidding process is usually based on bidding softwares that evaluate specific key-figures. One of the most important key-figures is the so-called Click-Though Rate (CTR), which is defined as the daily number of clicks per impression. The CTR estimates the current probability of receiving a click on a sponsored impression and therefore plays an important role in assisting the bidding process on a short-term basis (Geddes, 2014).

The economic success of ad campaigns, however, also depends on long-sighted bidding strategies taking into account product specific (time-global) seasonalities as well as (time-local) events, such as the importance of Valentine's Day for an online flower shop. Unfortunately, existing key-figures such as the CTR only provide a daily perspective and are not suitable for assisting in the implementation of long-sighted bidding strategies. Therefore, the functional linear regression model with points of impact is a suitable methodology to identify the (global and local) functional relationship between the *yearly* clicks and the *yearly* trajectories of daily impressions—leading to a long-sighted version of the CTR.

As a yearly measure of clicks, we use the logarithmized yearly sums of clicks, i.e., $Y_i = \log(C_i)$ with $C_i := \sum_{t=1}^{365} \texttt{clicks}_{it}$, where i indexes the ith keyword of the considered ad campaign. As a yearly measure of impressions, we use the yearly trajectories of daily logarithmized numbers of impressions, i.e., $X_i(t) = \log(\mathcal{I}_i(t))$ with $\mathcal{I}_i(t) := \texttt{impressions}_{it}$, where $t = 1, \ldots, 365$ indexes the days of the considered year. Our application uses data from a real Google AdWords campaign run by an online store selling outdoor equipment in the year from April 1st, 2012 to March 31st, 2013. The left plot in Figure 2.3 shows all trajectories $X_i(t)$ of the considered ad campaign. The middle plot shows three exemplary (logarithmized) impression trajectories $X_i(t)$. The right panel shows the (logarithmized) yearly sum of clicks Y_i, received on the impressions of the ith keyword.

The data are provided by *Crealytics* (www.crealytics.com), an online advertising service provider with offices in Berlin (Germany), London (UK), and New

guishable from ordinary Google search results.

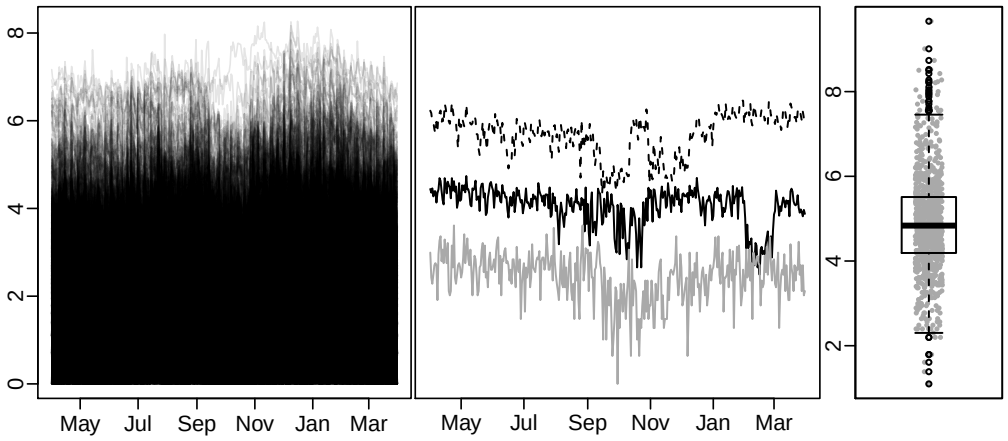


Figure 2.3: LEFT: Yearly trajectories of daily logarithmized numbers of impressions. MIDDLE: Three exemplary trajectories $X_i(t)$. RIGHT: Logarithmized yearly clicks Y_i.

York City (USA). The considered ad campaign is that of an online store selling outdoor equipment. (For reasons of confidentiality, we cannot publish the company's name). A lot of keywords received no impression during the considered time span of 365 days from April 1st, 2012, to March 31st, 2013. Therefore, we consider only the well established and relevant keywords that have been used on at least 320 days within the considered time span—leading to $n = 903$ trajectories observed at $p = 365$ grid points. The very few missing values in the logarithmized impression trajectories are imputed by zeros since a missing value means that the corresponding keyword did not receive an impression.

The considered functional linear regression model with PoIs in (2.1) is identifiable if the covariance function of the function-valued explanatory variable X_i is sufficiently non-smooth at the diagonal (see Section 2.2.1 and Theorem 3 in Kneip et al., 2016). Kneip et al. (2016) propose the following consistent estimator $\widehat{\kappa}$ for their κ controlling the smoothness at the diagonal of the covariance function:

$$\widehat{\kappa} = \log_2 \left(\frac{(1/(p-2k_\delta)) \sum_{j \in \mathcal{J}_{0,\delta}} \sum_{i=1}^{n} Z_{\delta,X_i}(t_j)^2}{(1/(p-2k_\delta)) \sum_{j \in \mathcal{J}_{0,\delta}} \sum_{i=1}^{n} Z_{\delta/2,X_i}(t_j)^2} \right).$$

An estimate of $\widehat{\kappa} < 2$ indicates identifiability, which is clearly fulfilled in our case where $\widehat{\kappa} = 0.03$.

The estimation results from applying our PES-ES estimation algorithm and the originally proposed KPS procedure are summarized in Figure 2.4 and Table 2.4. In case of the PES-ES estimate, the function-valued slope parameter $\widehat{\beta}(t)$ shows a peak in the late summer and a pronounced negative trend towards the

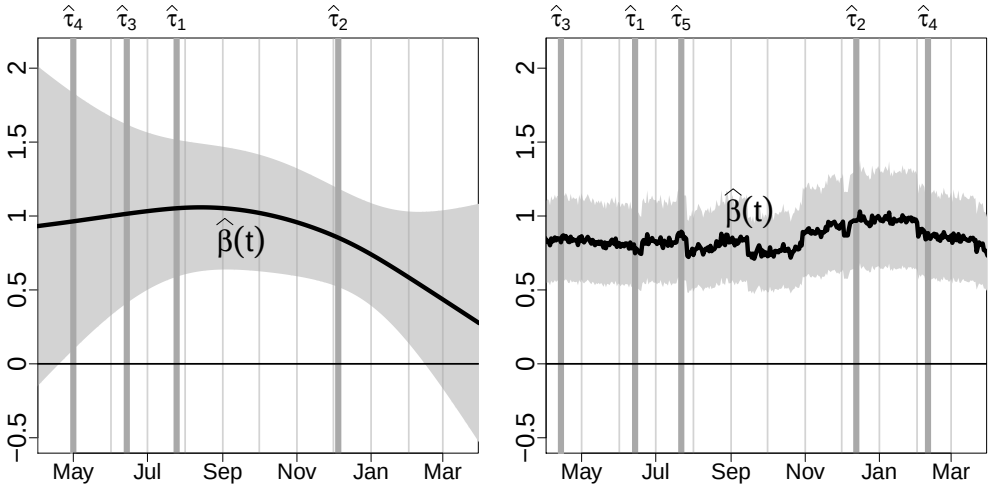


Figure 2.4: Result of the PES-ES (left panel) and KPS (right panel) estimate for $\beta(\cdot)$. The variabilities of the estimators are visualized using the gray shaded bands (see Remark 1).

Table 2.4: Estimate of PoI parameters β_r

PES-ES			KPS		
Location	Coef.	St.Err.	Location	Coef.	St.Err.
($\widehat{\tau}_4$) May 01	-0.17***	0.04	($\widehat{\tau}_3$) April 14	-0.10**	0.03
($\widehat{\tau}_3$) June 14	0.22***	0.03	($\widehat{\tau}_1$) June 14	0.22***	0.03
($\widehat{\tau}_1$) July 25	-0.15***	0.03	($\widehat{\tau}_5$) July 22	-0.17***	0.03
($\widehat{\tau}_2$) December 05	0.01	0.03	($\widehat{\tau}_2$) December 13	0.06*	0.03
			($\widehat{\tau}_4$) February 10	-0.11***	0.03

end of the considered period. The shape of $\widehat{\beta}(t)$ is in line with our expectations since the demand for outdoor equipment is generally greater during the summer months than during the winter months. The negative trend towards the end of the considered period is due to the strongly increased competition for outdoor equipment ads in Google AdWords during the considered period. Additionally, the estimation procedure identifies four PoIs (in order of the magnitude of $|\widehat{\beta}_s|$): June 14th ($\widehat{\tau}_3$; $\widehat{\beta}_3 = 0.22$), May 1st ($\widehat{\tau}_4$; $\widehat{\beta}_4 = -0.17$), July 25th ($\widehat{\tau}_1$; $\widehat{\beta}_1 = -0.15$), and December 5th ($\widehat{\tau}_2$; $\widehat{\beta}_2 = 0.01$), where the effect of the PoI at $\widehat{\tau}_2$ seems to be of lower importance.

Remark. *Drawing inference about the function-valued slope coefficient and the PoI parameters is a difficult issue in regression models with functional predictors.*

This is due to the fact that estimation in such models involves an ill-posed inversion problem and the estimator of the function-valued slope parameter is not asymptotically normal in the strong topology (Cardot et al., 2007). In addition, it is difficult to construct confidence regions for random elements in infinite dimensional Hilbert spaces with proper coverage probability (Choi and Reimherr, 2018). All we can do is to visualize the variability of the estimator that is due to the error term ϵ_i. For this purpose, we approximate the sampling variance of the composite parameter vector $\beta^{\rho}_{\widetilde{\mathcal{T}}}$ using Eq. (15.16) in Ramsay and Silverman (2005), Ch. 15, and show Bonferroni-adjusted Gaussian (invalid) confidence intervals in Figure 2.4.

The PoI $\widehat{\tau}_3$ on June 14th, with coefficient $\widehat{\beta}_3 = 0.22$, summarizes two positive effects. On the one hand, the store started a contest on May 23rd, 2012, giving away outdoor gear. This contest ended on June 13th, i.e., one day before the PoI which resulted in an increased click-through ratio of contest participants looking for the winners. On the other hand, the closest competitor started the spring sale, which led to a spillover bringing many interested buyers onto the homepage to compare prices.

The two other significant PoIs are explained by effects specific to the German calendar (about 80 percent of the customers live in Germany). The PoI $\widehat{\tau}_4$ on May 1st, with coefficient $\widehat{\beta}_4 = -0.17$, marks the Labor Day (commemorating the Haymarket Riot in Chicago in 1886), a national holiday in Germany which is typically an opportunity for family outings. Similar in interpretation, the PoI $\widehat{\tau}_1$ on July 25th, with coefficient $\widehat{\beta}_1 = -0.15$, marks the beginning of the official summer holidays in Baden-Württemberg and Lower Saxony—two large German states. Both PoIs show a negative sign, which is due to a higher volume in search queries related to outdoor activities, however, the users do not click on the sponsored impressions since they do not intend to buy something—they are only searching the Internet for (free) information on hiking trails etc., which results in a lower CTR.

By contrast to the PES-ES estimate of $\beta(\cdot)$, the KPS estimate of $\beta(\cdot)$ is difficult to interpret and does not fit to our expectations (see right panel of Fig. 2.4): the trajectory is very unstable, does not show the expected peak in late summer, and does not show the plausible negative trend towards the end of the considered period. Regarding the PoI selections, the KPS approach identifies essentially the same PoIs as the PES-ES approach, but favors one additional PoI at February 10 (see Table 2.4), which has a significant negative impact ($\widehat{\beta}_4 = -0.11$) on the outcome variable. This additional PoI may reflect a compensation for the missing negative trend in the KPS estimate of $\beta(\cdot)$; see our discussion in Section 2.2.2.

The log-transformations in $Y_i = \log(C_i)$ and $X_i(t) = \log(\mathcal{I}_i(t))$ allow us to interpret the estimated slope coefficients as elasticities. Taking derivatives with respect to $\mathcal{I}_i(t)$ at a single time point t leads to the following *time-local* elasticity:

$$\frac{\%\Delta C_i}{\%\Delta \mathcal{I}_i(t)} \approx \begin{cases} \widehat{\beta}_s & \text{if} \quad t = \widehat{\tau}_s \\ 0 & \text{else.} \end{cases}$$

That is, time-local changes in $\mathcal{I}_i(t)$ generally have no (i.e., practically negligible) effects on the yearly clicks C_i, except at PoIs, i.e., if $t = \widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{S}}$. For instance, a 1% increase in the impressions at the time point of the after-contest PoI ($t = \widehat{\tau}_3$) causes (on average) a 0.22% ($\widehat{\beta}_3 = 0.22$) increase in the yearly clicks.

The function-valued slope parameter $\widehat{\beta}(t)$ does not contribute to the time-local elasticities; however, it determines the elasticities with respect to time-global changes in the impressions, for instance, over the course of a month. The following Riemann sum allows for a simple, approximative approach to interpret such *time-global* elasticities:

$$\widehat{\log(C_i)} \approx \frac{1}{365} \sum_{t=1}^{365} \widehat{\beta}(t) \log\big(\mathcal{I}_i(t)\big) + \sum_{s=1}^{\widehat{S}} \widehat{\beta}_s \log\big(\mathcal{I}_i(\widehat{\tau}_s)\big).$$

For instance, the total elasticity of C_i with respect to $\mathcal{I}_i(t)$ for *all* $t \in$ August is given by

$$\sum_{t \in \text{August}} \frac{\%\Delta C_i}{\%\Delta \mathcal{I}_i(t)} \approx \frac{1}{365} \sum_{t \in \text{August}} \widehat{\beta}(t) + \sum_{s=1}^{\widehat{S}} \widehat{\beta}_s \mathbf{1}_{(\widehat{\tau}_s \in \text{August})},$$

where $\mathbf{1}_{(\texttt{TRUE})} = 1$ and $\mathbf{1}_{(\texttt{FALSE})} = 0$. That is, a 1% increase in the impressions $\mathcal{I}_i(t)$, simultaneously for all $t \in$ August, causes a 0.1% increase in the yearly clicks since $365^{-1} \sum_{t \in \text{August}} \widehat{\beta}(t) + \sum_{s=1}^{\widehat{S}} \widehat{\beta}_s \mathbf{1}_{(\widehat{\tau}_s \in \text{August})} \approx 0.1$. Hence, the time-global August-elasticity is half the size of the elasticity of the after-contest PoI. This is absolutely plausible since the super-imposed influence of the contest and the spillover definitely outperforms a high-season month such as August in terms of clicks-per-impressions.

2.5 Conclusion

In this work we propose an improved algorithm for estimating the unknown model components of the functional linear regression model with points of Kneip et al. (2016). Our estimation algorithm decouples the estimation of the points of im-

pact from the estimation of the function-valued slope parameter. The first step of the estimation algorithm, allows for a consistent estimation of the points of impact without knowledge (or pre-estimation) of the slope function. Given the consistent estimates of the points of impact, the second step of the estimation algorithm consists of an essentially classical estimation of the function-valued slope parameter. For this latter step we propose a generalization of the penalized smoothing splines estimator of Crambes et al. (2009), which allows to incorporate the estimates of the points of impacts. A further minor finite sample improvement is achieved by repeating the estimation of the points of impact, given the estimate of the function-valued slope parameter from the second step and by a finial repetition of the estimation of the slope parameter, given the updated estimates of the points of impact.

The new estimation algorithm significantly improves the original estimation procedure by Kneip et al. (2016). Using an extensive simulation study, we assess the robustness of our estimation algorithm for different data generating processes, different signal-to-noise ratios, different sample sizes and different sampling resolutions for discretizing the function-valued predictors.

The paper was originally motivated by an interesting case study on a Google AdWords ad campaign. Our proposed functional linear regression model with points of impacts allows for data-based insights into the (time-global) seasonal factors and the (time-local) events influencing the yearly number of clicks on impressions of the considered Google AdWords online ad campaign.

2.A Appendix

Additional simulation setups

Table 2.A.1: Squared bias and variance of the estimators. Lowest/highest MSE has the darkest/lightest gray-scale. **Scenario:** No standardization of the functions in preselection step and $p = 300$ grid points.

		Easy		Complicated		NoPoI		OnlyPoI	
$\int \widehat{\beta}(t)$		Bias^2	Var.	Bias^2	Var.	Bias^2	Var.	Bias^2	Var.
$n = 250$	PES	0.05	0.89	2.02	12.36	0.00	0.02	0.00	0.08
	PES-ES	0.05	0.74	1.81	11.64	0.00	0.01	0.00	0.07
	PES-2ES	0.04	0.72	1.85	11.69	0.00	0.01	0.00	0.06
	KPS	3.98	60.37	139.62	301.13	0.01	0.02	0.12	10.97
$n = 500$	PES	0.01	0.23	0.87	6.04	0.00	0.01	0.00	0.01
	PES-ES	0.01	0.2	0.89	5.55	0.00	0.01	0.00	0.02
	PES-2ES	0.01	0.19	0.9	5.46	0.00	0.01	0.00	0.02
	KPS	0.69	23.39	82.42	241.99	0.01	0.01	0.01	1.77
$\frac{1}{S}\sum \widehat{\beta}_s$									
$n = 250$	PES	0.02	0.55	0.06	0.42	-	-	0.00	0.02
	PES-ES	0.02	0.46	0.04	0.33	-	-	0.00	0.02
	PES-2ES	0.02	0.45	0.04	0.32	-	-	0.00	0.02
	KPS	0.04	0.65	1.01	3.31	-	-	0.00	0.14
$n = 500$	PES	0.00	0.1	0.03	0.14	-	-	0.00	0.00
	PES-ES	0.00	0.09	0.02	0.11	-	-	0.00	0.00
	PES-2ES	0.00	0.09	0.02	0.1	-	-	0.00	0.00
	KPS	0.01	0.22	0.49	2.25	-	-	0.00	0.02

Table 2.A.2: Squared bias and variance of the estimators. Lowest/highest MSE has the darkest/lightest gray-scale. **Scenario:** With standardization of the functions in preselection step and $p = 500$ grid points.

		Easy		Complicated		NoPoI		OnlyPoI	
$\int \widehat{\beta}(t)$		Bias^2	Var.	Bias^2	Var.	Bias^2	Var.	Bias^2	Var.
$n = 250$	PES	0.04	0.37	0.16	1.43	0.01	0.04	0.00	0.06
	PES-ES	0.03	0.3	0.09	0.94	0.00	0.02	0.00	0.04
	PES-2ES	0.04	0.3	0.08	0.94	0.00	0.02	0.00	0.03
	KPS	2.62	46.83	135.19	288.08	0.01	0.02	0.09	8.14
	CKS	-	-	-	-	0.00	0.01	-	-
$n = 500$	PES	0.01	0.09	0.06	0.41	0.00	0.02	0.00	0.01
	PES-ES	0.01	0.08	0.05	0.38	0.00	0.01	0.00	0.02
	PES-2ES	0.01	0.08	0.04	0.38	0.00	0.01	0.00	0.02
	KPS	0.43	17.82	91.76	238.1	0.01	0.01	0.01	2.71
	CKS	-	-	-	-	0.00	0.00	-	-
$\frac{1}{S}\sum \widehat{\beta}_s$									
$n = 250$	PES	0.01	0.08	0.01	0.02	-	-	0.00	0.04
	PES-ES	0.01	0.11	0.01	0.02	-	-	0.00	0.05
	PES-2ES	0.01	0.11	0.01	0.02	-	-	0.00	0.05
	KPS	0.03	0.54	1.04	3.26	-	-	0.00	0.14
$n = 500$	PES	0.00	0.01	0.00	0.01	-	-	0.00	0.00
	PES-ES	0.00	0.03	0.00	0.01	-	-	0.00	0.00
	PES-2ES	0.00	0.03	0.00	0.01	-	-	0.00	0.00
	KPS	0.01	0.16	0.62	2.39	-	-	0.00	0.05

Table 2.A.3: Squared bias and variance of the estimators. Lowest/highest MSE has the darkest/lightest gray-scale. **Scenario:** No standardization of the functions in preselection step and $p = 500$ grid points.

		Easy		Complicated		NoPoI		OnlyPoI	
$\int \widehat{\beta}(t)$		Bias^2	Var.	Bias^2	Var.	Bias^2	Var.	Bias^2	Var.
$n = 250$	PES	0.09	1.18	1.32	11.64	0.00	0.02	0.00	0.08
	PES-ES	0.06	0.94	1.36	10.69	0.00	0.01	0.00	0.06
	PES-2ES	0.06	0.9	1.32	10.42	0.00	0.01	0.00	0.06
	KPS	2.05	41.9	145.2	291.28	0.01	0.02	0.04	6.06
$n = 500$	PES	0.02	0.26	0.48	4.67	0.00	0.01	0.00	0.01
	PES-ES	0.01	0.18	0.46	4.1	0.00	0.01	0.00	0.01
	PES-2ES	0.01	0.18	0.48	4.04	0.00	0.01	0.00	0.01
	KPS	0.47	19.09	81.27	229.75	0.01	0.01	0.01	2.77
$\frac{1}{S}\sum \widehat{\beta}_s$									
$n = 250$	PES	0.01	0.23	0.04	0.23	-	-	0.00	0.02
	PES-ES	0.01	0.28	0.04	0.22	-	-	0.00	0.04
	PES-2ES	0.01	0.3	0.04	0.22	-	-	0.00	0.04
	KPS	0.03	0.5	1.17	3.2	-	-	0.00	0.06
$n = 500$	PES	0.00	0.13	0.02	0.08	-	-	0.00	0.00
	PES-ES	0.00	0.12	0.02	0.07	-	-	0.00	0.00
	PES-2ES	0.00	0.11	0.02	0.09	-	-	0.00	0.00
	KPS	0.01	0.2	0.66	2.49	-	-	0.00	0.04

Table 2.A.4: Mean squared bias and variance. Lowest/highest MSE has the darkest/lightest gray-scale. DGP "Complicated" with different standard deviations σ_ϵ.

		$\sigma_\epsilon = 0.5$		$\sigma_\epsilon = 1$		$\sigma_\epsilon = 2$		$\sigma_\epsilon = 5$	
$\int \widehat{\beta}(t)$		Bias^2	Var.	Bias^2	Var.	Bias^2	Var.	Bias^2	Var.
$n = 250$	PES	0.22	1.78	0.28	3.76	0.58	13.76	0.67	31.37
	PES-ES	0.2	1.67	0.23	2.37	0.3	9.21	0.23	38.64
	KPS	19.51	95.41	1.52	35.09	0.56	27.38	0.21	46.26
$n = 500$	PES	0.12	0.67	0.15	1.36	0.29	5.59	0.5	21.84
	PES-ES	0.09	0.31	0.13	0.55	0.2	2.9	0.27	21.9
	KPS	11.47	80.98	0.6	20.17	0.2	10.18	0.14	25.74
$\frac{1}{\lvert S\rvert}\sum \widehat{\beta}_s$									
$n = 250$	PES	0.01	0.05	0.02	0.18	0.03	1.04	2.51	9.02
	PES-ES	0.01	0.05	0.01	0.18	0.01	0.69	1.9	11.37
	KPS	0.31	2.16	0.17	2.15	0.11	3.08	1.79	14.06
$n = 500$	PES	0.00	0.02	0.01	0.06	0.01	0.27	0.97	6.06
	PES-ES	0.00	0.02	0.01	0.05	0.01	0.2	0.42	5.86
	KPS	0.19	1.61	0.04	0.9	0.03	1	0.62	8.33

Table 2.A.5: Percentage of replications with correct detection of all points of impact $\tau_1, \ldots, \tau_S$.

		300 grid points			500 grid points		
		Easy	Compl.	OnlyPoI	Easy	Compl.	OnlyPoI
$n = 250$	PES	86.1	28.4	98.3	85	28	98.6
	PES-ES	87.4	30.4	98.3	86.5	30.4	98.8
	PES-2ES	87.5	30.4	98.3	86.6	30.8	98.8
	KPS	87.9	25.7	98.4	90.9	24.3	98.7
$n = 500$	PES	97.6	54.5	100	97.4	58.9	99.9
	PES-ES	97.8	56.2	100	97.8	61.1	99.9
	PES-2ES	97.8	56.2	100	97.8	61.2	99.9
	KPS	95.6	44.5	99.8	96.9	45.4	99.4

Chapter 3

The spatial decay of human capital externalities

3.1 Introduction

Workers interact with each other within and across firms. They share their knowledge, discuss ideas and adopt procedures and technologies. All of these interactions potentially increase the productivity of workers through *'human capital externalities'* (Davis and Dingel, 2019; Acemoglu, 1996; Lucas, 1988; Marshall, 1890). Although a large empirical literature supports the existence of geographically bounded human capital externalities (Cornelissen et al., 2017; Ciccone and Peri, 2006; Moretti, 2004; Rauch, 1993), little is known about the exact spatial extent of human capital externalities. For several reasons, human capital externalities are likely to decline with distance. For instance, distance raises the costs of planned social interactions, such as meetings. Similarly, distance reduces the likelihood of unintended encounters that lead to the exchange of knowledge. Moreover, because distance generally raises the number of intermediaries between individuals in a social network and an increasing number of intermediaries impedes information flows, distance depresses indirect information flows. Consequently, individuals are likely to benefit more from proximate than from distant neighbors.

It is the aim of this paper, to shed some light on the question how human capital externalities attenuate with distance. To this end, we draw on a large and novel micro panel dataset that features the exact coordinates of nearly all German establishments and rich information on individual workers over one and a half decades. Furthermore, we introduce a novel estimation procedure to the urban economic literature that is capable of evaluating such detailed geodata.

This allows us to estimate the spatial attenuation of human capital externalities with high precision. In line with previous studies we assume that wages reflect the productivity of workers such that we can measure human capital externalities based on external wage effects from the local concentration of high-skilled workers. External effects may arise from knowledge exchange (Marshall, 1890; Lucas, 1988) or the diffusion of new technologies (Nelson and Phelps, 1966; Acemoglu, 1998).

Previous empirical research provides initial evidence for spatially decreasing human capital externalities. Using cross-sectional data from the US, Rosenthal and Strange (2008) construct concentric rings around workers that measure the concentration of human capital within 5 miles and between 5 to 25 miles. To explore how human capital externalities attenuate with distance, they regress individual wages on the concentration of human capital within these rings. They find that human capital externalities from the inner ring are notably larger than externalities from the outer ring. A closely related study by Fu (2007) adopts the strategy of Rosenthal and Strange (2008) to analyze cross-sectional data from the Boston metropolitan area. Using more precise geocoded data, Fu (2007) measures the concentration of human capital within finer rings (i.e., 0-1.5, 1.5-3, 3-6 and 6-9 miles). Fu (2007) finds evidence that human capital externalities may vanish after only three miles. Recent findings from the Netherlands in a setting with panel data and concentric rings of 0-10, 10-40, and 40-80 kilometers' distance suggest that human capital externalities reach 10 kilometers (Verstraten, 2018). Although these studies provide evidence for the spatial attenuation of human capital externalities, the exact decay of the effects remains unclear because the literature is constrained either by relatively imprecise geo-information or by specific data from a single area. Furthermore, most empirical evidence is restricted to cross-sectional data, which complicates causal inference. Additionally, the described studies overlook that human capital externalities from high-skilled workers are entangled with labor market supply and demand effects (Katz and Murphy, 1992; Card and Lemieux, 2001; Borjas, 2003; Moretti, 2004; Ciccone and Peri, 2006).

To fully exploit the information from exact geocodes of workplaces, we adopt a methodologically fresh approach and measure the magnitude of human capital externalities (or spillovers) with respect to distance in a continuous manner. Recent developments in functional data analysis (FDA) provide particularly suitable frameworks. FDA is a branch of statistics that extends classical statistical methods to random variables with a functional nature, such as curves or surfaces over a continuous domain. Typical examples of such data are temperature curves, growth curves or the continuous evolution of stock prices over time. The continuity of curves entails that adjacent values are somehow related. In many

applications, exploiting this information makes FDA more efficient than classical multivariate methods on discretized data.

While statisticians employ FDA in a wide range of applications (see Ullah and Finch, 2013 for a systematic overview), FDA is applied quite rarely in economics (examples include Ramsay and Ramsey, 2002, Wang et al., 2008 and Caldeira and Torrent, 2017).[1] This paper, therefore, illustrates the potential of FDA in economic research with high-dimensional variables. Our approach relies on a functional linear regression model in which a scalar outcome variable (log wage) is regressed on observations of a functional random variable (share of high-skilled workers as a function of distance to a worker's workplace). For this purpose, we augment the classical scalar-on-function regression model to incorporate further scalar-valued explanatory variables and use an estimation procedure, suggested by Crambes et al. (2009), that is based on smoothing splines and makes it possible to model the function-valued spillover parameter very flexibly. The estimated spatial spillover function relates wages to the share of high-skilled workers as a function of distance, which is evaluated at 500 meter intervals up to 50 kilometers.

The previous literature that estimates the spatial attenuation of economic effects follows a semi-parametric approach (e.g., Rosenthal and Strange, 2008; Fu, 2007; Verstraten, 2018; Gibbons et al., 2021; Faggio et al., 2019; Faggio, 2019).[2] In the semi-parametric approach, econometricians estimate linear models in which the main explanatory variable is measured in several geographically concentric rings or circles around observations. Although the semi-parametric approach is generally well suited to measure the spatial attenuation of economic effects and is a straightforward application of the linear OLS model, it is less precise compared to our FDA approach. The reason is that multicollinearity issues usually do not allow to estimate effects from a large and fine-graded series of measurement points. To circumvent multicollinearity issues researches are therefore forced to construct relatively broad rings or circles that measure the spatial distribution of the explanatory variable. Our FDA approach solves this issue by regularizing the parameter estimates. This enables us to exploit geographically extremely fine graded data and to estimate the spatial attenuation of economic effects with detail.

[1] Readers with a general interest in FDA are referred to the textbooks of Ramsay and Silverman (2005); Ferraty and Vieu (2006); Horváth and Kokoszka (2012) and Hsing and Eubank (2015).

[2] Some examples of studies that investigate the spatial patterns of agglomeration effects are: Arzaghi and Henderson (2008), who study networking effects within the advertising agency industry in Manhattan; Ahlfeldt et al. (2015), who examine productivity externalities in Berlin; Andersson et al. (2019), who evaluate productivity effects from industry specialization and diversity in Swedish cities; and Faggio (2019) and Faggio et al. (2019), who assess the local labor market impact of relocations of public sector jobs in the UK and Germany.

There are two major challenges in identifying regional human capital externalities, namely, confounding labor market supply and demand effects and the sorting of high-skilled workers into high-wage regions. We aim to address both problems with an extensive set of time-varying fixed effects. If high- and low-skilled workers are imperfect substitutes, standard supply and demand models indicate that an increase in the share of high-skilled workers raises (lowers) the wages of low-skilled (high-skilled) workers (see Ciccone and Peri, 2006 and Moretti, 2004 for detailed explanations in our context). Thus, spillovers are potentially entangled with labor market supply and demand effects. To disentangle spillover from supply and demand effects we follow Eppelsheimer and Möller (2019) and exploit the different spatial natures of the two effects. While it is plausible that supply and demand effects are spatially very equally distributed within local labor markets (i.e., supply and demand effects originating in one part of the city very uniformly affect wages throughout the city), the intensity of spillover effects truly depends on distance (i.e., spillovers affect close neighbors more than distant neighbors). Thus, in the data, we aim to purge spillover effects from supply and demand effects by eliminating variation that is common within regional labor markets. To do so, we include time-varying labor-market-area fixed effects in the econometric specification (i.e., a specific intercept for every labor market area in each year). Because supply and demand effects may have different impacts on high- and low-skilled workers, we further interact these labor-market-area-year fixed effects with a skill dummy.

Following Cornelissen et al. (2017), who, in a related context, address worker sorting at the firm level (Abowd et al., 1999; Card et al., 2013), we address sorting of high-skilled workers into high-wage regions (Acemoglu and Angrist, 2000) by including a comprehensive set of fixed effects. In particular, the above-introduced labor-market-area-year fixed effects nullify unobserved regional heterogeneity that might attract high-skilled workers, such as (changes in) average wages, general labor market conditions and amenities. Importantly, labor-market-area-year fixed effects also cover temporal labor market shocks that might pull or push skilled workers into or out of regions—a concern raised by Moretti (2004). Additionally, we aim to account for locational advantages within regions (e.g., proximity to infrastructure and facilities) and unobserved individual heterogeneity with worker-firm match fixed effects.

We find significant spillover effects from the local concentration of high-skilled workers. Moreover, our estimates reveal that spillover effects decay with distance. Human capital externalities from direct neighbors (i.e., high-skilled workers who are located within a 0.5 kilometer radius) are roughly twice as large as spillovers from high-skilled workers that are located 10 kilometers apart. After 15 kilo-

meters, spillover effects vanish completely. Overall, an evenly distributed, one-standard-deviation increase in the local share of high-skilled workers leads to wage gains of 2%. The magnitude of this effect is comparable to *classical* estimates at the aggregate level (NUTS-3), for which we provide results in the appendix. Reassuringly, semi-parametric estimates with our data lead to similar results. In general, our findings are in line with the urban economic literature and support the existence of human capital externalities. Additionally, our results imply that human capital externalities cover entire cities. However, the majority of their effect is bounded within the near neighborhood of high-skilled workers. Workers at firms located in, or very close to, a skilled neighborhood, therefore, benefit most from spillovers. Those who work further away from skilled neighbors gain less, and workers in very remote regions do not profit from human capital externalities at all.

The remainder of the paper is organized as follows. The next section explains the estimator and our identification strategy. Section 3.3 summarizes the data. Section 3.4 presents our main findings and compares our estimation procedure to the semi-parametric approach. The section also illustrates the statistical properties of the estimator in a simulation study and provides an overview of several robustness checks. Section 3.5 concludes the paper.

3.2 Estimation strategy

This paper seeks to measure the spatial attenuation and reach of human capital externalities. Thus, we aim to measure external productivity effects from the local concentration of human capital. As productivity itself cannot be observed directly, we use individual wages as a proxy. To model the concentration of human capital, we compute the share of high-skilled workers around each workplace in our data as a continuous function that depends on distance. Consequently, each workplace has a unique functional description of the surrounding concentration of human capital. To fully exploit the available information, we also model the treatment effect as a continuous function that depends on distance.

In the following, we present our estimator, describe how we adjust it to meet the requirements of our application, discuss statistical inference and define the functional representation of the concentration of high-skilled workers around workplaces in more detail. Finally, we specify our identification strategy that addresses endogenous sorting of workers and confounding labor market supply and demand effects. For notational simplicity, we will formulate the estimation framework only for the cross-section. The empirical model, of course, takes the panel structure of our data into account.

3.2.1 The estimator

The spatial allocation of human capital varies considerably across and within administrative boundaries. For a given location, say worker i's workplace, the concentration of high-skilled workers in the immediate neighborhood, therefore, may differ from the concentration in the greater neighborhood. In principle, it is even possible to measure the concentration of high-skilled workers at any distance to worker i's workplace. Thus, one can naturally regard the concentration of high-skilled workers with respect to the distance to worker i's workplace as a curve. We use these function-valued observations as an explanatory variable to assess how the concentration of human capital influences productivity in space.

The functional linear regression model with a scalar response variable is a suitable framework to measure such a relationship. With Y_i being the scalar dependent variable, log wage in our empirical analysis, the model is defined as

$$Y_i = \int_0^1 \beta(z) X_i(z) \, \mathrm{d}z + \varepsilon_i, \; i = 1, \ldots, n \tag{3.1}$$

where $X_i \in L^2([a,b])$, $i = 1, \ldots, n$, in the classical setup are n independent and identically distributed (iid) random functions defined on a common domain, which we set to $[0,1]$ without loss of generality. In our application, $X_i(z)$ is the share of high-skilled workers among all workers being located z units away from worker i's workplace. The function-valued coefficient parameter $\beta \in L^2([0,1])$ is the quantity of interest and describes the influence of X_i on Y_i which can be different for different z. This coefficient function measures the magnitude of the productivity spillover induced by human capital located z units away from worker i's workplace. The error term ε_i is independently distributed and has a mean of zero and homoscedastic variance (we will later consider heteroscedastic and autocorrelated errors).

Model (3.1) has received considerable attention in the FDA literature (see Morris, 2015, for an overview). Classically, the estimation of β is based on the Karhunen-Loève decomposition of the empirical covariance operator of the observed curves X_i. A drawback of the classical approach is that the expansion of the so-called functional principal component (FPC) estimator heavily depends on the random curves' correlation structure. In this paper, we instead build on the smoothing spline estimator proposed by Crambes et al. (2009). This approach has the advantage that the basis functions are independent of the curves X_i, which results in a more flexible function space for $\widehat{\beta}$. From an asymptotic perspective, both estimators have minimax-optimal convergence rates (Hall and Horowitz, 2007; Crambes et al., 2009).

In the following, $\mathbf{X}$ denotes the $n \times p$ matrix holding all n curves $X_i(z)$ observed at p grid values $z_1, \ldots, z_p$, and $\mathbf{Y}$ denotes the n-vector with observations of the dependent variable. To estimate β, the approach of Crambes et al. (2009) minimizes the penalized sum of squared residuals

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{1}{p}\sum_{j=1}^{p}\beta(z_j)X_i(z_j)\right)^2 + \rho\left(\frac{1}{p}\sum_{j=1}^{p}\pi_\beta^2(z_j) + \int_0^1 (\beta^{(m)}(z))^2\,\mathrm{d}z\right). \quad (3.2)$$

Expression (3.2) comprises two terms: While the first one quantifies how well the model can fit the data, the second term measures curvature of β (via its m-th derivative). The term $\frac{1}{p}\sum_{j=1}^{p}\pi_\beta^2(z_j)$, where $\pi_\beta(z)$ is the best approximation of $\beta(z)$ by a polynomial of degree $m-1$, is not common in traditional smoothing splines regressions. However, this term is necessary to ensure uniqueness of the solution without imposing further assumptions on the random function X_i. The penalty parameter $\rho \geq 0$ controls the flexibility of the estimated parameter function $\widehat{\beta}$. With $\rho = 0$, as one extreme, equation (3.2) coincides with the least-squares criterion and, with $\rho \to \infty$ as the other extreme, $\widehat{\beta}$ is constrained to be a function of which the m-th derivative is zero. If, for instance, $m = 2$, the coefficient function will be a straight line. The minimization problem of equation (3.2) has a closed-form solution, given by

$$\left(\widehat{\beta}(z_1), \ldots, \widehat{\beta}(z_p)\right)' = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}'\mathbf{X} + \rho\mathbf{A}\right)^{-1}\mathbf{X}'\mathbf{Y}, \quad (3.3)$$

where the penalty matrix $\mathbf{A}$ corresponds to the second term of the minimization criterion (3.2).[3] Traditional smoothing splines penalize second derivatives. Thus, we set $m = 2$, which results in an expansion of cubic natural splines with knots at

[3] The matrix $\mathbf{A} = \mathbf{P} + p\mathbf{A}^*$ is a combination of a classical regularization matrix $\mathbf{A}^* \in \mathbb{R}^{p\times p}$ and a nonstandard projection matrix $\mathbf{P} \in \mathbb{R}^{p\times p}$ projecting into the space spanned by polynomial functions of degree $m-1$. The latter ensures the invertibility of $\mathbf{X}'\mathbf{X} + \rho\mathbf{A}$ and is defined by $\mathbf{P} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'$, where $\mathbf{W} = (z_j^q)_{j,q} \in \mathbb{R}^{p\times m}$, $q = 0, \ldots, m-1$. The regularization matrix $\mathbf{A}^*$ is the traditional penalty matrix of smoothing splines, defined by

$$\mathbf{A}^* = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\left(\int_0^1 \mathbf{b}^{(m)}(z)\mathbf{b}^{(m)}(z)'\,\mathrm{d}z\right)(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B},$$

where $\mathbf{B}$ denotes the $p \times p$ matrix of the p basis functions, evaluated at the p grid values, and $\mathbf{b}^{(2)}(z)$ is, for given value of $z \in [0,1]$, a p-vector of second derivatives of the p basis functions. Using this notation and approximating the integral via the Riemann sum, the penalty term of equation (3.2) becomes

$$\frac{1}{p}\sum_{j=1}^{p}\pi_\beta^2(z_j) + \int_0^1 (\beta^{(m)}(z))^2\,\mathrm{d}z = \left(\beta(z_1), \ldots, \beta(z_p)\right)\mathbf{A}\left(\beta(z_1), \ldots, \beta(z_p)\right)'.$$

$z_1, \ldots, z_p$. A computational convenient basis for the corresponding p-dimensional function space is formed by cubic B-splines with knots $z_1, \ldots, z_p$ and the natural boundary condition.[4] Let $\mathbf{b} : [0,1] \to \mathbb{R}^p$ denote the p-dimensional basis system, then, for any $\left(\widehat{\beta}(z_1), \ldots, \widehat{\beta}(z_p)\right) \in \mathbb{R}^p$, there is a $\theta \in \mathbb{R}^p$ with $\widehat{\beta}(z_j) = \theta' \mathbf{b}(z_j)$ for all $j = 1, \ldots, p$, and the estimated coefficient functioned is expanded accordingly $\widehat{\beta}(z) = \theta' \mathbf{b}(z), z \in [0,1]$. In addition to providing a more 'natural' way of modelling β (as a smooth function), the smoothing splines approach has the nice property that the curves X_i can enter the model 'as is' and it is not necessary to approximate them as it is the case for the FPC estimator.

The estimation framework proposed by Crambes et al. (2009) does not cover the case of additional (scalar) covariates. Therefore, to account for the influence of further explanatory variables, we have to expand model (3.1) with a k-vector of scalar-valued explanatory variables Z_i and a corresponding parameter vector γ:

$$Y_i = \int_0^1 \beta(z) X_i(z) \, \mathrm{d}z + Z_i' \gamma + \varepsilon_i. \tag{3.4}$$

Accordingly, we augment the smoothing spline estimator of Crambes et al. (2009) to incorporate scalar-valued explanatory variables. Let $\mathbf{X_Z}$ denote the compound data matrix $(\mathbf{X}, p\mathbf{Z})$, where the matrix $\mathbf{Z}$ holds the sample values of the k additional scalar explanatory variables. After approximating the integral via the Riemann sum and using matrix notation, model (3.4) can also be written as

$$\begin{aligned} \mathbf{Y} &= \frac{1}{p} \mathbf{X} \left(\beta(z_1), \ldots, \beta(z_p)\right)' + \mathbf{Z}\gamma + \boldsymbol{\varepsilon} \\ &= \frac{1}{p} (\mathbf{X}, p\,\mathbf{Z}) \left(\beta(z_1), \ldots, \beta(z_p), \gamma_1, \ldots, \gamma_k\right)' + \boldsymbol{\varepsilon}. \end{aligned}$$

In order to write the estimator for β and γ in a compact way, it is necessary to augment the penalty matrix accordingly. The parameters of the k scalar predictors are not penalized, hence, we have to append k zero columns and k zero rows to the matrix $\mathbf{A}$ to get a properly dimensional matrix $\mathbf{A_Z}$:[5]

$$\mathbf{A_Z} = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{(p+k)\times(p+k)}.$$

[4] For readers not familiar with the construction of spline basis systems and regularization techniques used here, we refer to Hastie and Tibshirani (1990) or De Boor (1978).

[5] Let $\alpha = \left(\beta(z_1), \ldots, \beta(z_p)\right)'$ and $\alpha_{\mathbf{Z}} = \left(\beta(z_1), \ldots, \beta(z_p), \gamma_1, \ldots, \gamma_k\right)'$. It is obvious, that the quadratic forms $\alpha' \mathbf{A} \alpha$ and $\alpha_{\mathbf{Z}}' \mathbf{A_Z} \alpha_{\mathbf{Z}}$ are equivalent.

The compound estimator of (discretized) β and γ then is

$$\widehat{\boldsymbol{\beta}} = \big(\widehat{\beta}(z_1), \ldots, \widehat{\beta}(z_p), \widehat{\gamma}_1, \ldots, \widehat{\gamma}_k\big)' = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}_{\mathbf{Z}}'\mathbf{X}_{\mathbf{Z}} + \rho\mathbf{A}_{\mathbf{Z}}\right)^{-1}\mathbf{X}_{\mathbf{Z}}'\mathbf{Y}, \tag{3.5}$$

where ρ again plays the role of a tuning parameter to control the complexity of the estimate of the function-valued slope parameter β. The parameter ρ itself has no meaningful interpretation. Instead, a well-established measure for the complexity of the estimate $\widehat{\beta}$ is the *effective number of degrees of freedom* (edf):

$$\text{edf}(\rho) = \text{trace}\big(\mathbf{H}_{\mathbf{Z}}^{\rho}\big), \tag{3.6}$$

where $\mathbf{H}_{\mathbf{Z}}^{\rho} = (np)^{-1}\mathbf{X}_{\mathbf{Z}}\left((np)^{-1}\mathbf{X}_{\mathbf{Z}}'\mathbf{X}_{\mathbf{Z}} + \rho\mathbf{A}_{\mathbf{Z}}\right)^{-1}\mathbf{X}_{\mathbf{Z}}'$ is the so-called *smoothing matrix* of model (3.4). Given a predefined number of degrees of freedom, equation (3.6) allows us to determine ρ. For instance, $\rho = 0$ results in $\text{edf}(\rho) = p + k$, which is the most flexible fit, while large values for ρ will yield values of $\text{edf}(\rho)$ close to 2, which corresponds to a straight line. In our preferred specification, we set $\text{edf}(\rho) = 2.5$, such that the estimate can be more flexible than a straight line. This choice mainly is driven by the expectation that the true spillover mechanism has a monotonically decreasing shape and approaches zero at some point. Therefore, allowing more flexibility (3 might be similar to a parabola), in our view is not necessary. We experiment with different penalties in appendix 3.A.3. Qualitatively, our results do not depend on the exact choice of the penalty term ρ.

3.2.2 Inference

From a theoretical perspective, drawing local inference about the slope parameter β in the functional linear regression model is a difficult issue. When $X_i(z)$ are elements of the infinite-dimensional Hilbert space L^2, the estimator $\widehat{\beta}$ is not asymptotically normal (w.r.t. the strong topology on L^2). The reason is that such models belong to the class of ill-posed inversion problems, that is, the (compact) covariance operator of the random curves $X_i(z)$ has no bounded inverse (see Cardot et al., 2007, for details). As a consequence, it is not possible to construct appropriate pointwise confidence bands around $\widehat{\beta}$ when the curves lie in L^2. This fundamental problem also can not be solved using resampling techniques such as the bootstrap, the latter being computationally quite demanding.

To quantify the estimation uncertainty, we therefore proceed as in the classical linear regression framework. In classical linear regression, inference about the model parameters builds on the variance of the parameter estimates conditional

on the observed regressors. Similarly, the (pointwise) variance of the compound parameter vector $\widehat{\boldsymbol{\beta}}$ for given observations of curves and covariates, $\mathbf{X_Z}$, and the regularization parameter, ρ, can be computed by (see also Ramsay and Silverman, 2005, equation 15.16)

$$\operatorname{Var}(\widehat{\boldsymbol{\beta}}|\mathbf{X_Z}, \rho) = \frac{1}{n^2}\left(\frac{1}{np}\mathbf{X_Z'X_Z} + \rho\mathbf{A_Z}\right)^{-1}\mathbf{X_Z'\Omega X_Z}\left(\frac{1}{np}\mathbf{X_Z'X_Z} + \rho\mathbf{A_Z}\right)^{-1}. \tag{3.7}$$

Here, $\boldsymbol{\Omega}$ is the covariance matrix of the error term, which does not necessarily have to be diagonal. By replacing this matrix with an appropriate estimate $\widehat{\boldsymbol{\Omega}}$, we obtain an estimate for the variance of the parameter vector $\widehat{\boldsymbol{\beta}}$. Furthermore, we estimate the $\mathbf{X_Z'\Omega X_Z}$ based on clustered standard errors at the firm level (see, for instance, Abadie et al., 2017, equation 2.3).

We use the variance (3.7) to visualize the pointwise variability of the estimate $\widehat{\boldsymbol{\beta}}$ with confidence bands. We obtain confidence bands by multiplying the square-root of the corresponding diagonal entry of $\operatorname{Var}(\widehat{\boldsymbol{\beta}}|\mathbf{X_Z}, \rho)$ by appropriate quantiles of the normal distribution. To account for the family-wise error rate, we divide the significance level by the effective degrees of freedom. The simulation exercise (section 3.4.3) supports such a procedure and shows that it indeed controls size when the (global) null is a linear function. Even if the true parameter β_0 is more complex, the estimator is able to resemble β_0 quite well, although a local bias leads to a pointwise violation of the nominal coverage probability of the confidence bands.

3.2.3 Calculation of curves

A key feature of our analysis is the representation of the spatial density of high-skilled workers around workplaces as curves. For the actual measure of the concentration of high-skilled workers we follow the recent literature on regional human capital externalities (for example Moretti, 2004) and use the share of high-skilled workers out of all workers. To calculate curves from geocoded data, we compute the values of the functions $X_i(z)$ for each worker i on an equidistant grid $z_1, \ldots, z_p$:

$$X_i(z_j) = \frac{n^{hs}_{i,[z_j - h,\, z_j)}}{n_{i,[z_j - h,\, z_j)}}. \tag{3.8}$$

Here, $n^{hs}_{i,[z_j - h,\, z_j)}$ refers to the number of high-skilled individuals for which the spherical distance between their working location and the workplace of worker i is at least as large as $z_j - h$ and smaller than z_j. Similarly, $n_{i,[z_j - h,\, z_j)}$ is the number of all workers (high-skilled and low-skilled) within the same distance window. In

other words, the value of the curve X_i at distance z_j indicates the share of high-skilled workers in all workers within the distance window $[z_j - h, z_j)$, where h is a fixed bandwidth. To ensure that a firm's own skill structure does not affect measurements of its neighborhood, we compute $X_i(z_1)$ without its own number of workers. Thus, we only measure regional human capital externalities without firm-internal spillovers. To balance analytical precision and computational costs, we choose a bandwidth of $h = 500$ meters and compute $X_i(z_j)$ on the grid $z_j = 500m, 1000m, \ldots, 50000m$.

3.2.4 Identification

Having explained the estimator, we will now address confounding labor market demand and supply effects and endogenous sorting of individuals. The empirical literature has established that high- and low-skilled labor are imperfect substitutes (e.g., Autor et al., 2008; Ciccone and Peri, 2005; Card and Lemieux, 2001; Krusell et al., 2000). As Acemoglu and Angrist (1999), Moretti (2004) and Ciccone and Peri (2006) illustrate, apart from potential externalities, changes in the supply of high-skilled labor therefore entail a market mechanism that affects wages. Due to these labor market demand and supply effects, an increase in the share of high-skilled workers in the labor market depresses the wages of high-skilled workers and raises the wages of low-skilled workers. Consequently, changes in the local concentration of high-skilled workers might simultaneously influence wages through labor market effects and human capital externalities.

To disentangle human capital externalities from labor market supply and demand effects, we follow Eppelsheimer and Möller (2019) and exploit the different spatial nature of the two effects. On the one hand, the intensity of human capital externalities should be highly localized and decay with distance. We therefore expect larger spillovers from close neighbors than from distant neighbors. On the other hand, labor market supply and demand effects arguably uniformly affect larger areas. Consequently, purging the data from variation that is common within larger areas potentially eliminates labor market supply and demand effects while not affecting highly localized externalities.

In our estimation framework we aim to achieve such a disentanglement of human capital externalities and labor market supply and demand effects by including local labor market area fixed effects. To identify local labor markets we follow the definition from the German Federal Ministry for Economic Affairs and Energy (BMWi) that classifies 258 local labor market areas based on commuter links (Kosfeld and Werner, 2012). Although these areas are designed to identify local labor markets we cannot guarantee that supply and demand effects are per-

fectly uniform within each area. Thus, some correlation between human capital externalities and labor market supply and demand effects may still remain in the data.

As labor market supply and demand effects vary over time and effects might be different for different skill groups, we expand equation (3.4) to include time-varying labor-market-area fixed effects for each skill group π_{rst} (i.e., an intercept for each labor market area and skill group in every year). Our full estimation equation is:

$$Y_{it} = \int_0^1 \beta(z) X_{it}(z)\,\mathrm{d}z + Z'_{it}\gamma + \theta_{if} + \tau_t + \omega_o + \pi_{rst} + u_{it}. \tag{3.9}$$

Here, Y_{it} is the individual log wage of worker i in year t, and $X_{it}(z)$ is the share of high-skilled workers, described as a continuous curve around the workplace of individual i that depends on distance z. Note that all workers of a given establishment in year t share the same locational characteristics, specifically they all have the same curve $X_{it}(z)$. $\beta(z)$ is the associated spillover function that we seek to retrieve from the data. The model controls for time-varying observable individual, establishment and regional characteristics Z_{it} and a series of fixed effects. θ_{if} is a worker-firm match fixed effect, τ_t is a year fixed effect and ω_o is an occupation fixed effect.

Endogenous sorting of workers (Acemoglu and Angrist, 2000) constitutes another challenge in identifying human capital externalities. Although the empirical literature finds that workers do not sort into cities based on their (unobserved) abilities (De la Roca and Puga, 2017; Glaeser and Mare, 2001), there is evidence of ability-driven sorting of workers into firms (Card et al., 2013; Abowd et al., 1999). If more-productive firms locate in neighborhoods with high concentrations of human capital, sorting of workers would thus create a spurious relationship between wages and the concentration of high-skilled workers around the workplace.

Inspired by Cornelissen et al. (2017), we aim to address sorting with an extensive set of fixed effects. To ensure that neither sorting of workers nor sorting of firms biases our estimates, we include worker-firm match fixed effects (θ_{if}) in our model. Worker-firm match fixed effects eliminate unobservable characteristics of workers and firms that are time-constant during the matched employment period.[6] An additional benefit of worker-firm match fixed effects is that they also remove neighborhood characteristics from the data that are time-constant

[6] Different to the literature on assortative matching we are not interested in describing worker-firm matches. As such, we do not aim to separate match specific effects from worker and firm productivity. Instead, our goal is merely to remove inherent worker, firm and match effects from the data.

during the matched employment period. These characteristics include locational advantages that might influence productivity, like proximity to infrastructure or market access. Other neighborhood specific influences on wages, like housing and commuting costs, are also covered.

A further challenge is that high-wage areas might attract high-skilled workers. Such a behavior would reverse the direction of causality in our estimates (Moretti, 2004). Our identification strategy aims to overcome the issue by removing all time-constant and time-varying variation on the local labor market level by including local labor market area fixed effects (π_{rst}). These fixed effects erase push and pull factors that might attract or distract high-skilled workers. Thus, reversed causality is quite unlikely.[7]

In summary, equation (3.9) allows us to estimate human capital externalities that are unrelated to labor market demand and supply effects that are spatially constant within local labor market areas and endogenous sorting of individuals. We also purge the data from potentially confounding neighborhood characteristics that are relatively stable over time. The remaining variation of $X_{it}(z)$ in equation (3.9) stems from temporal intra-regional changes in the concentration of high-skilled workers around workplaces. Note that our approach only measures human capital externalities at the workplace, not the place of residence.

3.3 Data and descriptive statistics

3.3.1 Data

In the empirical analysis, we combine administrative data on almost all German firms and rich data from a representative sample of workers over a period of 15 years. Our panel data include exact geo-coordinates of establishments and therefore allow us to describe the distribution of high-skilled workers as spatial functions around individual workplaces. We evaluate the share of high-skilled workers at 500-meter intervals up to a distance of 50 kilometers.

Our main meso-level data sources are the *Establishment History Panel* (BHP 7516) and *IEB GEO* from the Institute for Employment Research (IAB).[8] The *Establishment History Panel* comprises all German establishments with at least one employee on June 30 of each year. The dataset provides establishment-level

[7] One might be tempted to think that reversed causality also threatens identification on the intra-regional level. However, it does not seem plausible that high-skilled workers systematically sort into high-wage neighborhoods within regions. Instead, high-skilled workers might sort into high-wage firms. However, on the treatment level such a sorting process would not materialize into wages of neighboring firms and thus not reverse the direction of causality.

[8] For a detailed description of the Establishment History Panel, see Schmucker et al. (2016)

information on, among other metrics, the number of employees and the number of employees with tertiary education. To measure the distribution of high-skilled workers, we classify employees holding a degree from a university or a university of applied sciences as high skilled.[9]

We expand the dataset with exact geo-coordinates from IEB GEO. IEB GEO is a novel data source that includes addresses of establishments in the *Establishment History Panel* between 2000 and 2014 as geo-coordinates. In Germany, firms are obliged to register at least one of their establishments per municipality and industry. In general, the registration of one establishment per municipality provides a detailed description of the geographic landscape of workplaces. In some cases, however, firms might actually have multiple establishments within the same industry in a single municipality, which they do not report. In these cases, we cannot confirm that individuals work where they are registered. We therefore exclude the following chain-store industries from our data: construction, financial intermediation, public service, retail trade, temporary agency work and transportation. With the remaining set of establishments, we compute the density of high-skilled workers as spatial functions around establishments as described in section 3.2.3. However, as we only can exclude firms for which multiple establishments within one municipality are common, we can not assure that the remaining establishments are unique within one municipality. Consequently, especially in larger cities, it may happen that workers are registered at a headquarter while working at a different establishment in the same municipality. This can, potentially, result into a biased estimation of the spillover effect, which we, however, do not expect to be very significant. First, because we assume that these cases are relatively rare and, second, because such an effect should, at least to a large extent, be captured by worker-firm match fixed effects.

In the econometric analysis of human capital externalities, we merge the constructed spatial functions of high-skilled workers with micro-level data from the *Sample of Integrated Labour Market Biographies* (SIAB 7514).[10] The Sample of Integrated Labour Market Biographies is a 2% random sample of social security records. The dataset contains, among other data, information on wages, age, work experience and education with daily precision. To join the individual-level data to the establishment-level data we transform the spell dataset into a yearly

[9] There are two types of universities in the German tertiary education system: traditional universities and universities of applied sciences (*Fachhochschulen*). Compared to traditional universities, universities of applied sciences focus more on practical topics. Universities of applied science usually also have a stronger focus on engineering and technology. Both kinds of universities award bachelor's and master's degrees.

[10] For a detailed description of the Sample of Integrated Labour Market Biographies, see Antoni et al. (2016)

panel with June 30 as the reference date and link workers and firms with their unique firm identifiers.

Because employers face legal sanctions for misreporting, information on wages in German social security data is generally highly reliable. However, one limitation is that roughly 10% of earnings are right-censored at the social security maximum. Therefore, we impute top-coded wages following Dustmann et al. (2009) and Card et al. (2013) (see the Appendix for details). Further, we improve information on education following Fitzenberger et al. (2005) and restrict the sample to full-time workers aged between 18 and 64. As we are only interested in the effects on individuals in regular employment, we exclude apprentices, interns, marginally employed workers and trainees. The final dataset consists of 3,498,536 observations from 539,179 individuals between 2000 and 2014.

To assign workplaces to local labor markets, we use the *de facto* standard definition of local labor market areas in Germany from the Federal Ministry for Economic Affairs and Energy (BMWi). The goal in designating these local labor market areas is to design regions with strong internal commuter links but clear detachment from other areas. The construction is based on Kosfeld and Werner (2012), who use factor analysis on commuter flows to identify local labor market areas in Germany. The BMWi partitions Germany into 258 local labor market areas with an average radius of 21 kilometers. The size of these local labor market areas corresponds well to the findings of Manning and Petrongolo (2017), implying that 80% of the effects of local labor demand shocks are measurable within 20 kilometers. As a rule of thumb, the authors further suggest that treatment areas for labor demand shocks should be 2.5 times the median commute. In our case the rule of thumb would suggest 24 kilometers and is therefore close to the actual size of the labor market areas from the BMWi (Dauth and Haller, 2018, own calculations). Because labor market areas consist of multiple counties (*Stadt- und Landkreise*, NUTS-3), we complete our dataset with county-level indicators on population density, unemployment and number of hotel beds (as a proxy for amenities) from the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR).

3.3.2 Descriptive statistics

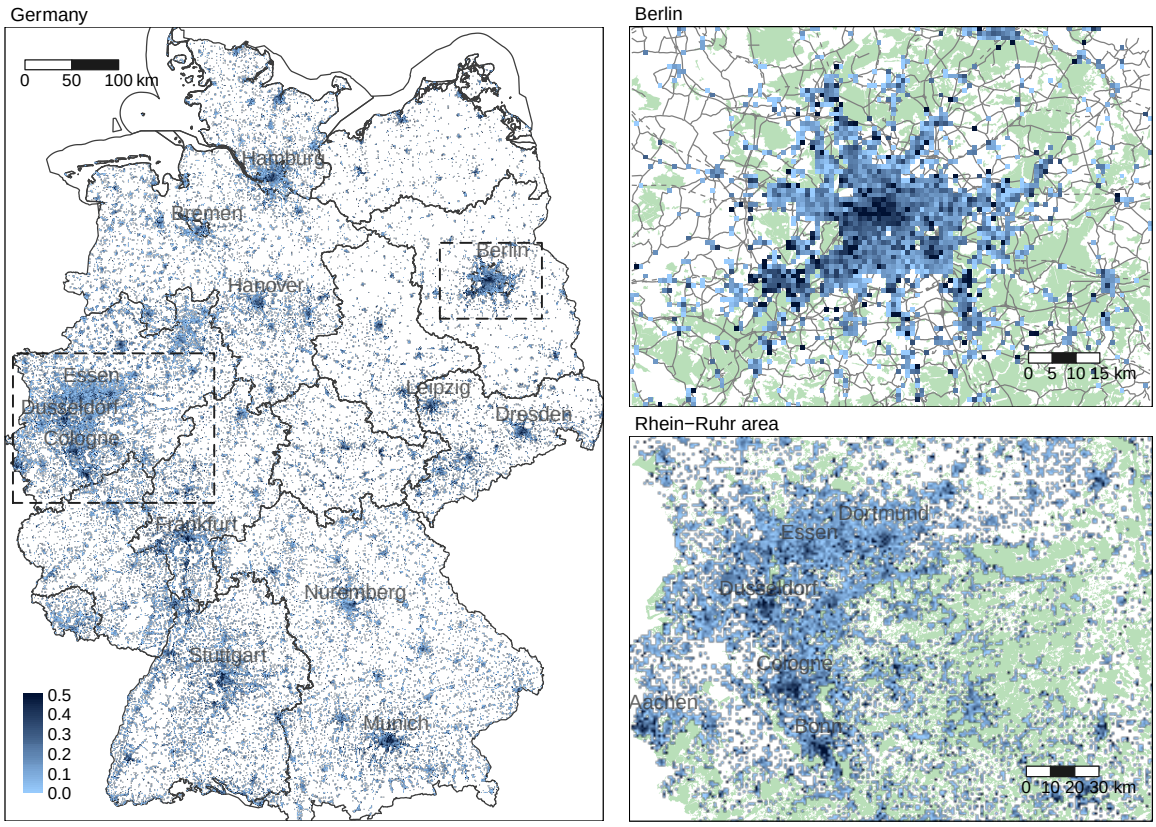
Figure 3.3.1 provides an overview of the distribution of high-skilled workers in Germany. For data protection reasons, the map shows the share of high-skilled workers in 1×1 kilometer grid cells. Note that the data used in the econometric analysis are more precise and offer exact coordinates. The map illustrates the considerable diversity in the distribution of high-skilled workers in Germany. For

instance, among the largest cities, there is a massive concentration of high-skilled workers in Munich, Hamburg and Berlin. By contrast, Nuremberg and Bremen exhibit significantly lower shares of high-skilled workers. Moreover, apart from metropolitan areas, there are several hot spots for skilled labor. For example, in Erlangen (15 kilometers north of Nuremberg), Darmstadt (25 kilometers south of Frankfurt) and Jena (70 kilometers south east of Leipzig) over 30% of full-time workers hold a degree from a university or university of applied sciences. Moreover, the distribution of high-skilled workers also varies considerably within administrative regions. The upper-right panel of Figure 3.3.1 shows a substantial cluster of high-skilled workers in the city center of Berlin. Additionally, there are several smaller clusters along the main traffic connections. The bottom-right panel focuses on the Rhein-Ruhr area. While high-skilled workers are evenly distributed in Essen and Dortmund, they appear to be very concentrated in the city centers of Düsseldorf, Cologne and Bonn. There are numerous small hot spots between the cities.

To capture the heterogeneous distribution of high-skilled workers, we compute a spatial function that relates the share of high-skilled workers to distance for each workplace in our data. Figure 3.3.2 illustrates the resulting curves. The light gray curves are 100 random examples and provide an impression of the variability in the data. The solid line shows the average share of high-skilled workers around establishments, and the dashed lines indicate the pointwise standard deviation around the mean. Although individual curves have strong variation, the average share of high-skilled workers around workplaces is stable in space. On average, the share of high-skilled workers is 17% in the direct neighborhood of establishments and gradually declines to 14.5% 50 kilometers away. The graph shows that there is no inherent distance at which the share of high-skilled workers suddenly falls. Instead, irregular city sizes and distances between settlements lead to a stable mean of the intensity of human capital over the whole domain. Note that the slight decline in the standard deviation is an artifact: The share of high-skilled workers within a distance window $[z_j - 500m, z_j)$ is the average of a binary variable, and since the absolute number of workers in $[z_j - 500m, z_j)$ increases with z, the uncertainty of the estimate for the mean of this binary variable becomes smaller with z. As the variation related to the estimation error becomes smaller, also the variance of the share declines with z. Also refer to appendix 3.A.1 for illustrative examples on the distribution of high-skilled workers around workplaces.

To obtain a first impression of the relationship between individual earnings and the spatial concentration of human capital, Figure 3.3.3 shows the correlation between log wages and the share of high-skilled workers within distance windows

Figure 3.3.1: Distribution of high-skilled workers in Germany

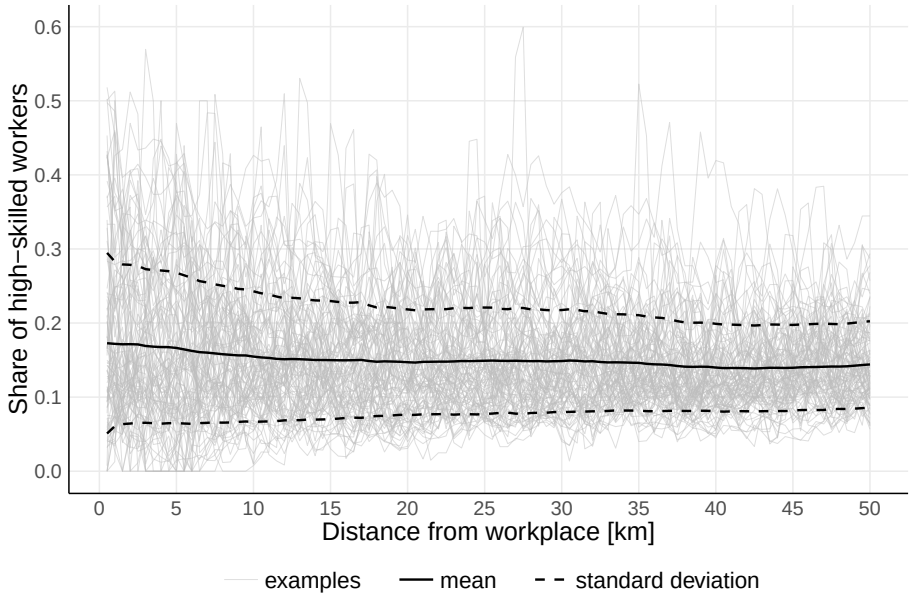


Notes: The figure depicts the share of high-skilled workers in 1×1 kilometer grid cells in Germany (left panel), Berlin (upper-right panel), and the Rhein-Ruhr area (bottom-right panel) in 2014. For data protection reasons, the maps depict aggregated data in grid cells. For the same reason, we removed cells with fewer than four establishments from the graphs. Note that the data for our statistical analysis are more precise and provide the exact coordinates of workplaces. Light blue cells indicate low shares of high-skilled workers, and dark cells signal high shares (see the scale at the bottom left). For the sake of clarity, values are capped at 50%. In the left panel, black lines depict the boundaries of federal states. In the right panels, green areas depict forests, and in the upper-right panel, gray lines and dashed gray lines illustrate streets and railways, respectively.

$[z_j - 500m, z_j), z_j = 500m, 1000m, ..., 50000m$. While the magnitude of the *ordinary* correlation has no direct interpretation, the declining trend signals that the relationship between income and the spatial concentration of high-skilled labor decays with distance.

One reason that the magnitude of the correlation coefficients has no direct interpretation is that the functions for the share of high-skilled workers are spatially autocorrelated. Figure 3.3.4 illustrates this issue. The graph depicts the

Figure 3.3.2: Spatial functions of the share of high-skilled workers

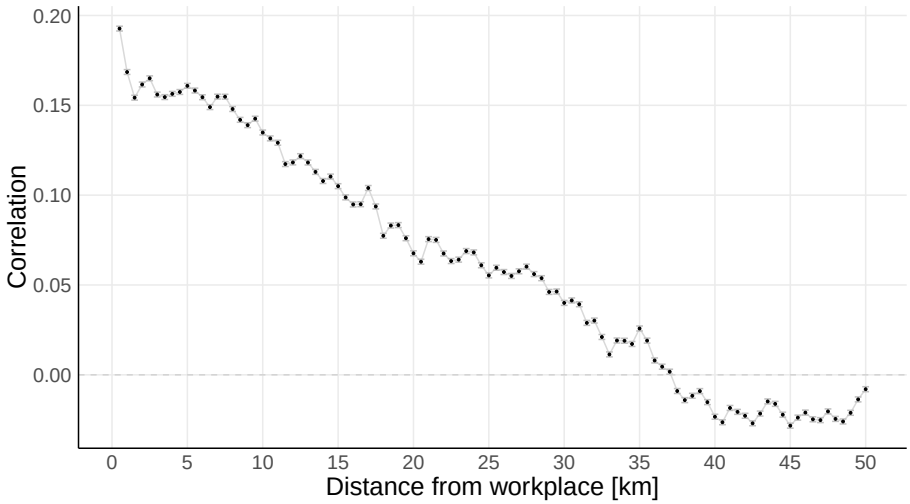


Notes: The figure shows the pointwise mean (solid line) and standard deviation (dashed lines) of the share of high-skilled workers around workplaces. Throughout the paper, we describe the share of high-skilled workers with spatial functions that map the share of high-skilled workers to the distance from a workplace. The graph also illustrates the variability of the spatial functions with 100 randomly selected curves (light gray lines). Each gray line depicts the spatial distribution of high-skilled workers around an establishment.

correlation between the share of high-skilled workers in three selected distance windows with the remaining 99 measurement points. For instance, the first panel presents the correlation of the share of high-skilled workers between measurement point t_1 and the random curve's value at $t_2, \ldots, t_{100}$. As the figure shows, adjacent values have a very high correlation compared to more distant measurement points. It is in principle possible to use all grid values of the functional predictor as regressors to measure the partial effect. However, as shown in the next section, the strong correlation between adjacent measurement points leads to a multicollinearity problem. As a consequence, the effects can not be measured any more.

For further summary statistics on individual wages and other covariates in our dataset, we refer to appendix 3.A.2.

Figure 3.3.3: Correlation of individual wages and the share of high-skilled workers around workplaces



Notes: The figure illustrates the correlation between log wages and the share of high-skilled workers within distance windows $[z_j - 500m, z_j], z_j = 500m, 1000m, ..., 50000m$. The graph suggests that the correlation between individual earnings and the intensity of human capital attenuates with distance. Note that the magnitude of the correlation coefficients cannot be interpreted directly.

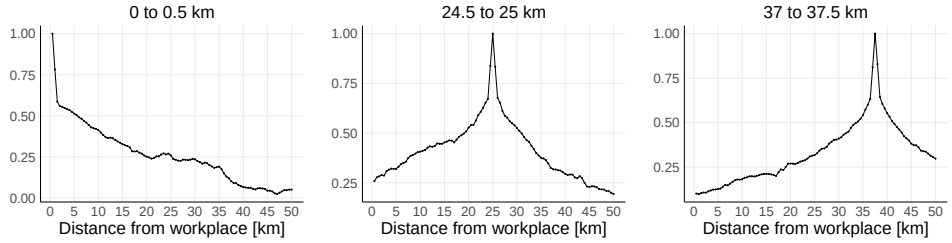
3.4 Results

Our main results show that spillover effects from the local concentration of high-skilled workers significantly increase individual wages. The spillover effects decay with distance, and the point estimates suggest that after 10 kilometers, the effects are reduced by half. Beyond 15 kilometers, the effects are no longer distinguishable from zero. In the following, we present the estimation results and discuss our findings. We also present effects on different skill groups. Next, we corroborate the robustness of our estimates with a simulation study. Finally, we summarize several additional robustness checks.

3.4.1 Main findings

We illustrate estimates of the spatial intensity of human capital externalities from high-skilled workers in Figure 3.4.1 and Figure 3.4.2. Figure 3.4.1 depicts an unrestricted estimate of equation (3.9) (i.e., $\rho = 0$ in equation (3.5)), which coincides with standard OLS regression. Figure 3.4.2 presents penalized estimates of equation 3.9 (i.e., $\rho > 0$). Both estimates control for labor market demand

Figure 3.3.4: Spatial autocorrelation at selected measurement points



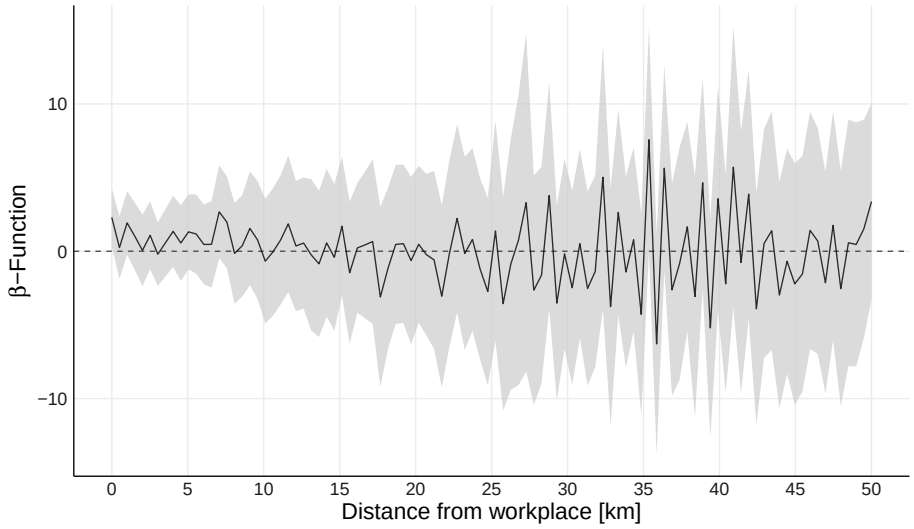
Notes: The graphs shows the spatial autocorrelation of the spatial functions of high-skilled workers at different measurement points. For instance, the panel in the middle shows the correlation of the share of high-skilled workers 24.5 to 25 kilometers away from workplaces with the share of high-skilled workers at the other 99 measurement points. The focal points in the remaining two panels are 0 to 0.5 and 37 to 37.5 kilometers, respectively. As is typical with functional data, values close to the focal point have high correlation. The correlation declines with distance from the focal point. Note that the three selected focal points well illustrate the general pattern of the underlying three-dimensional correlation function.

and supply effects and endogenous sorting of individuals with an extensive set of fixed effects. In addition to standard controls from the labor literature, our models include worker-firm match fixed effects and skill-specific yearly labor-market-area fixed effects. In the graphs, black lines display the estimated spillover functions. The gray area indicates the associated 99% confidence band. Note that OLS estimates of equation (3.9) would be mis-scaled by the number of discretization points of $X_{it}(z)$. By contrast, our estimates provide an approximation via a Riemann sum and are thus correctly scaled.

As Figure 3.4.1 shows, the unpenalized estimate of equation (3.9) identifies no significant link between the spatial concentration of high-skilled workers and individual earnings. The point estimates are very unstable, and the confidence bands include the null over the whole domain. There are two reasons for the unstable behavior of the curve. First, as described in the previous section, the measurement points of the share of high-skilled workers are highly correlated. Because the unrestricted estimator is (up to a scale) identical to the standard OLS estimator, high correlation among a large set of regressors poses multicollinearity problems. Consequently, the estimates exhibit high variance. Second, an unrestricted estimator allows one to compute unnecessarily complex functions and is therefore potentially prone to overfitting the data by modeling noise.

By contrast, the penalized estimates in Figure 3.4.2 reveal a clear influence of the spatial concentration of high-skilled workers on individual wages. The spatial spillover function depicted in the figure was obtained with 2.5 effective degrees of freedom. With such a specification, the estimate can be substantially

Figure 3.4.1: Unrestricted estimates of spatial human capital externalities from high-skilled workers

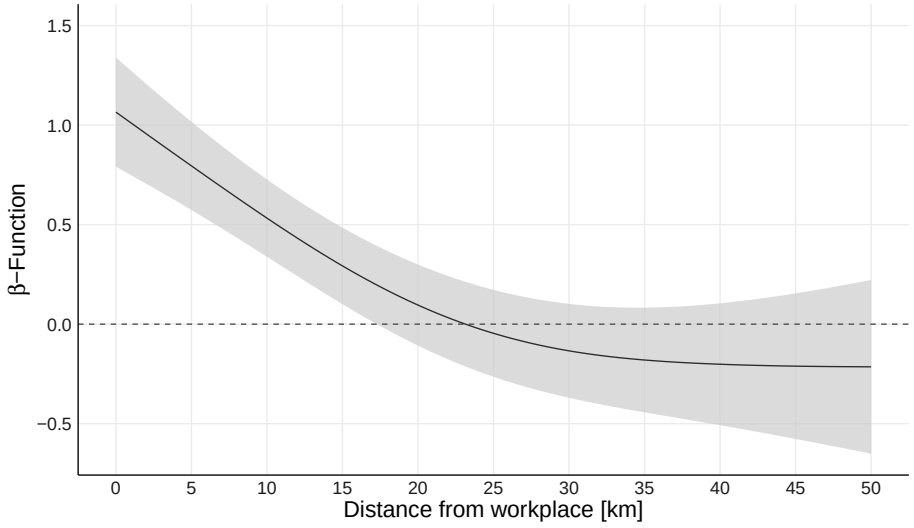


Notes: The figure presents an unrestricted estimation of spatial human capital externalities from high-skilled workers into individual log wages (equation (3.9)). We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The unrestricted estimator (equation (3.5), with $\rho = 0$) coincides with the standard OLS estimator. Due to multicollinearity and overfitting, the estimator cannot retrieve valid estimates of $\beta(z)$ from the data. The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 3,498,536$

more complex than a straight line. Estimates with more (fewer) effective degrees of freedom are qualitatively similar but are of course more (less) flexible (see appendix 3.A.3).

Our estimates in Figure 3.4.2 reveal economically significant spillover effects from the local concentration of high-skilled workers. The spillover effects decay with distance and vanish after approximately 15 kilometers. The magnitude of the effects from direct neighbors is roughly twice as large the size of effects from high-skilled workers located ten kilometers away. In the graph, the effect of a p-percentage-point increase in the share of high-skilled workers within distance z_j and $z_{j'}$ (in a 0 to 1 range), is p times the area bellow the estimated spillover function from z_j to $z_{j'}$. For instance, a 20-percentage-point increase in the concentration of high-skilled workers within 5 kilometers leads to wage gains of 1.75%

Figure 3.4.2: Spatial human capital externalities from high-skilled workers



Notes: The figure shows spatial human capital externalities from high-skilled workers into individual log wages. We measure the concentration of high-skilled workers as the share of high-skilled workers z units away from individuals workplaces. To compute the spatial spillover function ($\beta(z)$) we estimate equation (3.9) with the estimator (3.5). For this estimate, we restrict the β curve by setting $\text{edf}(\rho) = 2.5$ such that it is flexible enough to capture a monotonously decreasing function which can approach zero. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The graph shows significant spillover effects that decay with distance. The effect of a p-percentage-point increase in the share of high-skilled workers within distance z_0 and z_1 (in a 0 to 1 range) is p times the area bellow the estimated spillover function from z_0 to z_1. For instance, a 20-percentage-point increase in the concentration of high-skilled workers within 5 kilometers ($z_0 = 0, z_1 = \frac{5}{50}$) leads to wage gains of 1.75%. The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). Refer to Table 1 in the Online Appendix for a complete list of parameter estimates. $N = 3,498,536$

($\approx 20 \times \{0.75 \times \frac{5}{50} + \frac{1}{2}\left[(1 - 0.75) \times \frac{5}{50}\right]\}$). An evenly distributed ten-percentage-point (one standard deviation) increase in the share of high-skilled workers over the whole domain raises individual wages by 2% ($\approx 10 \times \frac{1}{2}\left(1 \times \frac{20}{50}\right)$). Reassuringly, *classical* estimates at an aggregate level, where we use OLS to model the wage effect of the share of high-skilled workers within counties and identical covariates as in equation (3.9), suggest effects of the same magnitude (see the Appendix).

Our results are also similar to the findings of Rosenthal and Strange (2008) for the US. The authors regress wages on the number of workers with a college

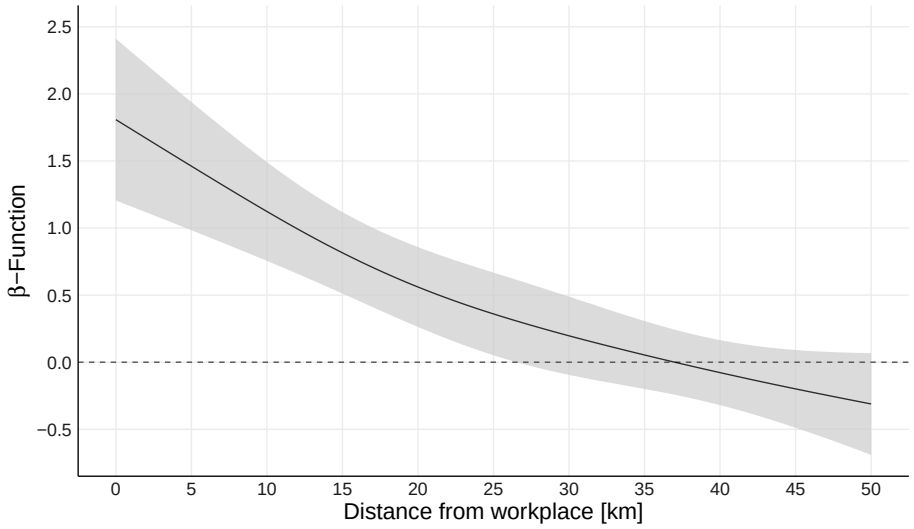
degree or higher education within 5 miles' distance and within 5 to 25 miles' distance. They report that spillovers from high-skilled workers within 5 miles' distance are up to 3.5 times larger than spillovers from high-skilled workers 5 to 25 miles away. Averaging our estimates within the same distance windows yields a ratio of 6. Although we follow a different estimation approach with different data, our findings seem to be consistent with those of Rosenthal and Strange (2008).

If we compare our results to findings from studies that analyze human capital externalities on administrative levels, our estimates are at the lower end of the range. The reason likely is our demanding battery of fixed effects, which lead to rather conservative estimates. Our results imply that an evenly distributed one-percentage-point increase in the share of high-skilled workers raises wages of other workers by 0.2%. With data of US metropolitan regions Moretti (2004) finds that a one-percentage-point increase in the share of college graduates raises wages between 0.4% and 1.2%.[11] With Russian survey data, Muravyev (2008) finds a 1.5%-response of wages. Based on the very similar data we use in our analysis, Heuermann (2011) estimates a wage reaction of 1.8% on highly qualified workers and 0.6% on other workers. In a dynamic context and using data from Sweden, Mellander et al. (2017) find a reaction of wages by 0.08% to a one-percentage-point increase of the share of high skilled in the local labor market.

Let us now briefly discuss the importance of removing demand and supply effects when estimating human capital externalities. Figure 3.4.3 reports estimates of our model (equation (3.9)) without skill-specific yearly labor-market-area fixed effects (π_{rst}) and thus includes labor market demand and supply effects that stem from imperfect substitution of high- and low-skilled labor (see Moretti, 2004; Ciccone and Peri, 2006). Compared to our main findings, the estimated relationship between individual wages and the concentration of high-skilled workers appears stronger in these estimates. Specifically, there is a global upward shift of the estimated $\beta(z)$ by, roughly, a factor of two. Although π_{rst} also nullifies other confounders (e.g., temporal effects from sorting of high-skilled workers), the uniform upward shift of $\beta(z)$ corresponds well to Ciccone and Peri (2006). They also find that the bias from demand and supply effects in Mincerian estimates of human capital externalities are large.

[11] Note that the following selection only includes studies with comparable measurements of human capital. Clearly, there are other important contributions using, for instance, average years of schooling to measure human capital.

Figure 3.4.3: Spurious estimates of spatial human capital externalities from high-skilled workers



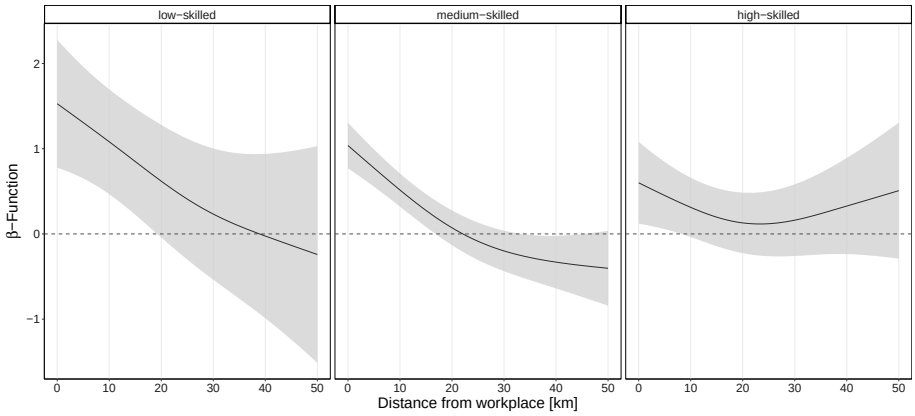
Notes: The figure presents estimates of the spatial human capital externalities from high-skilled workers into individual log wages without nullifying labor market demand and supply effects that stem from imperfect substitution of high- and low-skilled workers. Specifically, the graph depicts estimates of the spatial spillover function ($\beta(z)$) from equation (3.9) without skill-specific yearly labor-market-area fixed effects (π_{rst}). We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z and compute the model with the estimator (3.5). We restrict the capacity of the β curve to a parabola-like function that may remain flat over some interval, and we set the penalty parameter ρ accordingly. The black line illustrates the estimated spillover function ($\beta(z)$), and the light gray area indicates the 99% confidence band. The graph shows a significant relationship between the spatial concentration of high-skilled workers and wages. However, approximately half of the relationship is attributable to labor market supply and demand effects and other confounders. The underlying model controls for worker-firm match fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 3,498,536$

3.4.2 Different skill groups

Figure 3.4.4 presents estimation results for different skill groups. The panel on the left hand side shows estimated human capital externalities for workers without vocational training (low skilled), the middle panel for workers with completed vocational training (medium skilled), and the right panel for workers with a degree from a university or university of applied science (high skilled).

The estimates are in line with our main findings in all three cases. For all skill groups, the overall effect is significantly positive and decreases with distance. As

Figure 3.4.4: Spatial human capital externalities from high-skilled workers for different skill groups



Notes: The figure shows spatial human capital externalities from high-skilled workers into individual log wages for different skill groups based on sample-splits. We measure the concentration of high-skilled workers as the share of high-skilled workers z units away from individuals workplaces. To compute the spatial spillover function ($\beta(z)$) we estimate equation (3.9) with the estimator (3.5). For this estimate, we restrict the β curve by setting $\text{edf}(\rho) = 2.5$ such that it is flexible enough to capture a monotonously decreasing function which can approach zero. The black lines illustrates the estimated spillover function ($\beta(z)$) for each skill group, and the gray area indicates the 99% confidence band respectively. The graph shows significant spillover effects that decay with distance for all skill levels. The estimates even suggest that groups with a lower skill level are affected more strongly than those with a higher skill level. However, these differences are not statistically significant as the confidence bands overlap each other.
The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). The splitted samples have size $N = 663,661$ (high-skilled), $N = 2,552,942$ (medium-skilled), and $N = 281,993$ (low-skilled).

in the pooled case, the spillover effect can be measured up to 15 kilometers, at least for low- and medium-skilled workers. For the group of high-skilled workers, the confidence bands are large because the according sample size is relatively small. The estimated functional coefficients suggest that the effects are even stronger for lower skilled groups than for high-skilled workers. However, these differences are not statistically significant as the corresponding confidence bands have a large overlap. Overall we conclude that there are no systematic heterogeneities of the spillover effect with respect to skills.

3.4.3 Simulation study

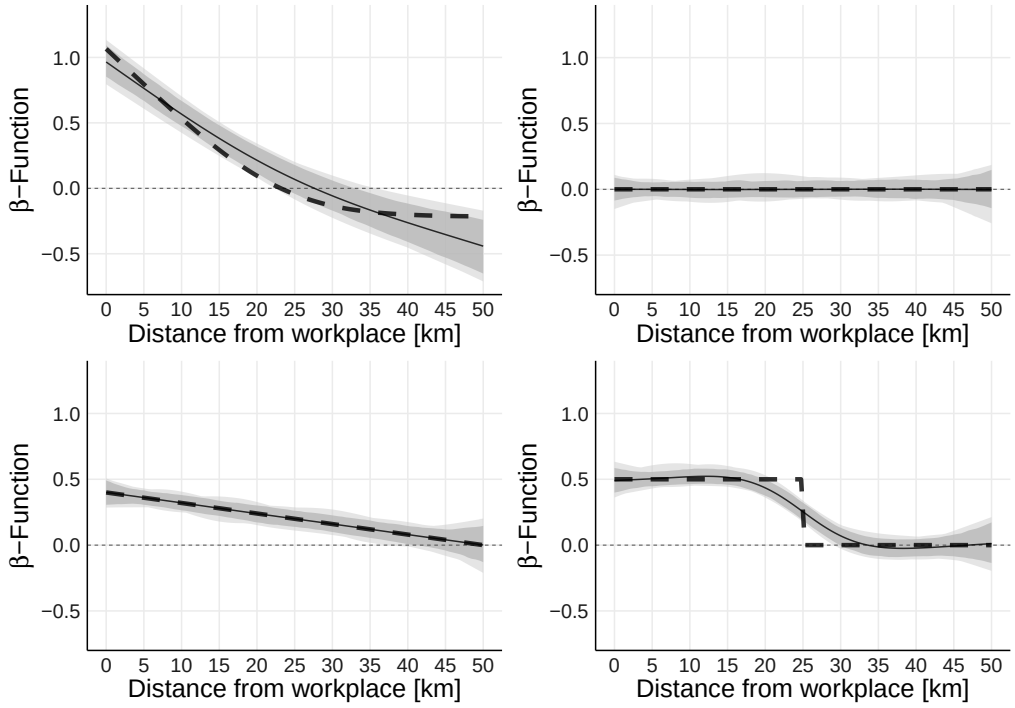
As outlined in section 3.2.2, drawing local inference about the function-valued parameter β is difficult. The following simulation exercise, therefore, is intended to evaluate the statistical properties of our estimation framework. The results show that our estimation framework, although yielding locally biased estimates, is reliable in the sense that it is able to reproduce the structure of the true curve well. We also show that the inference procedure controls size when the null is a linear function.

In the simulation study, we consider four scenarios. First, we evaluate the estimator's properties in a situation where the data generating process (DGP) resembles the particular real-world problem. Therefore, we take the DGP from the preferred estimate (Figure 3.4.2). We also incorporate parameter estimates from all covariates and generate artificial observations of the dependent variable based on iid errors that are drawn from $N(0, \widehat{\sigma}_u^2)$. Here, $\widehat{\sigma}_u$ denotes the standard error of the residuals of the estimated model. The structure of the simulated dataset (e.g., sample size, number of firms, number of workers per firm), therefore, is the same as in the original sample. The remaining three scenarios assess the statistical properties of the estimator in different extreme situations. Here, we simulate data that have a similar structure as the real dataset. In particular, we replicate the first two moments of the original data.[12] The second and third scenarios evaluate the accuracy of the inference procedure when the null is the zero function or a linear function. The fourth and most extreme setting analyzes the performance of the estimator when the true parameter is a non-smooth step function. To asses the statistical properties of the estimator, we simulate 1000 replications in each scenario.

Figure 3.4.5 summarizes the results of the four simulations. In each panel, the bold dashed line depicts the true parameter function $\beta_0(z)$ of the DGP, the light gray areas show pointwise minimum and maximum of all estimates, and the dark gray areas show the first and the 99^{th} percentiles of all estimates of the parameter function. The solid line represents the pointwise mean over all replications. In general, the estimates follow the true parameter function well, and no replication deviates substantially from the DGP. However, as is typical for penalized (or nonparametric) models, the estimates deviate from the true curve in regions with complex structure (i.e., in regions with strong nonlinearities). In such regions, the estimator possesses a local bias. As one might expect, this behavior is especially pronounced at the jump discontinuity of the step function in the bottom-right

[12] To replicate this part of the simulation study, refer to the code in the online supplement of this article.

Figure 3.4.5: Performance of the estimator in different simulations



Notes: The figure shows four Monte-Carlo simulations. The bold dashed line depicts the true parameter function $\beta_0(z)$, the light gray areas show pointwise minimum and maximum of all estimates, and the dark gray areas show the first and 99th percentile of all estimates of the parameter function. The solid line represents the pointwise mean over all replications. Simulated replications of the estimator were obtained by estimating model (3.9) based on simulated data. The setup corresponding to the top-left panel uses the predictors from the real-data application, and observations of the dependent variable are simulated based on estimated coefficients and iid normally distributed errors. All other setups are based solely on simulated data that mimic the original sample but use different specifications for the functional parameter $\beta(z)$. In the top-right panel $\beta(z) = 0$, bottom-left: $\beta(z) = 0.4(1-z)$ and bottom-right $\beta(z) = 0.5 \cdot \mathbb{1}(z < 0.5)$.

panel of Figure 3.4.5. By construction, however, the smoothing splines estimator never produces estimates different from zero in regions where the true curve is zero in a larger neighborhood. Therefore, if the underlying functional shape of the spatial decay of human capital externalities is monotonically decreasing and zero beyond a certain distance, the regularized estimation captures the true curve well. This appears to be a reasonable assumption in our application.

Table 3.4.1 provides the integrated squared bias, integrated variance, and the coverage probability of the confidence bands for each scenario. The integrated (squared) bias is largest for the setup in which the function-valued parameter is taken from the real-data application because the true parameter is curved over

Table 3.4.1: Performance measurements in different simulations

	Specification for β_0			
	I	II	III	IV
Integrated squared bias	0.0096	0.0000	0.0000	0.0055
Integrated variance	0.0030	0.0009	0.0009	0.0010
Coverage probability of 99%-CIs	0.7290	0.9920	0.9930	0.0000

Notes: The table contains integrated variance, integrated squared bias and the coverage probability of confidence bands of the parameter estimate for the functional coefficient for all four setups considered in the simulation exercise. In the first setup, the data were generated based on the regressors and functional predictors with corresponding coefficients taken from the original estimate. The other setups are based solely on simulated data but with similar characteristics. In setup II, the functional coefficient of the DGP is zero; in setup III it is a linear function. The coefficient in the last setup (column IV) is discontinuous and possesses a discrete jump in the interior of its domain. We compute integrated variance as $1000^{-1} \int \sum_{r=1}^{1000} \left(\widehat{\beta}_r(z) - \overline{\beta}(z)\right)^2 \mathrm{d}z$ and integrated squared bias as $\int \left(\overline{\beta}_r(z) - \beta_0(z)\right)^2 \mathrm{d}z$, where $\overline{\beta}(z) = 1000^{-1} \sum_{r=1}^{1000} \widehat{\beta}_r(z)$.

the whole domain (column 1). Similarly, the variance is the largest in this setup. The two scenarios with linear parameter functions, by the construction of the estimator, show favorable properties and exhibit the lowest variance and no bias (columns 2 and 3). In this situation, confidence bands based on equation (3.7) have proper coverage probability which, however, no longer holds with more complex parameter functions. In the most extreme case (discontinuous β_0), the bias at the jump discontinuity is so large that the confidence bands are unable to cover the true parameter over the whole domain (column 4).

The implications from the simulation study for our main findings are as follows. If the true spatial decay of human capital externalities is not too complex, our estimates and confidence bands are generally reliable. However, because the estimator is locally biased in regions with a more complex β_0, identifying the exact distance at which human capital externalities cease is difficult. A conservative strategy would be to choose a threshold somewhat lower than indicated by the confidence bands. Regarding our main findings, such a strategy suggests that human capital externalities might already be statistically insignificant after 15 kilometers.

3.4.4 Semi-parametric OLS estimates with broader rings

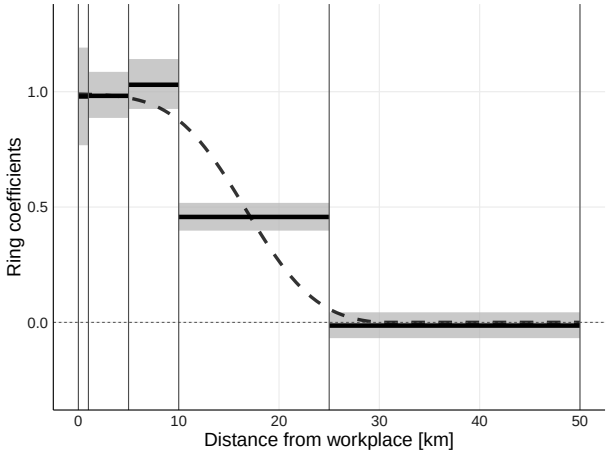
The previous literature that measures the spatial attenuation of economic effects uses a semi-parametric framework, in which the main explanatory variable is measured in a series of concentric rings or circles. The outcome variable is then regressed on the series of measurements (e.g., Rosenthal and Strange, 2008; Fu, 2007; Verstraten, 2018; Gibbons et al., 2021; Faggio et al., 2019; Faggio, 2019). The beauty of the semi-parametric framework is that it is a straightforward application of the linear OLS model and in principle can be applied to any geographical data. The drawback of the semi-parametric framework compared to our FDA approach is that estimates of the spatial attenuation of effects are less precise. The reason is that multicollinearity issues (usually) do not allow to estimate effects from a large or fine-graded series of measurements. To circumvent multicollinearity issues, researchers construct relatively broad rings or circles that measure the spatial distribution of the explanatory variable. We corroborate our main findings by applying the semi-parametric framework to our research question. Specifically, we estimate the effects from the shares of high-skilled workers in 0-1, 1-5, 5-10, 10-25 and 25-50 kilometers distance on log wages using OLS. Albeit less precise, the estimated effects are of similar magnitude as our main findings and support our procedure.

Before explaining the corresponding econometric specification, let us briefly discuss the properties of the semi-parametric approach by means of a small simulation exercise. To this end, we generate 1000 replications of the DGP (3.1) using predictors resembling first and second moments of our real data application. The functional coefficient β_0 corresponds to the dashed line of Figure 3.4.6. We then compute averages of the simulated curves with respect to larger intervals of the domain.[13] We obtain the spillover parameters by regressing the (simulated) dependent variable on these averages and normalizing the respective coefficient with the ring's width. The aggregation scheme is equivalent to the one used in (3.10).

In Figure 3.4.6 we illustrate the results of the simulation study. The coefficient function of the DGP is depicted by the dashed line and vertical solid lines indicate boundaries of the rings used in our specification. The grey areas illustrate first and 99^{th} percentiles of all replications and the horizontal black lines represent the mean over all replications. In general, the results show that the approximation via a Riemann sum works also quite well but the outcome heavily depends on how

[13] By aggregating the curves in such a manner, the resulting *rings* no longer reflect shares of high-skilled workers in a particular ring but a weighted average where, assuming a uniformly populated area, the more central observation get a larger weight compared to the more distant observations in each ring. In our real data application, we are of course able to compute the shares of high-skilled workers in the distance windows.

Figure 3.4.6: Simulation results of semi-parametric OLS estimates



Notes: The figure shows a Monte-Carlo simulation for the semi-parametric OLS estimation. The bold dashed line depicts the true parameter function $\beta_0(z)$. The vertical solid lines depict the boundaries of the rings and the horizontal black lines illustrate the mean over all replications of the approximation of the functional coefficient via a Riemann sum. The grey areas reflect the range between 1st and 99th percentile of all estimated coefficients of the Riemann sum. The Riemann sum coefficients are obtained by dividing the raw regression coefficient of the aggregated rings by the ring's width. Simulated replications were obtained by estimating model (3.10) on data generated by DGP (3.1) but with the same predictors used in the Monte-Carlo exercise described in section 3.4.3.

the rings are defined. In addition, such an estimation framework does not allow to learn from the data, how the coefficient function behaves inside the intervals.

Now, let us compare our estimates to the semi-parametric approach. To this end, we estimate the following model:

$$\begin{aligned} Y_{it} = {} & \alpha_1 x_{1\text{km},it} + \alpha_2 x_{5\text{km},it} + \alpha_3 x_{10\text{km},it} + \alpha_4 x_{25\text{km},it} + \alpha_5 x_{50\text{km},it} \\ & + Z_{it}'\gamma + \theta_{if} + \tau_t + \omega_o + \pi_{rst} + u_{it}. \end{aligned} \tag{3.10}$$

Here, Y_{it} is the individual log wage of worker i in year t. $x_{1\text{km}}$ is the share of high-skilled workers within 0 to 1 km distance of i's workplace, $x_{5\text{km}}$ is the share of high-skilled workers within 1 to 5 km distance of i's workplace, $x_{10\text{km}}$ is the share of high-skilled workers within 5 to 10 km distance of i's workplace and so on. Accordingly, $\alpha_1, \ldots, \alpha_5$ are the spillover coefficient we seek to estimate. In line with our main model, we control for time-varying observable characteristics of individuals, establishments and regions (Z_{it}) and a series of fixed effects. θ_{if} is a worker-firm match fixed effect, π_{rst} is a skill-specific yearly labor-market-area fixed effect, τ_t is a year fixed effect, and ω_o is an occupation fixed effect.

Table 3.4.2: Semi-parametric OLS estimates with broader rings

	raw		per km	
	(1)	**(2)**	**(3)**	**(4)**
Share of high-shilled workers in ...				
0–1km	0.050***	0.030***	0.050***	0.030***
	(0.003)	(0.003)		
1–5km	0.074***	0.070***	0.018***	0.017***
	(0.005)	(0.007)		
5–10km	0.078***	0.089***	0.016***	0.018***
	(0.006)	(0.010)		
10–25km	0.085***	−0.051**	0.006***	−0.003**
	(0.009)	(0.019)		
25–50km	0.004	−0.052	0.000	0.000
	(0.013)	(0.028)		
Worker-firm match fixed effects	Yes	Yes	Yes	Yes
Labor-market-area × year × skill fixed effects	No	Yes	No	Yes

Notes: The table summarizes estimates of the human capital externalities from high-skilled workers in broad concentric rings into individual log wages. The estimates replicate our main model in a less precise manner and serve as a comparison of the magnitude of the effects. The first two columns show raw coefficient estimates. Columns three and four show estimated effects within one kilometer bands. The underlying models further control for occupation fixed effects, time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). Cluster-robust standard errors are in parentheses.*** indicates significance at the 0.1%-level. $N = 3,498,536$

Table 3.4.2 summarizes the results. Columns 1 and 2 of table 3.4.2 show the strength of human capital externalities from five different distances (i.e., 0-1km, 1-5km, 5-10km, 10-25km and 25-50km). The effects are statistically significant up to the ring covering a distance of 10 to 25 kilometers.

Due to different bandwidths we cannot directly compare the magnitude of the raw estimates. As an illustration, consider that the parameter estimate on the first ring measures wage effects from a one-percentage-point increase in the share of high-skilled workers within one kilometer around individuals. The parameter estimate on the second ring expresses the effects of an one-percentage-point increase in one to five kilometers distance. Both estimates implicitly assume that

the one-percentage-point increase in the share of high-skilled workers is uniformly distributed within each bandwidth (i.e., the share of high-skilled workers increases by one-percentage-point in each kilometer). Thus, by construction, the second ring captures a treatment that is five times stronger than the first ring does. To make the parameter estimates comparable across rings we therefore divide the raw estimates by their underlying bandwidth in column 4. The according numbers give the effect of a one-percentage-point increase in the share of high-skilled workers within one kilometer within a certain bandwidth.

In line with our main findings columns 3 and 4 show that human capital externalities decay with distance. Also similar to our main findings, human capital externalities loose their economic significance between 10 to 25 kilometers distance. Also the magnitude of the estimated effects are similar to those of our main model. For instance, according to our main model a 20-percentage-points increase in the share of high-skilled workers within five kilometers leads to wage gains of 1.75%. According to our semi-parametric estimates with broader rings the same increase in the share of high-skilled workers raises wages by 2%. The difference between the two estimates is minor. In summary, the semi-parametric estimates buttress our main findings.

3.4.5 Further robustness checks

We have undertaken several robustness checks which can be found in the Appendix. In this section, we briefly summarize the results of these exercises.

As the data source is based on register data from the German social security system, information on high-skilled workers outside of Germany is not available. Consequently, in border regions, we construct our measure of the spatial concentration of human capital with partly truncated information. However, excluding border regions from our model yields similar results to our main findings. We conclude that truncated information from border regions does not affect our results.

Another concern may be that global labor market shocks influence our findings through local industry or occupation clusters. If, for instance, wages and the demand for skilled labor temporarily rise within a sector and firms in this sector tend to cluster locally, our estimates would capture a spurious relation between wages and the local concentration of high-skilled workers. To rebut these concerns, we augment our model with year-specific industry and occupation fixed effects. Reassuringly, absorbing industry and occupation trends does not affect our results.

Following Cornelissen et al. (2017), who identify human capital externalities

in the workplace, we additionally corroborate our findings with a placebo test, in which we expand our model with a one-year lead of the spatial distribution of high-skilled workers. Because workers cannot receive spillovers from neighbors who have not yet moved in, the future concentration of high-skilled workers serves as a placebo. The test indicates that the future concentration of high-skilled workers is almost unrelated to wages. Moreover, estimates of the human capital externalities from the current share of high-skilled workers change only slightly compared to our baseline specification.

Plausibly, the strength of human capital externalities differs in urban and rural areas. We therefore separately estimate our model in urban and rural areas. The according estimates imply that human capital externalities are considerably stronger in urban areas than in rural areas. In fact, we find only weak evidence for human capital externalities in rural areas. We therefore conclude that our main findings are mostly driven by urban areas. Finally, estimating the spillover effect on the county-level, we find that the magnitude of the effects from our functional model is close to comparable estimates at the county level.

3.5 Conclusions

This paper studies the impact of human capital externalities from the regional concentration of high-skilled workers into the individual wages of neighboring workers. We use, for the first time, precise geocoded register data of an entire economy and a novel estimation method from the field of functional data analysis (FDA) to compute the spatial decay of human capital externalities. We find significant spillover effects from the local concentration of high-skilled workers that attenuate with distance. Human capital externalities from the direct neighborhood of firms are roughly twice as large as those from high-skilled workers that are located 10 kilometers away. After 15 kilometers, the effects vanish. Overall, an evenly distributed one-standard-deviation increase in the local share of high-skilled workers leads to wage gains of 2%.

Two developments in modern social science are primarily responsible for our ability to derive a precise functional relationship between the concentration of high-skilled workers and individual earnings. First, the availability of exact geospatial data enables us to describe the distribution of high-skilled workers around workplaces as functional objects with high resolution. Specifically, we evaluate the concentration of high-skilled workers every 500 meters within a radius of 50 kilometers around almost all establishments in Germany. Second, FDA provides tools to fully exploit such detailed data. We employ the estimator of Crambes et al. (2009) to regress a scalar outcome (log wage) on a continuous

functional variable (the concentration of high-skilled workers depending on distance). Our application illustrates the potential of FDA in economic research. FDA is particularly beneficial when the variable of interest can be regarded as a function over some continuum.

Generally, our findings imply that education creates positive externalities in local labor markets. Thus, regions benefit from attracting and training skilled workers. Moreover, to maximize these external effects, firms should settle close to one another. Although spillover effects cover entire cities, workers and firms benefit most from the skill distribution in their near neighborhood. Because the effects vanish after 15 kilometers, firms in remote regions do not gain from human capital externalities. Overall, our findings support Rosenthal and Strange (2008), who argue that the physical concentration of human capital remains important for economic development. Among other agglomeration effects, human capital externalities help to explain differences in productivity between densely populated cities and rural areas.

3.A Appendix

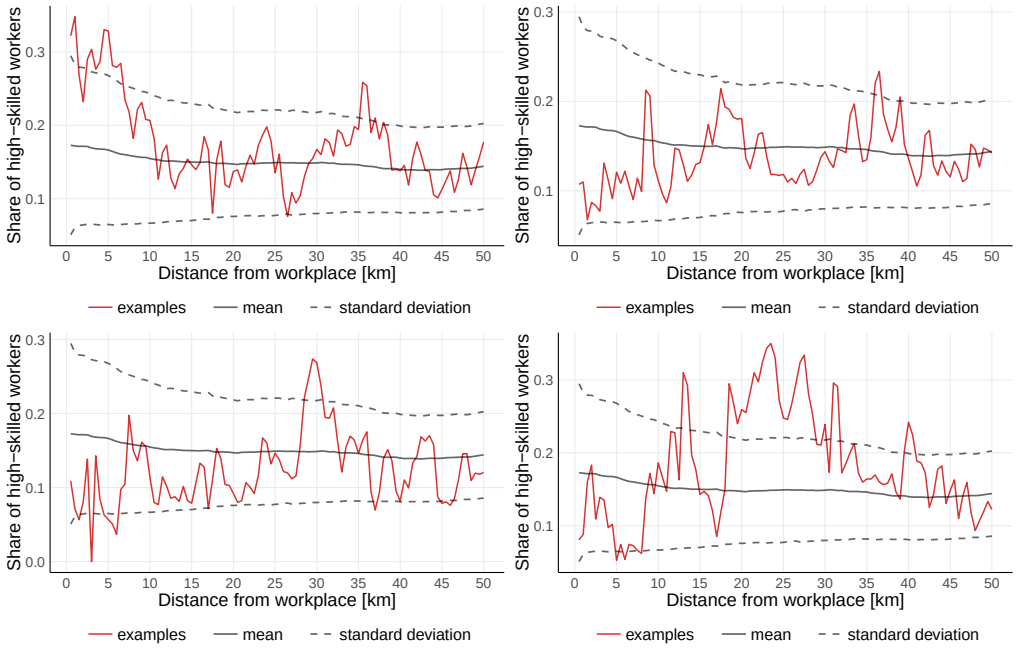
3.A.1 Examples of spatial functions of high-skilled workers

In the paper, we describe the distribution of high-skilled workers as continuous curves. More precisely, we define spatial functions that map the share of high-skilled workers to the distance from the workplace. To illustrate these functional objects, Figure 3.A.1 provides four randomly drawn examples. In each of the four graphs, red lines represent the share of high-skilled workers around an establishment. The light gray lines in the background indicate the pointwise mean and standard deviation in our dataset. For instance, in the first panel, we observe a high concentration of skilled labor of 30% in the near neighborhood of the workplace. Between 5 and 15 kilometers' distance, the share of high-skilled workers declines to 15%. After a decline around 25 kilometers away from the workplace, the share of high-skilled workers increases again. At the end of the domain,the share of high-skilled workers is approximately 15%. The remaining three panels illustrate different patterns.

3.A.2 Summary statistics

The dataset used in our econometric analysis covers 15 years and consists of 3.5 million records of 540,000 workers. Table 3.A.1 summarizes the dependent variable (log wage) and numerical control variables. In the dataset, the mean daily wage is 111 euros, and the first and second quartile range from 68 to 129 euros.

Figure 3.A.1: Examples of spatial functions of the share of high-skilled workers



Notes: The figure shows the distribution of high-skilled workers around four randomly drawn workplaces (red lines). The light gray lines indicate the pointwise mean and standard deviation of the share of high-skilled workers in the dataset. Throughout the paper, we describe the share of high-skilled workers as spatial functions that map the share of high-skilled workers to the distance from a workplace.

On average, individual are 41 years old and have 15 years of work experience. The median population density is 119 inhabitants per square kilometer (exp(4.78)). Furthermore, 36% of the observations are from females and 7% are from workers with foreign nationality. The shares of low-, medium- and high-skilled workers are 8%, 73% and 19%, respectively.

3.A.3 Estimates with different penalties

In our preferred specification, we estimate equation (3.9) with the estimator (3.5) and a penalty ρ that corresponds to 2.5 degrees of freedom, which restricts estimates of the spillover curve $\beta(z)$ to smooth parabola-like functions that may remain flat over some interval. To demonstrate the behavior of the estimator with different penalties, Figure 3.A.1 reports estimates with alternative values of ρ. Panels A and B allow for more flexible curves than our preferred specification, panel C repeats our preferred specification, and panel D restricts $\beta(z)$ to a linear function. Qualitatively, all models lead to similar results. The response of

Table 3.A.1: Summary statistics

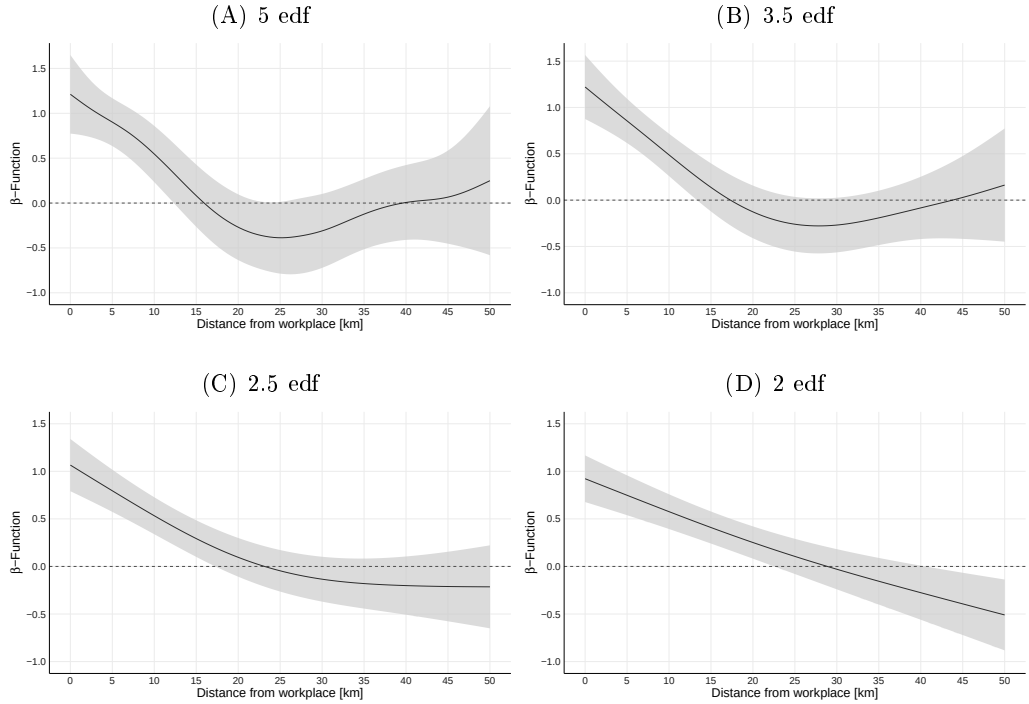
	Mean	Std. Dev.	25^{th} Perc.	Median	75^{th} Perc.
daily wage	111.37	78.05	68.17	94.64	129.02
daily log wage	4.55	0.56	4.22	4.55	4.86
age	41.14	10.65	33.00	41.00	49.00
work experience (days)	5528.31	3305.44	2860.00	5105.00	7974.00
tenure (days)	3059.98	2796.97	883.00	2160.00	4398.00
log firm size	4.68	2.10	3.14	4.63	6.10
log population density	3.71	2.38	0.97	4.78	5.66
log hotel beds	3.16	0.70	2.68	3.14	3.53
unemployment rate	8.74	4.11	5.60	7.90	11.00

Notes: The table presents summary statistics of wages and (numerical) control variables. The underlying dataset contains 3,498,536 observations of 539,179 individuals over a period of 15 years. Regional characteristics come from 402 counties.

individual wages to an increase in the share of high-skilled workers in the direct neighborhood is close to unity. When we reach 10 kilometers from the workplace, the effects are only approximately half the size. In all models, the spillovers become statistically insignificant after 13 to 23 kilometers. The confidence bands of the four estimates overlap on the whole domain.

However, depending on the hyperparameter ρ, the estimates of the spillover function are of course more or less flexible. Up to 20 kilometers' distance, the more volatile models in panels A and B are similar to our preferred specification and suggest that human capital externalities decline with distance. After 20 kilometers, however, the point estimates increase. Statistically, the rise at the end of the domain is accompanied by broad confidence bands. Thus, these estimates are imprecise. Moreover, it seems economically implausible that the intensity of human capital externalities follows a U-shaped pattern. Therefore, we regard the estimates from panels A and B as overly flexible. By contrast, the curve in panel D is forced to be linear. Again, up to 20 kilometers away from the workplace, the estimates are similar to our preferred model. Farther away, the point estimates diverge from our preferred specification and proceed to decline even after intersecting the abscissa. Similar to panels A and B, these estimates are less precise at the end of the domain. Moreover, theoretically, it seems implausible that human capital externalities follow a linear function. Thus, we regard the estimated spillover function from panel D as overly inflexible.

Figure 3.A.1: Estimates of spatial human capital externalities with different penalties



Notes: The figure shows estimates of the spatial human capital externalities from high-skilled workers into individual log wages based on four different penalty parameters. To compute the spatial spillover function ($\beta(z)$), we estimate equation (3.9) with the estimator (3.5). Each panel summarizes estimates with a different penalty ρ. The different penalty terms correspond to 5 (top left panel), 3.5 (top right panel), 2.5 (bottom left panel) and 2 (bottom right panel) effective degrees of freedom. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 3,498,536$

3.A.4 Imputation of wages

A common limitation of social security data is the right-censoring of earnings. To address this issue, we follow Dustmann et al. (2009) and Card et al. (2013) and impute censored wages with a two-step procedure.

In the first step, we group observations by year, East and West Germany, and three levels of education (i.e., no vocational training, vocational training and degree from a university or university of applied science). Within each group,

we fit a Tobit model with the following list of explanatory variables: age, age^2, tenure, tenure^2, work experience, (work experience)^2, firm size, and indicators for gender, being older than 40 years and being foreign born. Additionally, we include interaction terms of age and age^2 with the indicator variable *older than 40*. At the county level, we further include the predictors population density, the unemployment rate, the number of hotel beds and the share of high-skilled workers. With the parameters from the Tobit estimates ($\hat{\zeta}$), we impute wages by $X\hat{\zeta} + \hat{\sigma}\Phi^{-1}\left[k + u(1-k)\right]$, where $\hat{\sigma}$ is the estimated standard error of the regression, Φ is the standard normal density, u is a random value from a uniform distribution between zero and one, $k = \Phi\left[(c - X\hat{\zeta})/\hat{\sigma}\right]$ and c is the censoring point.

In the second step, we compute the lifetime average wages of each worker and firm, excluding the focal period. For workers and firms with only one observation, we assign the sample mean. With the period-specific lifetime average wages as additional predictors, we repeat the Tobit estimates. Finally, we impute censored wages by $X\hat{\zeta} + \hat{\sigma}\Phi^{-1}\left[k + u(1-k)\right]$.

3.A.5 Estimates of spatial human capital externalities: full table

Table 3.A.1 presents parameter estimates from our preferred specification and accompanies Figure 6 of our manuscript. In accordance with Figure 6, the table shows strong human capital externalities from high-skilled workers from nearby areas. The effects decay with distance and become statistically insignificant after 17 to 18 kilometers. The parameter estimates of worker characteristics are in line with the labor literature. Due to the extensive set of fixed effects in the model (Equation (9)), the parameter estimates for county-level variables are statistically insignificant.

3.A.6 County-level effects

In our paper, we model the distribution of high-skilled workers as continuous curves around workplaces and estimate human capital externalities with a functional regression model based on Crambes et al. (2009). To evaluate the magnitude of our results, let us now estimate a *classical* OLS model, in which we estimate spillovers from high-skilled workers at an aggregate level. Specifically, we calculate spillovers from the share of high-skilled workers within counties (NUTS-3, *Landkreise* and *kreisfreie Städte*). Apart from this, our estimation equation is identical to our main model (see also Equation (9) of the paper):

$$Y_{it} = \alpha x_{it} + Z'_{it}\gamma + \theta_{if} + \tau_t + \omega_o + \pi_{rst} + u_{it}. \tag{3.11}$$

Y_{it} is the individual log wage of worker i in year t, and x_{it} is the share of high-skilled workers within the county of i's workplace. Accordingly, α is the spillover coefficient we seek to measure. Identical to our main specification, the model controls for time-varying observable characteristics of individuals, establishments and regions (Z_{it}) and a series of fixed effects. θ_{if} is a worker-firm match fixed effect, π_{rst} is a skill-specific yearly labor-market-area fixed effect, τ_t is a year fixed effect, and ω_o is an occupation fixed effect.

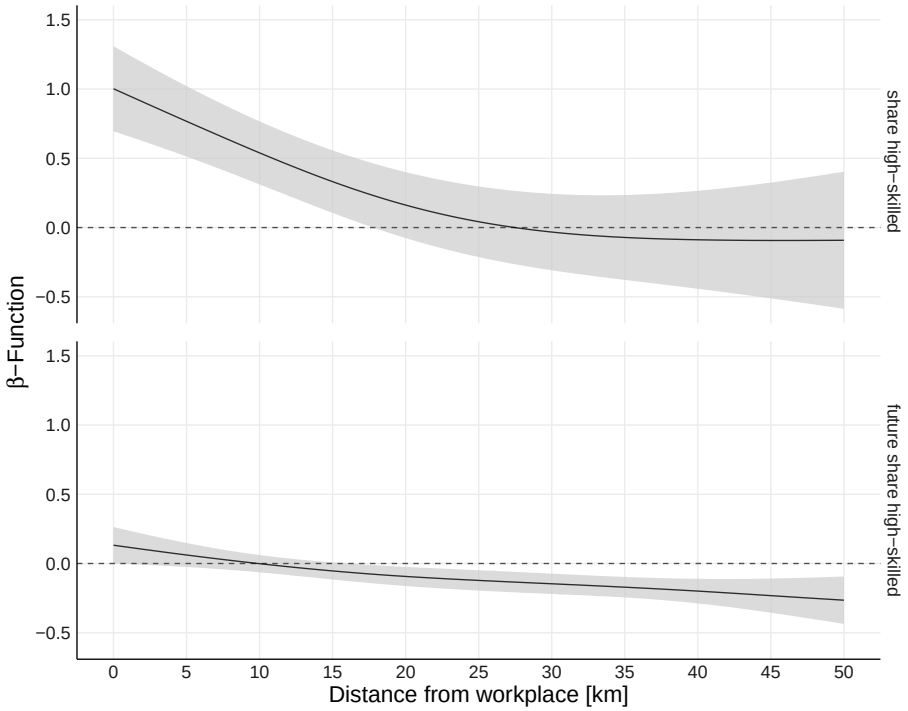
To estimate equation (3.11), we use the same dataset as in the paper and cluster standard errors at the county-level. Table 3.A.1, column 2 summarizes the results. Our model suggests significant positive spillovers from high-skilled workers into individual wages. The coefficient of 0.323 indicates that a one-standard-deviation increase in the regional share of high-skilled workers (7.2 percentage points) raises the wages of incumbent workers by 2.3%. The magnitude of this effect is close to our main findings, which imply that an evenly distributed one-standard-deviation increase in the share of high-skilled workers increases wages by 2%. Moreover, and similar to our main findings, neglecting skill-specific labor-market-area-year fixed effects significantly increases the computed coefficient (column 1). In summary, the predicted magnitude of spillover effects from an overall increase in the share of high-skilled workers is almost identical in county-level estimates and estimates based on the exact spatial distribution of workers.

3.A.7 Robustness

Placebo test: future concentration of high-skilled workers

Following Cornelissen et al. (2017), who identify human capital externalities in the workplace, we corroborate our findings with a placebo test, in which we expand our model with a one-year lead of the spatial distribution of high-skilled workers. Because workers cannot receive spillovers from neighbors who have not yet moved in, the future concentration of high-skilled workers serves as a placebo. As figure 3.A.1 indicates, the future concentration of high-skilled workers is almost unrelated to wages (bottom curve). Only after 17 kilometers' distance from the workplace does the model detect a small and economically negligible negative relationship between wages and the future concentration of high-skilled workers. Moreover, estimates of the human capital externalities from the current share of high-skilled workers change only slightly relative to the baseline specification (top curve). Overall, the placebo test buttresses our main findings.

Figure 3.A.1: Estimates of human capital externalities from the current and the future distribution of high-skilled workers



Notes: The figure depicts estimates of the human capital externalities from the current and future distributions of high-skilled workers on individual log wages. We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z and define the future concentration of high-skilled workers as the one-year lead of the share of high-skilled workers within distance z. We estimate Equation (9) in our paper, expanded with the lead of $X_{it}(z)$, with the smoothing spline estimator. The top panel presents estimates of the contemporaneous spillover function. The bottom panel depicts estimates of the link between log wages and the future concentration of high-skilled workers, which serves as the placebo. Black lines illustrate computed β functions, and gray areas indicate 99% confidence bands. The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 2,959,357$

Non-border regions

Because we have no data on workers outside of Germany, measurements of the distribution of high-skilled workers in border regions are partly truncated. For instance, establishments in the city center of Passau are only two kilometers from the Austrian border. Therefore, past two kilometers' distance, we observe the

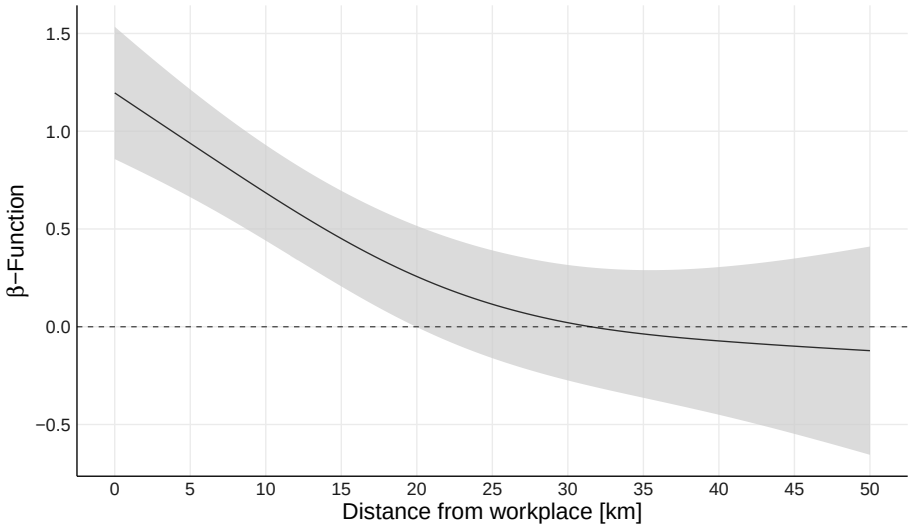
concentration of high-skilled workers only in southwest to northeast directions. Consequently, information on the distribution of high-skilled workers comes solely from these data points. Ignoring the partial truncation, we implicitly assume that the distribution on the Austrian side of the border is the same as on the German side of the border and that there are no costs from frictions in information flows across the border. To assess whether these assumptions influence our estimates, we now remove border regions from our dataset and re-estimate our main model with establishments that are at least 50 kilometers from the German border.

Figure 3.A.2 summarizes the results. Generally, the estimated curve resembles the spillover function from the full sample. Identically to our main findings, the function value is slightly above unity in the direct neighborhood of establishments. However, the graph implies that spillovers in non-border regions are slightly higher, and the point estimates reach seven kilometers farther than in the full sample. There are several explanations for the stronger effects in non-border regions. First, due to labor market barriers, spillovers in border regions might generally be lower, which would reduce measurements of the overall effect. Second, the concentration of high-skilled labor behind the German border might be lower than on the German side of the border, which would oppose our assumption of similar skill distributions on both sides of the border. Third, there are institutional differences between border and non-border regions that depress human capital externalities in border regions. Fourth, by chance, cities in border regions benefit less from human capital externalities than other cities do. Given the multitude of possible explanations, it seems plausible that estimates in non-border regions differ slightly from those in the full sample. Reassuringly, the point estimates of the spillover function are nonetheless similar in both samples, and the confidence bands overlap over the whole domain. Overall, the robustness exercise therefore confirms our main findings.

Labor market trends and industry clusters

Another concern may be that industry- or occupation-specific trends in the labor market influence our results through local clusters. To illustrate this issue, consider the following scenario. Industry b experiences an economic upswing that raises wages and the demand for skilled labor. If firms in industry b tend to cluster geographically, wages and the concentration of high-skilled labor would simultaneously rise in these areas. In our estimates, a global labor market shock at the industry level would therefore create a spurious relationship between wages and the regional concentration of high-skilled workers. The same applies to labor market shocks to occupations.

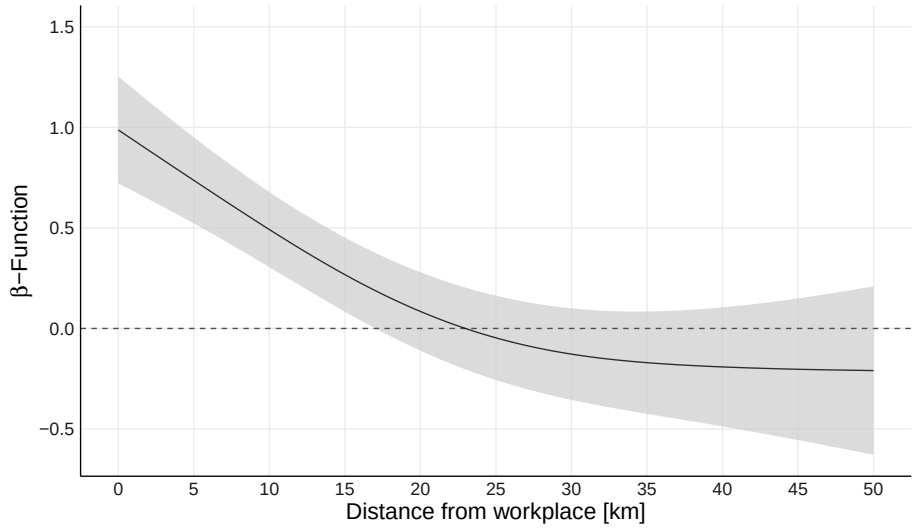
Figure 3.A.2: Spatial human capital externalities from high-skilled workers (without border regions)



Notes: The figure shows spatial human capital externalities from high-skilled workers into individual log wages in regions that are at least 50 kilometers from the German border. We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. To compute the spatial spillover function ($\beta(z)$), we estimate the model corresponding to Equation (9) in the paper using the smoothing spline estimator. We restrict the capacity of the β curve to a parabola-like function that may remain flat over some interval, and we set the penalty parameter ρ accordingly. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The effect of a p-percentage-point increase in the share of high-skilled workers within distance z_0 and z_1 (in a 0 to 1 range) is p times the area bellow the estimated spillover function from z_0 to z_1. The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 2,489,083$

To assess whether industry or occupation trends in the global labor market affect our results, we augment our estimation equation (Equation (9) of the manuscript) with year-specific industry and occupation fixed effects. These fixed effects absorb changes in wages and the concentration of high-skilled workers that stem from industry- or occupation-wide shifts in the labor market. Figure 3.A.3 shows the resulting spillover function. The curve is almost identical to that from our main specification (Figure 6 in our paper). We therefore conclude that trends at the industry or occupational level do not influence our results.

Figure 3.A.3: Spatial human capital externalities from high-skilled workers (removing industry and occupation trends)



Notes: The figure shows spatial human capital externalities from high-skilled workers into individual log wages. We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. To compute the spatial spillover function ($\beta(z)$), we estimate the model corresponding to Equation (9) in the paper using the smoothing spline estimator. To control for industry- and occupation-specific trends in the labor market, we additionally control for time-varying industry and occupation fixed effects. We restrict the capacity of the β curve to a parabola-like function that may remain flat over some interval, and we set the penalty parameter ρ accordingly. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The graph shows significant spillover effects that decay with distance. The effect of a p-percentage-point increase in the share of high-skilled workers within distance z_0 and z_1 (in a 0 to 1 range) is p times the area bellow the estimated spillover function from z_0 to z_1. The underlying model further controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 3,498,536$

Effects in urban and rural areas

Plausibly, marginal travel costs for physical distance differ in cities and rural areas. Additionally, social interactions in sparsely populated regions might be more costly than those in dense urban areas. Thus, the intensity and spatial reach of human capital externalities in cities and rural areas might differ. To assess these considerations, we separately estimate human capital externalities in urban and rural areas.

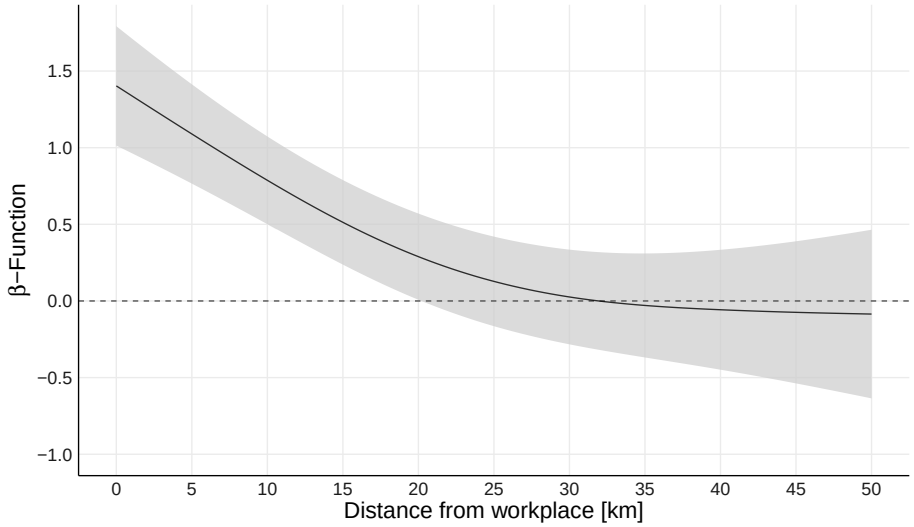
Figure 3.A.4 and figure 3.A.5 illustrate the estimates of human capital exter-

nalities within urban and rural areas. Estimates of human capital externalities in urban areas are generally similar to our main findings. However, compared to the overall population human capital externalities in urban areas are stronger and reach slightly further than in the average population. For instance, an increase of the share of high-skilled workers within five kilometers distance increases wages of workers in cities by 2.5%. The same increase in the share of high-skilled workers raises wages of the average worker by only 1.75%.

Contrarily, as figure 3.A.5 indicates, estimates of human capital externalities in rural areas are insignificant. These results suggest that workers in rural areas do not benefit from human capital externalities. Our identification strategy relies on a extensive set of fixed effects that remove all variation in the data that comes from the labor market area and time-invariant individual and establishment characteristics. Thus, we only measure human capital externalities form changes in the concentration of high-skilled workers in closer areas. Common variation in the intensity of human capital on the labor market area level, as well as time-invariant regional differences are not captured in our estimates. Apparently, our identification strategy is very demanding. Since the number of observations in rural areas is considerably smaller than in urban areas we cannot rule out that insignificant results in the rural sample might be due to efficiency issues. Figure 3.A.6 shows estimates where we replace worker-firm match fixed effects by worker fixed effects. Consequently, we do not control for time-invariant neighborhood characteristics in this estimation. Estimates in figure 3.A.6 are therefore less demanding because they not only use time-variant variation in the data but also variation between workplaces. Allowing, between-variation leads to significant estimates of human capital externalities. However, estimate are still considerably smaller than in the urban sample (even with less demanding controls). Moreover, since we no longer control for worker-firm match fixed effects, estimates might be confounded by other neighborhood characteristics.

Overall, our findings imply that human capital externalities are considerably stronger in urban areas than in rural areas. In fact, we find only weak evidence for human capital externalities in rural areas. Although these findings support our main results, they also suggest that they are mostly driven by urban areas.

Figure 3.A.4: Spatial human capital externalities from high-skilled workers (urban areas)



Notes: The figure shows the spatial human capital externalities from high-skilled workers into individual log wages in rural areas. We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. To compute the spatial spillover function ($\beta(z)$), we estimate the model corresponding to Equation (9) in the paper using the smoothing spline estimator. We restrict the capacity of the β curve to a parabola-like function that may remain flat over some interval, and we set the penalty parameter ρ accordingly. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 2.601.624$

Table 3.A.1: Spatial human capital externalities from high-skilled workers (full table)

Distance	Value	Sig.	SE	Distance	Value	Sig.	SE	Distance	Value	Sig.	SE
0.5	1.0654	***	0.1178	20.5	0.0890		0.0876	40.5	-0.2024		0.1332
1.0	1.0380	***	0.1151	21.0	0.0723		0.0882	41.0	-0.2037		0.1355
1.5	1.0106	***	0.1125	21.5	0.0562		0.0888	41.5	-0.2049		0.1379
2.0	0.9831	***	0.1100	22.0	0.0407		0.0894	42.0	-0.2060		0.1404
2.5	0.9558	***	0.1076	22.5	0.0258		0.0900	42.5	-0.2070		0.1430
3.0	0.9284	***	0.1052	23.0	0.0115		0.0907	43.0	-0.2080		0.1456
3.5	0.9011	***	0.1029	23.5	-0.0023		0.0913	43.5	-0.2088		0.1482
4.0	0.8739	***	0.1008	24.0	-0.0155		0.0920	44.0	-0.2096		0.1509
4.5	0.8467	***	0.0987	24.5	-0.0281		0.0926	44.5	-0.2103		0.1537
5.0	0.8196	***	0.0968	25.0	-0.0401		0.0933	45.0	-0.2109		0.1566
5.5	0.7926	***	0.0949	25.5	-0.0516		0.0940	45.5	-0.2115		0.1594
6.0	0.7656	***	0.0932	26.0	-0.0625		0.0946	46.0	-0.2120		0.1624
6.5	0.7387	***	0.0916	26.5	-0.0729		0.0953	46.5	-0.2124		0.1653
7.0	0.7119	***	0.0901	27.0	-0.0828		0.0961	47.0	-0.2128		0.1683
7.5	0.6852	***	0.0887	27.5	-0.0921		0.0968	47.5	-0.2131		0.1714
8.0	0.6585	***	0.0875	28.0	-0.1009		0.0976	48.0	-0.2134		0.1745
8.5	0.6320	***	0.0864	28.5	-0.1093		0.0983	48.5	-0.2137		0.1776
9.0	0.6057	***	0.0854	29.0	-0.1171		0.0991	49.0	-0.2139		0.1808
9.5	0.5795	***	0.0846	29.5	-0.1245		0.1000	49.5	-0.2142		0.1839
10.0	0.5535	***	0.0838	30.0	-0.1314		0.1008	50.0	-0.2144		0.1872
10.5	0.5277	***	0.0832	30.5	-0.1379		0.1018	**Controls**			
11.0	0.5021	***	0.0827	31.0	-0.1440		0.1027	Age	-0.6766		1178.6
11.5	0.4768	***	0.0824	31.5	-0.1496		0.1037	Age^2	-0.0003	***	0.0000
12.0	0.4518	***	0.0821	32.0	-0.1548		0.1048	Exper.	0.0814	***	0.0016
12.5	0.4270	***	0.0819	32.5	-0.1597		0.1059	$Exper.^2$	-0.0001	***	0.0000
13.0	0.4026	***	0.0818	33.0	-0.1642		0.1071	Tenure	0.0042	***	0.0009
13.5	0.3785	***	0.0818	33.5	-0.1684		0.1083	$Tenure^2$	-0.0001	***	0.0000
14.0	0.3548	***	0.0819	34.0	-0.1723		0.1096	l. firm size	0.0258	***	0.0009
14.5	0.3315	***	0.0821	34.5	-0.1758		0.1109	l. p. dens.	0.0011		0.0006
15.0	0.3086	***	0.0823	35.0	-0.1792		0.1124	l. hotel b.	0.0059		0.0034
15.5	0.2861	***	0.0826	35.5	-0.1822		0.1139	Unemp.	0.0009		0.0006
16.0	0.2641	***	0.0829	36.0	-0.1851		0.1155				
16.5	0.2425	***	0.0833	36.5	-0.1877		0.1171				
17.0	0.2214	**	0.0838	37.0	-0.1901		0.1189				
17.5	0.2009	**	0.0842	37.5	-0.1923		0.1207				
18.0	0.1809	*	0.0847	38.0	-0.1943		0.1226				
18.5	0.1614		0.0853	38.5	-0.1962		0.1245				
19.0	0.1424		0.0858	39.0	-0.1980		0.1266				
19.5	0.1241		0.0864	39.5	-0.1996		0.1287				
20.0	0.1063		0.0870	40.0	-0.2011		0.1309				

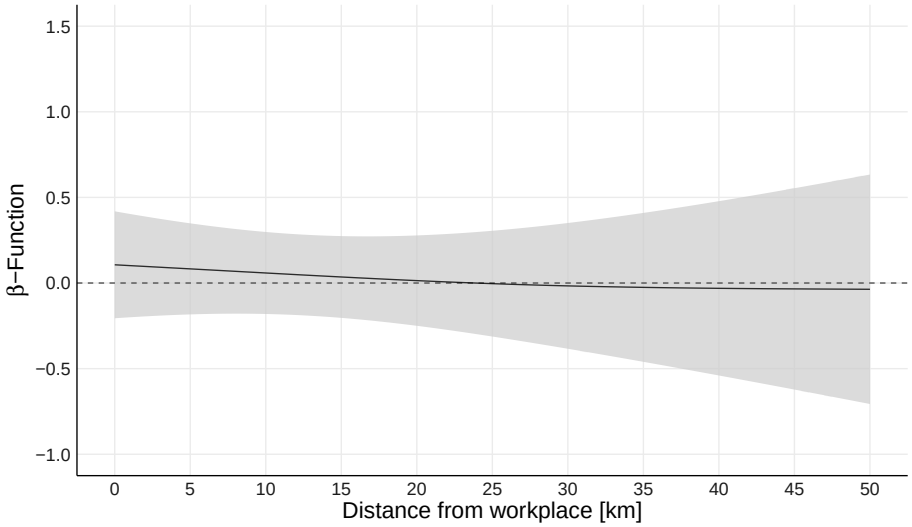
Notes: The table accompanies Figure 6 of our paper and shows the strength of spatial human capital externalities from high-skilled workers at numerous distances on individual log wages. To compute the spatial spillover function ($\beta(z)$), we estimate Equation 9 of the paper using the smoothing spline estimator. We restrict the capacity of the β curve to a parabola-like function that may remain flat over some interval, and we set the penalty parameter ρ accordingly. The table also reports coefficient estimates for the control variables. The underlying model further controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation fixed effects and time fixed effects. Standard errors are clustered. ***, ** and * indicate significance at the 1%-, 5%- and 10%-level, respectively. $N = 3,498,536$

Table 3.A.1: human capital externalities at the county-level

	(1)	**(2)**
Share of high-shilled workers	0.409^{***}	0.323^{***}
	(0.095)	(0.045)
Worker-firm match fixed effects	Yes	Yes
Labor-market-area × year × skill fixed effects	No	Yes

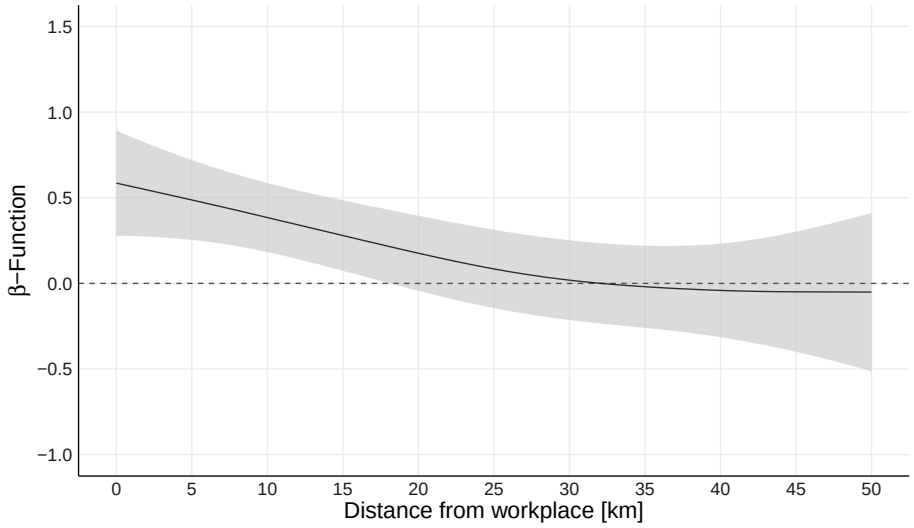
Notes: The table summarizes estimates of the human capital externalities from high-skilled workers into individual log wages at the county level. The estimates replicate our main model at an aggregate level and serve as a comparison of the magnitude of the effects. The underlying models further control for occupation fixed effects, time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). Cluster-robust standard errors are in parentheses.*** indicates significance at the 0.1%-level. $N = 3,498,536$

Figure 3.A.5: Spatial human capital externalities from high-skilled workers (rural areas)



Notes: The figure shows the spatial human capital externalities from high-skilled workers into individual log wages in rural areas. We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. To compute the spatial spillover function ($\beta(z)$), we estimate the model corresponding to Equation (9) in the paper using the smoothing spline estimator. We restrict the capacity of the β curve to a parabola-like function that may remain flat over some interval, and we set the penalty parameter ρ accordingly. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The underlying model controls for worker-firm match fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 896.912$

Figure 3.A.6: Estimates of the spatial human capital externalities from high-skilled workers (rural areas, no worker-firm match fixed effects)



Notes: The figure shows estimates of the spatial human capital externalities from high-skilled workers into individual log wages in rural areas without nullifying worker-firm match fixed effects (but worker fixed effects only). We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. To compute the spatial spillover function ($\beta(z)$), we estimate the model corresponding to Equation (9) in the paper using the smoothing spline estimator. We restrict the capacity of the β curve to a parabola-like function that may remain flat over some interval, and we set the penalty parameter ρ accordingly. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99% confidence band. The underlying model controls for worker fixed effects, skill-specific yearly labor-market-area fixed effects, occupation and time fixed effects and worker characteristics (age, work experience, tenure and the respective second-order polynomials), log establishment size and county characteristics (unemployment rate, log population density and the log number of hotel beds as a proxy for amenities). $N = 896.912$

Chapter 4

Directed local testing in the functional linear model

4.1 Introduction

With the increasing availability of data collected on a dense grid, *functional data analysis* (FDA) has become an important field in statistics in recent years. The corresponding literature comprises numerous theoretical contributions and various interesting applications. In particular, the functional linear regression model with a scalar dependent variable is quite popular and has seen many applications in different areas (see Ullah and Finch, 2013, for an overview of various fields of application).

It is well known, however, that regression models with a functional predictor belong to the class of ill-posed inversion problems and that every estimate has to use some type of regularisation. As a result, local inference in the functional linear model is a difficult and challenging issue. More specifically, Cardot et al. (2007) show that it is impossible to derive a CLT for the estimated coefficient function. As a consequence, it is also not possible to draw uniform inference about the functional coefficient, even asymptotically.

This does not mean that statistical tests for the parameter function are not possible in general. Indeed, there are several approaches to test whether the coefficient function deviates globally from a given null hypothesis. For instance, Cardot et al. (2003, 2004) use properties of the cross-covariance operator, Swihart et al. (2014); Kong et al. (2016) adapt classical methods such as likelihood-ratio and F tests to the functional case, and González-Manteiga et al. (2012) build on bootstrap techniques to construct such a global test. With these tests, however, it is only possible to find out whether the relationship between the functional

predictors and the scalar outcome variable deviates from a prespecified null. The most important specific case, for instance, is the test of association in the functional linear model, where the null hypothesis corresponds to the zero function as functional regression coefficient.

Though, if the global test leads to a rejection, it is still not clear, whether the functional coefficient deviates from the null on the whole domain or is only different on a subpart of the domain. Therefore, a rejection of the global test does not uncover parts of the domain where the functional coefficient differs almost everywhere from a null coefficient. Evidently, this is an important shortcoming of the scalar-on-function regression model for practitioners that are interested in identifying subdomains where a functional covariate is related significantly to the scalar response.

An important contribution to this direction is the paper by Hall and Hooker (2016). Using their method, it is possible to consistently estimate the support of the coefficient function (that is, a subinterval $[0, \theta]$ of the original domain where $\beta(t) \neq 0$). However, this method can not tell whether the coefficient function over the estimated support is statistically different from zero at a given significance level. Very briefly, Hall and Hooker (2016) also discuss the possibility of constructing a confidence interval for the boundaries of the support, based on the bootstrap. However, in their discussion, they also raise the issue that using the bootstrap is not appropriate in the context of smoothing. Again, this leaves the practitioner only with a point estimate.

We present an approach that shows one possibility to overcome these limitations under some additional assumptions about the structure of the coefficient function. Our testing procedure then is able to uncover a subinterval of the functional domain where the coefficient function is statistically different from zero almost everywhere. The underlying idea is to apply the global test sequentially on a family of subsets of the domain such that the family-wise error rate is controlled. The performance of available approaches for the global test is compared in Tekbudak et al. (2019), showing favorable properties of the F test suggested by Kong et al. (2016). Therefore, our method builds on this F test, although it might be replaced by other versions of the global test (see Tekbudak et al., 2019, for several approaches).

Inference in the functional regression model with scalar response has also been addressed by several other authors. In a generalised model framework, Müller and Stadtmüller (2005) propose (simultaneous) confidence bands for the coefficient function that can be used to compare an estimate with a prespecified slope function on the whole domain. Their confidence bands are based on the test statistic for the global test and ensure only coverage for a finite-dimensional sur-

rogate function (see also Section 2.3 of Imaizumi and Kato, 2019) and, therefore, cannot be used to find out at which regions the estimated functional coefficient is statistically different from zero. By weakening the requirements for the coverage probability, Imaizumi and Kato (2019) also propose confidence bands for the estimator based on functional principal component analysis, which due to the weaker requirements does not reflect a suitable testing framework. In order to illustrate the performance of different estimation methods, Reiss and Ogden (2007) use empirical (pointwise) confidence bands based on simulations, which, however, are not meant and can not be used for inference. Constructing confidence regions for random elements in infinite dimensional Hilbert spaces is in general a challenging problem and it is not naturally clear how such confidence bands have to be constructed (see, e.g. Choi and Reimherr, 2018, for one possible approach).

The literature related to the scalar-on-function regression model is extensive and we refer to Reiss et al. (2017) for an overview of different methods. For readers with a general interest in FDA, the books by Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), and Hsing and Eubank (2015) provide a broad overview of available methods.

The remainder of this article is organized as follows. In Section 4.2, we introduce the model framework. The main results are given in Section 4.3, where we extend the global F test to a spline basis and introduce the sequential testing procedure. In that section, we also present our theoretical results. In Section 4.4 we illustrate the properties of the suggested method by means of a simulation study, and present a real data application in Section 4.5. Section 4.6 concludes.

4.2 Model framework

We consider the functional linear model (FLM) with a scalar response variable

$$Y_i = \beta_0 + \int_{\mathcal{D}} \beta(t) X_i(t)\, \mathrm{d}t + \varepsilon_i, \tag{4.1}$$

reflecting the dependency between observations of a scalar variable, $Y_1, \ldots, Y_n$, and a functional covariate $X_1, \ldots, X_n$, the latter taking values in the Hilbert space of square intgrable functions $L^2(\mathcal{D})$. Without loss of generality, we set $\mathcal{D} = [0, 1]$ and assume for notational simplicity that $E(Y_i) = 0$ and $E(X_i) = 0 \in L^2$, omitting the intercept, β_0 hereafter. For the error term we assume that the ε_i's are i.i.d. centered ($E(\varepsilon_i) = 0$) random variables with variance $Var(\varepsilon_i) = \sigma^2 < \infty$, and are independent of the X_i. The function-valued slope parameter $\beta \in L^2(\mathcal{D})$ quantifies the effect of the functional predictors on the scalar outcome variable,

and, very often is of central interest. Due to the reasons outlined above, available methods only leave a practitioner with a point estimate $\widehat{\beta}$ and the information whether the relationship between X_i and Y_i is significant at a given confidence level.

In some applications, it can be reasonable to assume that if there is a linear dependence between the scalar response and the functional covariate, it must be strongest at one boundary of the domain. For instance, if one can assume that $|\beta|$ is a monotonically decreasing function, it is then possible to construct a sequential test procedure that is capable to identify a subset of the domain, $[0, \tau^*] \in [0, 1]$ where at a prespecified level α, it holds that the function $\beta(t) \neq 0$ for almost every $t \in [0, \tau^*]$.

Before we introduce the test approach, let us define the basic notation. To this end, it is assumed that all realizations of the functional covariate are observed on the same grid values $t_1, \ldots, t_p$, with $t_1 = \frac{1}{2p}$, $t_i - t_{i-1} = \frac{1}{p}$. In the $n \times p$ matrix $\mathbf{X}$, we collect the functional observations and the vector $\mathbf{Y} \in \mathbb{R}^n$ holds the observations of the dependent variable, where n denotes the sample size.

4.3 Testing procedure and theoretical results

To formulate our test procedure, we make the following assumption about the coefficient function β.

> The functional coefficient β is continuous on $[0, 1]$, the absolute value,$|\beta|$, is monotonically decreasing, and there exists a $\tau \in [0, 1]$ such that $\beta(t) \neq 0$ for all $t \in [0, \tau]$. (ASS 1)

Under this assumption, the procedure proposed in this paper is able to find the largest number $\tau^* \in [0, \tau]$ such that β is statistically different from zero on the interval $[0, \tau^*]$ at a predefined significance level α. The core idea of the test procedure is to split the domain into two parts, $[0, t_l]$ and $[t_l, 1]$, and test whether $\beta(t) = 0$ for almost all $t \in [t_l, 1]$ sequentially for all split points t_l with $l = 1, \ldots, p - 1$.

Assumption ASS 1 may appear quite restrictive, but in our view it reflects the most important case in practice. Monotonicity is not a necessary condition for the theoretical results, however, it is quite helpful for the structural interpretation of β. Qualitatively the same results are also possible without assuming monotonicity, for instance, assumption 6.1 of Hall and Hooker (2016) would be also sufficient to get the same results. By building on a basis system for estimating β, even continuity may not be necessary and one might allow for a finite number of discontinuities.

A consistent estimator for β also requires some additional regularity assumptions regarding the process X and the functional coefficient β. These requirements depend on the estimation context, see Hall and Horowitz (2007) or Hall and Hosseini-Nasab (2006) for the classical approach based on functional principal components and Crambes et al. (2009); Cardot et al. (2007) for an estimator based on smoothing splines. We do not discuss the consequences arising from the discretization of the curves. Hence, we assume that the number of observations points p is sufficiently large compared to the number of observations n, such that the approximation error can be neglected. Formal arguments can also be found, for example, in Crambes et al. (2009). Furthermore, we assume that the function β can be fully represented using B-splines. A more rigorous development of the corresponding theory is therefore left for future work.

4.3.1 Local test

For each $l = 1, \dots, p-1$ the null and alternative hypothesis of the local test are

$$\begin{aligned} H_0 &: \beta(t) = 0 \text{ for almost all } t \in [t_l, 1], \\ H_1 &: \beta(t) \neq 0 \text{ for some } t \in [t_l, 1], \end{aligned} \tag{4.2}$$

where, for the alternative, the set $\{t \in [t_l, 1] \,|\, \beta(t) \neq 0\}$ must not be a zero set with respect to the Lebesgue measure.

The test can also be formulated differently by partitioning the integral in Model (4.1) into two parts. To this end, let $\beta_{1,t_l}(t) = \mathbb{1}_{[0,t_l]}(t)\beta(t)$ and $\beta_{2,t_l}(t) = \mathbb{1}_{[t_l,1]}(t)\beta(t)$ denote the first and second part of $\beta(t)$ splitted at t_l. Model (4.1), using that notation, is then equivalent to the splitted model

$$Y_i = \int_{\mathcal{D}} \beta_{1,t_l}(t) X_i(t) \,\mathrm{d}t + \int_{\mathcal{D}} \beta_{2,t_l}(t) X_i(t) \,\mathrm{d}t + \varepsilon_i, \tag{4.3}$$

and the test (4.2) is then equivalent to test (globally) for $\beta_{2,t_l} = 0$ a.e. for which several methodologies have been suggested in the literature. Here, we build on the F test suggested by Kong et al. (2016) which builds on the classical Karhunen-Loève decomposition. However, using their approach straight away, it would be necessary to solve a relatively costly eigenvalue problem twice for every local test. To avoid this, we instead use a fixed spline basis which we have to assume to be appropriate for β. In the following section, we briefly explain how model (4.3) is fitted based on a B-spline expansion of the splitted β.

Spline estimator for splitted model

Let the integer k denote the dimension of the spline basis for the splitted model which should be chosen such that the corresponding basis is flexible enough to expand β. A function space corresponding to a cubic spline basis has at least four dimensions, hence, the function space for the splitted β has at least eight dimensions ($k \geq 8$). However, when $t_l < 4$ (or $t_l > p-3$), the matrix of a cubic spline basis evaluated at $0, \ldots, t_l$ (or $t_l, \ldots, 1$, respectively) are not regular and, therefore, these cases have to be treated separately.

Starting with the case $l \in \{4, \ldots, p-3\}$, the corresponding knot sequence $\tau_1, \ldots, \tau_{k+4}$, spanning $[0,1]$, for the spline basis is given by

boundary knots:

$$\tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$$

$$\tau_{k+1} = \tau_{k+2} = \tau_{k+3} = \tau_{k+4} = 1$$

interior knots:

$$\tau_j = \begin{cases} t_l & \text{if } t_l \in (\tau_{j-1} + 0.5\,\delta, \tau_{j-1} + 1.5\,\delta] \wedge \tau_{j-3} \neq t_l \\ \tau_{j-5} + 2\delta & \text{if } \tau_{j-1} = t_l \\ \tau_{j-1} + \delta & \text{otherwise,} \end{cases}$$

where $\delta = \frac{1}{k-6}$ is the distance between interior knots. Intuitively, the above knot sequence has four boundary knots at 0 and 1, respectively, $k-7$ equidistant interior knots, where the interior knot which is closest to t_l is replaced by four knots at t_l. The corresponding B-spline basis, by design, spans a function space of twice continuously differentiable functions on the intervals $(0, t_l)$ and $(t_l, 1)$ and a jump point (due to repeating the knot t_l four times) at t_l. We denote by $\mathbf{b}(t) = (b_1(t), \ldots, b_k(t))^\top$ the corresponding basis functions and by $BS(k, t_l) = \text{span}(b_1, \ldots, b_k)$ the k dimensional function space based on B splines with split point at t_l. Every function $f \in BS(k, t_l)$ can be represented by a k-dimensional parameter θ, such that under the assumption $\beta \in BS(k, t_l)$, model (4.1) can be rewritten as

$$Y_i = \int X_i(t)\mathbf{b}(t)^\top \theta \, \mathrm{d}t + \varepsilon_i.$$

If we approximate the integral by the corresponding Riemann sum (using the midpoint rule), we get

$$Y_i = \frac{1}{p} \sum_{j=1}^{p} X_i(t_j)\mathbf{b}(t_j)^\top \theta + \varepsilon_i.$$

Let $\mathbf{B}$ be the $p \times k$ matrix of the k spline functions evaluated at the p grid values, that is $\mathbf{B}_{ij} = b_j(t_i)$. The matrix $\mathbf{B}$ has a block structure, where the upper-left $l \times m$ block $\mathbf{B_1}$ (m is the number of knots preceding the 4 knots for the split point) is the evaluated basis for the interval $[0, t_l]$ and the lower-right $(p-l) \times (k-m)$ block $\mathbf{B_2}$ is the basis for $(t_l, 1]$. The remaining upper-right and lower-left blocks are zero matrices.

In the other case, where $l < 4$ or $l > p-3$, the cubic spline basis evaluated at $0, \ldots, t_l$ (or $t_l, \ldots, 1$) is no longer invertible and, therefore, this singular block is replaced by the identity matrix of size l (or $p-l$, respectively).

The least-squares estimate of the spline coefficients $\widehat{\theta}$ is obtained as usual as

$$\widehat{\theta} = p\left(\mathbf{B}^\top \mathbf{X}^\top \mathbf{X}\mathbf{B}\right)^{-1} \mathbf{B}^\top \mathbf{X}^\top \mathbf{y}.$$

Under the null, the last $k-m$ entries of θ are zero and the m remaining non-zero entries of $\widehat{\theta}$ in the restricted model become

$$\widehat{\theta}_1 = p\left(\mathbf{B_1}^\top \mathbf{X}^\top \mathbf{X}\mathbf{B_1}\right)^{-1} \mathbf{B_1}^\top \mathbf{X}^\top \mathbf{y}.$$

Analogous to the finite dimensional regression model, this results in the projection matrices

$$\mathbf{P_{XB}} = \mathbf{XB}\left(\mathbf{B}^\top \mathbf{X}^\top \mathbf{XB}\right)^{-1} \mathbf{B}^\top \mathbf{X}^\top$$
$$\mathbf{P_{XB_1}} = \mathbf{XB_1}\left(\mathbf{B_1}^\top \mathbf{X}^\top \mathbf{XB_1}\right)^{-1} \mathbf{B_1}^\top \mathbf{X}^\top$$

for the unrestricted and the restricted model, respectively.

Test statistic

The F test in Kong et al. (2016) is defined in terms of the residual sum of squares under the full and null models:

$$RSS_{\text{full}} = \mathbf{y}^\top (\mathbf{I}_n - \mathbf{P_{XB}})\mathbf{y}$$
$$RSS_{\text{null}} = \mathbf{y}^\top (\mathbf{I}_n - \mathbf{P_{XB_1}})\mathbf{y}$$

and the test statistic is

$$T_F = \frac{(RSS_{\text{null}} - RSS_{\text{full}})/(k-m)}{RSS_{\text{full}}/(n-k)} = \frac{\mathbf{y}^\top (\mathbf{P_{XB}} - \mathbf{P_{XB_1}})\mathbf{y}/(k-m)}{\mathbf{y}^\top (\mathbf{I}_n - \mathbf{P_{XB}})\mathbf{y}/(n-k)}. \tag{4.4}$$

Assuming that the discretization error is negligible, we can formulate the following result.

Algorithm 2 Directed sequential test

1: Let α denote global significance level and H_0^i denote the i-th local null hypothesis
2: Initialize $i = 1$
3: **while** $i < p \ \wedge H_0^{i-1}$ was rejected **do**
4: test H_0^i at level α
5: set $i = i + 1$
6: **end while**
7: With H_0^{i-1} being the last rejected hypothesis, conclude that β is statistically different from zero at level α over the interval $[0, \tau^*]$, with $\tau^* = t_{i-1}$.

Lemma 1. *If for fixed k and l, $\beta \in BS(k, t_l)$ (the B-spline space with k dimensions and split point at t_l) and the errors are centered i.i.d. normal, then, under the null, T_F is F-distributed with $k - m$ and $n - k$ degrees of freedom.*

Remark. *By requiring that the true parameter function is an element of $BS(k, t_l)$, the above result is limited as is does not consider the more general case that $\beta \in L^2(\mathcal{D})$. For many applications, this is a reasonable assumption, particularly when the true parameter function is sufficiently smooth. The general case requires that the model complexity (k in our notation) grows at a suitable rate with the number of observations n. As mentioned above, developing formal arguments for the general case, is left for future research.*

4.3.2 Sequential directed test

Based on the above result, we can formulate the testing procedure for detecting the interval $[0, \tau^*]$, on which the curves significantly influence the dependent variable Y. The central idea is to perform the above local test consecutively for every $l \in \{1, \ldots, p-1\}$ and stop once the test does not reject for the first time. The iterative procedure is given in Algorithm 2. The following result shows that our directed testing procedure controls the family-wise error rate (FWER) in the strong sense by design.[1]

Theorem 1. *The sequential test described in Algorithm 2 controls FWER in the strong sense, that is, for any $\tau \in [0, 1]$ for which $\beta \mathbb{1}_{t>\tau} = 0$ a.e., the probability to reject too many hypotheses is bounded by α, namely, $P(\tau^* > \tau) \leq \alpha$.*

[1] A test procedure is said to control the FWER *in the strong sense*, if the probability of making at least one type I error is controlled at level α for every combination of true or non-true individual null hypotheses, whereas the *weak sense* control refers to the case where all individual null hypothesis have to be true.

As shown in the proof, our test is an example for a test meeting the requirements for the closed testing principle (Marcus et al., 1976). By design, it controls FWER without any further corrections.

Remark. *For the above result, it is not even necessary that β belongs to all $BS(k, t_l)$, $l = 1, \ldots, p-1$. It is sufficient, if the requirements of Lemma 1 are fulfilled for the first local test in the test sequence for which the null hypothesis is true. This local test then takes the role of a gatekeeper, such that the size of the overall procedure is maintained at level α.*

Under Assumption ASS 1, $\widehat{\beta}$ is statistically different from zero on the interval $[0, \tau^*]$ and one can conclude that functional predictors on this interval are significantly associated to the scalar outcome variable.

4.4 Simulation study

In the following simulation study, we investigate the performance of the sequential testing procedure. Apart from a numerical validation of the above theoretical results, it is of interest to analyse power properties of the test under different scenarios, in particular with respect to the distance to τ. The simulation exercise considers two different Data Generating Processes (DGPs) which can be seen as extreme scenarios. While the first DGP is a step function that is useful to demonstrate how closely the procedure can approach the point τ where the true β becomes exactly zero, the second DGP is a smooth and monotonically decreasing function which approaches zero at $t = 1$ (see Figure 4.4.1 for an illustration of both DGPs). The actual specification for the first DGP is $\beta_{\texttt{step}} = \mathbb{1}_{[0,0.5]}$, and for the second DGP, $\beta_{\texttt{smooth}}$ is a continuously decreasing function based on B-splines with 8 equidistant interior knots. While $\beta_{\texttt{smooth}}$ by construction fulfills the assumptions of Lemma 1 for all individual tests, this is not true for $\beta_{\texttt{step}}$ due to the discontinuity at $t = 0.5$. It can be easily seen, however, that $\beta_{\texttt{step}}$ fulfills the requirements of Lemma 1 for one individual test, namely with split point $t_l = 0.5$, which is the first individual test for which the null hypothesis is true. Thus, both DGPs meet the necessary requirements for the procedure.

For each DGP we consider three different sample sizes ($n = 250, 500, 1000$) and six different signal-to-noise ratios ($\gamma = 0.1, 0.5, 1, 2, 5, 10$). For every scenario, we generate 10 000 replications of n random tuples (X_i, Y_i) via Model (4.1), where the random curves X_i are i.i.d. realizations of the functional random variable $\sum_{1<j<p} Z_j \, j^{-2} \phi_j$, where $\phi_j(t) = \sqrt{2}\cos((j-1)\pi t)$ and independent standard normal Z_j. In the appendix, we also consider random curves with a different covariance function that is more concentrated around the diagonal. Although the

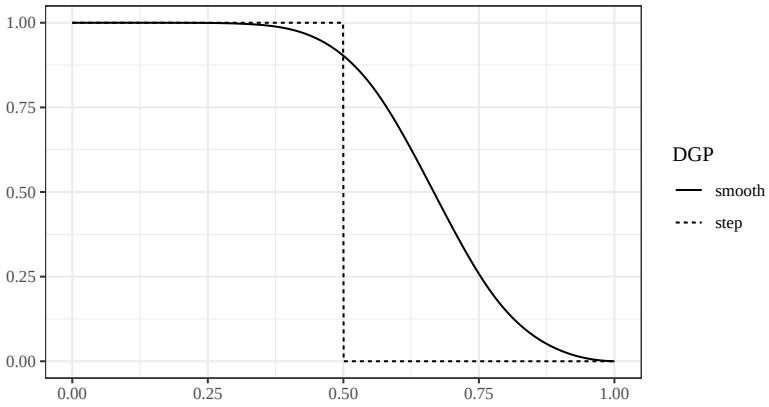


Figure 4.4.1: Coefficient functions used in the simulation exercise.

Table 4.4.1: Type-I error rates for DGP with $\beta = \beta_{\text{step}}$ and at global level $\alpha = 0.05$,

	$p = 100$				$p = 300$			
n	100	250	500	1000	100	250	500	1000
γ								
0.1	0.007	0.008	0.011	0.012	0.006	0.007	0.009	0.014
0.5	0.007	0.012	0.014	0.019	0.008	0.013	0.016	0.019
1	0.012	0.015	0.017	0.020	0.012	0.015	0.018	0.018
2	0.014	0.019	0.021	0.021	0.013	0.016	0.020	0.024
5	0.017	0.021	0.022	0.026	0.016	0.023	0.024	0.026
10	0.018	0.021	0.024	0.027	0.019	0.023	0.028	0.027

actual number of discretization points p is of minor importance, the simulation study also considers two cases ($p = 100, 300$). The error term ε_i is a centered normally distributed random variable with variance σ^2. The signal-to-noise ratio is specified as the quotient of the variance of $\int X_i(t)\beta(t)\,\mathrm{d}t$ and the error variance σ^2. For the test, one also has to specify the number of spline basis functions for the expansion of $\widehat{\beta}$ which we set to $k = 12$. The simulation is implemented in GNU R (R Core Team, 2017) and, together with an R-package implementing the testing procedure, the code is part of the online supplement to this article.

Simulation results for the type-I error rate are summarized in Table 4.4.1 for the DGP involving $\beta_{\texttt{step}}$. Note, that type-I-error rates for the DGP involving β_{smooth} cannot be reported since the tested H_0 is never true for this DGP. In case of the first DGP, a type-I error occurs if the test does not stop rejecting at or before $t = 0.5$. Every row of the table corresponds to a different signal-to-noise

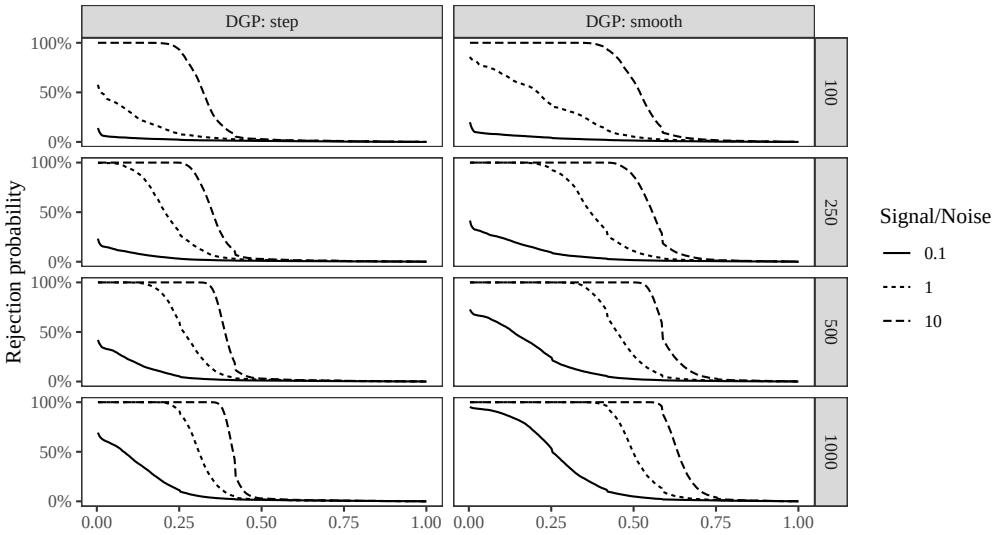


Figure 4.4.2: Rejection probabilities of the directed testing procedure for the two DGPs with $p = 300$.

ratio γ. Combinations of sample size n and number of discretization points p are found in the columns of the table, where columns 1-4 correspond to $p = 100$ and columns 5-8 correspond to $p = 300$. The global significance level here is set to $\alpha = 0.05$ and we can see that the procedure maintains size in all constellations, as expected.

Figure 4.4.2 shows the results of our power analysis. Left and right column correspond to $\beta_{\texttt{step}}$ and $\beta_{\texttt{smooth}}$, each row refers to a specific sample size, and different line types indicate different signal-to-noise ratios. The line of the graph depicts the probability of a rejection up to the respective value $t \in [0, 1]$ on the horizontal axis. It can be seen that sample size and signal-to-noise ratio have the expected influence on the rejection probability: the larger the sample size and the larger the signal-to-noise ratio, the larger is the power.

4.5 Application

With the following real data exercise, we want to demonstrate the applicability of the sequential testing method in practice. We are drawing on data from a sports biomechanics experiment. Using the functional data approach to model such measurements is quite natural and frequently adopted by empirical researchers in biomechanics (Vanrenterghem et al., 2012; Liebl et al., 2014; Hamacher et al., 2016; Warmenhoven et al., 2019).

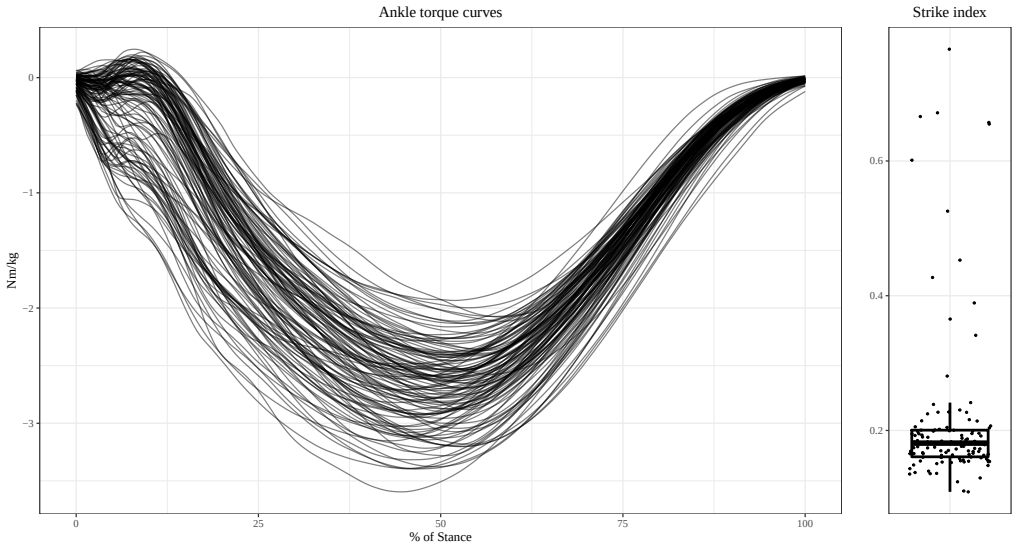


Figure 4.5.1: Sample of torque curves measured at the ankle joint and boxplot of the strike index of recreational runners.

The data shown in Figure 4.5.1 result from a sports biomechanics experiment conducted at the biomechanics lab of the German Sport University, Cologne, Germany. The sample comprises measurements of torque curves at the ankle joint as well as the strike index of $n = 119$ recreational runners, obtained while they were running without shoes. Each curve describes the torque measured at the right ankle joint in the sagittal pane during the *stance phase*, i.e. the phase during which the foot has contact to the ground. Using an affine transformation, the individual stance phases were standardized such that the initial ground contact takes place at $t = 0\%$ while the foot leaves the ground at $t = 100\%$. The torque measures are standardized by the participants' body weights to make the curves comparable. The right panel of Figure 4.5.1 shows the distribution of the so-called *strike index*, describing the center of pressure (CoP) at initial ground contact with respect to the long axis of the foot (Cavanagh and Lafortune, 1980; Altman and Davis, 2012). The strike index is used to classify runners into the groups rear foot, mid foot or fore food strikers, the majority of runners belonging to the group of rear foot strikers.

The torque curves as a whole show a similar pattern over all runners with a negative value throughout the largest part of the stance phase, representing a torque in the plentarflexion direction which is associated with the acceleration into the moving direction. Many curves, however, at the very beginning of the stance phase are positive which indicates a torque in dorsiflexion direction, associated

with a controlled lowering of the fore food immediately after ground contact.

It is therefore obvious that the torque curves at $t = 0\%$ depend on the center of pressure at initial ground contact. Runners with a very low strike index (center of pressure is very close to the heel) are likely to have a torque curve which is positive in the beginning of the stance phase, while runners with a higher strike index (center of pressure is located more in the center of the foot's long axis) need a negative torque at the ankle joint in order to accelerate the body mass upwards as well as into the moving direction. However, it is not clear which part of the stance phase has an association with the foot strike pattern and how large that part is.

With torque curves forming the functional predictor and the strike index as the scalar dependent variable, we applied the sequential test to find out for which part of the stance phase the dependence between torque curves and strike index is statistically different from zero. At first, however, we present the result of an application of a smoothing splines estimation (Crambes et al., 2009) of the corresponding functional linear model. Since the data is not standardized, we also include an intercept in the model. Figure 4.5.2 shows in the upper panel an estimated functional coefficient based on a smoothing parameter for which we set the effective degrees of freedom to 5. As expected, the estimate shows a negative relationship between torque curves and strike index and from 25% of the stance phase, the estimated functional coefficient approaches zero. As lined out in the introduction, such a point estimate cannot indicate anything about statistical significance of the result.

The result of applying our sequential testing procedure at global level $\alpha = 0.05$ is given in the lower panel of Figure 4.5.2. The grey area indicates the region where the test rejects the null of no association which in this application turns out to be the first 14% of the stance phase. This result makes sense if we once more take a look at the torque curves in Figure 4.5.1 where we can see that at the first 14% in particular, the shape of the curves vary the most. In the later stance phase, the curves have a rather similar pattern across individuals. Consequently, it is not very likely that the center of pressure during the first ground contact is related to the torque in later stance phase. The solid and dashed black lines of Figure 4.5.2 depict the estimation result of the splitted model which, however, the estimated slope function on the smaller domain, $[0\%, 14\%]$, should be interpreted with care since it is not quite clear whether the resulting oscillating pattern really reflects the relationship beween the foot strike pattern and the torque curves. On the other hand, if we can assume that the true β is monotonic which is supported by the estimation result on the whole domain, we can conclude from the outcome of the sequential testing procedure that the torque curves and the foot strike

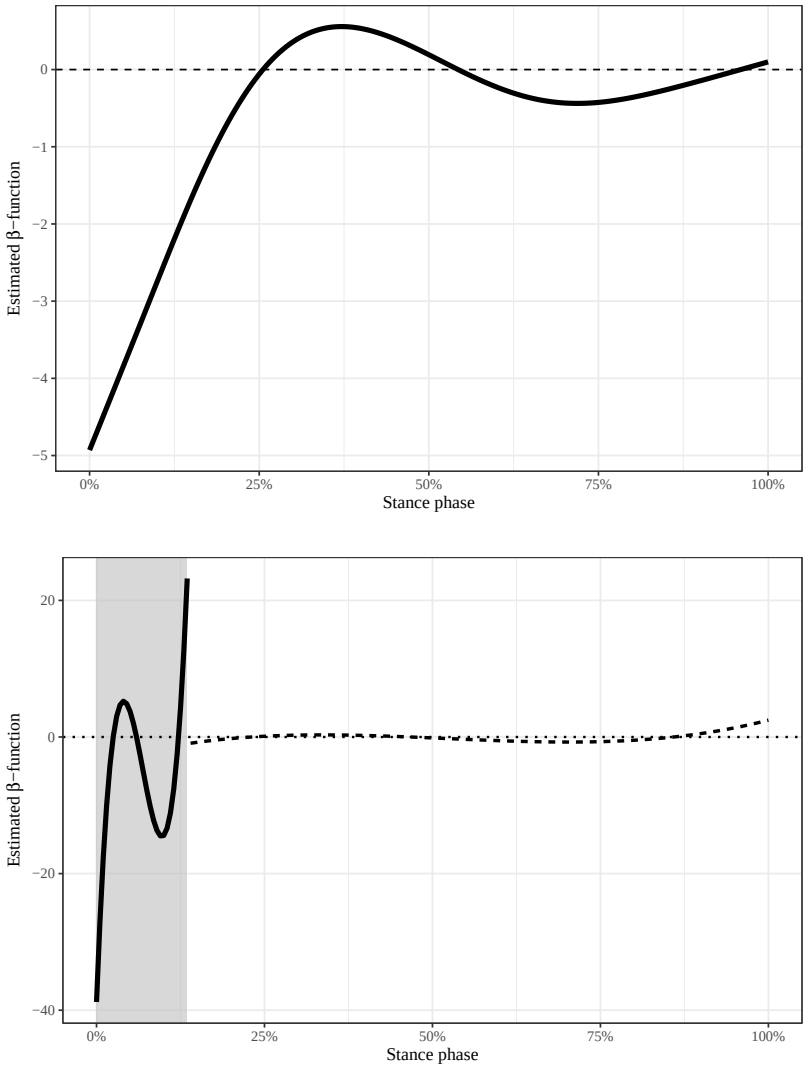


Figure 4.5.2: Estimation result for FLM where the scalar variable strike index is explained by the torque curves of barefoot runners (upper panel), and result of the sequential testing procedure (lower panel), where the rejection region is shaded in grey.

pattern are significantly related over $[0\%, 14\%]$.

4.6 Conclusion

In this paper, we propose a new testing method for the functional linear model in order to facilitate interpretability of model estimates. The proposed methodology can be helpful to detect a subregion of the functional predictors' domain where the relationship to a scalar outcome variable is statistically significant. The testing method builds on the (global) test of association in the functional linear model and performs this test on a sequence of decreasing domains.

The key results of our approach follow from the closed testing principle which ensures that the family-wise error rate (FWER) is maintained in the strong sense without any further corrections. For the theoretical analysis, we make some limiting assumptions in particular about the functional coefficient and the discretization error. Very likely, these assumptions can be relaxed to make the procedure valid in a more general setting. However, this requires a more thorough theoretical analysis that can be a future direction for this paper. In addition, the method can also be generalized to test against a general $\beta_0 \in L^2$ and, in principle, it is also not necessary to start the test at the left boundary of the domain. The procedure's numerical properties are also evaluated in a simulation exercise. We show that the global type-I error is maintained and explore the method's power properties under different scenarios with respect to sample size and signal-to-noise ratio.

We demonstrate the practical use of our method by applying it to data from a sports biomechanics experiment with recreational runners. In this example, we use a functional linear model to measure the dependency between the strike index as scalar outcome variable and torque curves measured at the ankle join as functional predictor. Using the proposed sequential test, we can show that the dependency between strike index and torque curves is statistically significant only at the very beginning of the stance phase (0%-14%) which corresponds to the first time after the foot's ground contact.

4.A Appendix

4.A.1 Proofs

Proof of Lemma 1

Under the null and the prerequisites of the lemma, we have that $\beta \in BS(k, t_l)$ with $\beta(t) = 0$ for $t \in (t_l, 1]$. The model then can then be written as

$$Y_i = \int_0^{t_l} \beta(t) X_i(t)\, \mathrm{d}t + \varepsilon_i,$$

which in matrix notation (assuming that the discretization error is negligible) for the sample is

$$\mathbf{y} = \frac{1}{p}\mathbf{X}\mathbf{B}_\mathbf{1}\theta_1 + \boldsymbol{\varepsilon}.$$

Now, observe that the kernel of the projection matrix $\mathbf{P}_\mathbf{XB} - \mathbf{P}_{\mathbf{XB_1}}$ is given by the space spanned by the columns of $\mathbf{B_1}$ and, similarly, the kernel of $\mathbf{I}_n - \mathbf{P}_\mathbf{XB}$ is the column space of $\mathbf{B}$.

It follows that

$$\mathbf{y}^\top (\mathbf{P}_\mathbf{XB} - \mathbf{P}_{\mathbf{XB_1}})\, \mathbf{y} = \boldsymbol{\varepsilon}^\top (\mathbf{P}_\mathbf{XB} - \mathbf{P}_{\mathbf{XB_1}})\, \boldsymbol{\varepsilon}$$

and

$$\mathbf{y}^\top (\mathbf{I}_n - \mathbf{P}_\mathbf{XB})\, \mathbf{y} = \boldsymbol{\varepsilon}^\top (\mathbf{I}_n - \mathbf{P}_\mathbf{XB})\, \boldsymbol{\varepsilon}$$

which are scaled (by the variance of ε) versions of χ^2-distributed variables with $k - m$ and $n - k$ degrees of freedom, since the matrices $(\mathbf{P}_\mathbf{XB} - \mathbf{P}_{\mathbf{XB_1}})$ and $(\mathbf{I}_n - \mathbf{P}_\mathbf{XB})$ have rank $k - m$ and $n - k$. It is easy to see that the matrices $(\mathbf{P}_\mathbf{XB} - \mathbf{P}_{\mathbf{XB_1}})$ and $(\mathbf{I}_n - \mathbf{P}_\mathbf{XB})$ are orthogonal projections, completing the proof. □

Proof of Theorem 1

We will show that our sequential testing procedure is an example for a closed test, such that the FWER is controlled in the strong sense by design.

The family of hypotheses in our testing procedures is $\{H_i\}_{i=1,\dots,p}$. Now, consider H_i for some $i \in \{1, \dots, p\}$ that is rejected at local level α. It is easy to see that all possible intersection hypotheses involving H_i are the hypotheses $H_1, \dots, H_i$ which, by the design of our testing procedure, must have been rejected already at local level α. With the closed testing principle (Marcus et al., 1976), it follows that H_i can be rejected at global level α. □

Table 4.A.1: Type-I error rates for DGP with $\beta = \beta_{\text{step}}$ and at global level $\alpha = 0.05$,

		$p = 100$				$p = 300$			
	n	100	250	500	1000	100	250	500	1000
γ									
0.1		0.021	0.029	0.026	0.032	0.020	0.027	0.030	0.032
0.2		0.027	0.031	0.032	0.037	0.022	0.030	0.033	0.033
0.4		0.033	0.034	0.036	0.032	0.033	0.031	0.037	0.035
0.6		0.037	0.028	0.035	0.032	0.028	0.038	0.034	0.034
0.8		0.034	0.034	0.034	0.033	0.040	0.034	0.037	0.038
0.9		0.040	0.037	0.035	0.036	0.040	0.037	0.035	0.034

4.A.2 Additional Simulation Results

To complement the simulation results of the main text, we present in this appendix additional simulation results with the same setup as in Section 4.4, but with different signal-to-noise ratios $\gamma = 0.1, 0.2, 0.4, 0.6, 0.8, 0.9$ and with functions X_i that are generated from a p-dimensional spline basis with independent standard normal coefficients. The covariance function of these curves is only different from zero close to the diagonal and it holds that $\mathbb{E}[X_i(t)X_i(s)] = 0$ for $|t - s| > 4\delta$, where δ is the distance between adjacent knots. It can be observed that the power in this setting is much higher, which is a consequence of the curves' covariance structure. Even in the case where the signal-to-noise ratio is 10 and the number of observations is 1000, the power in the setting of Section 4.4 is lower than for any of the examples shown in Figure 4.A.1.

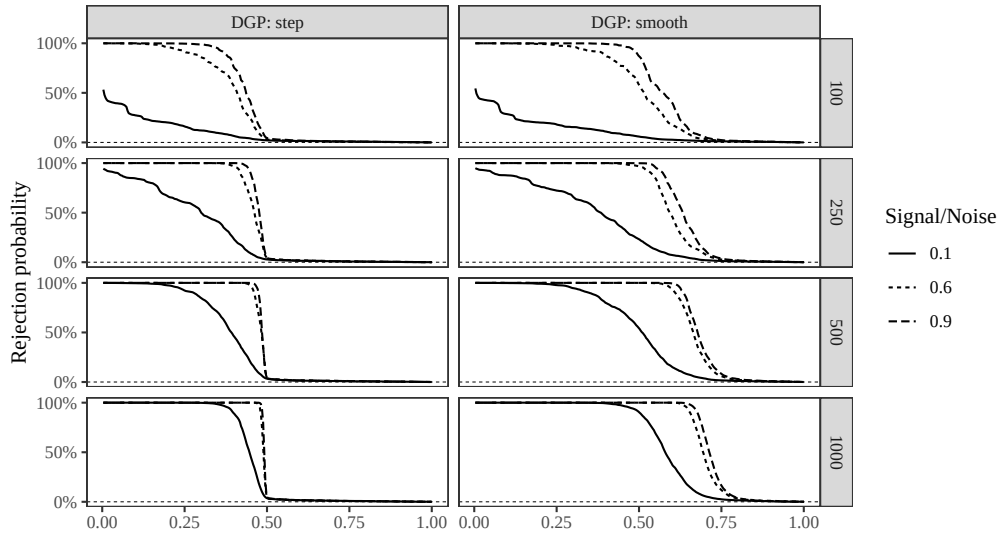


Figure 4.A.1: Rejection probabilities of the directed testing procedure for the two DGPs with $p = 300$.

Conclusion and outlook

The present thesis intends to contribute to the literature on functional data analysis (FDA) but it also aims to demonstrate that FDA provides useful methods to address empirical questions in relation to economic data. All the chapters of this thesis deal with or make use of the functional linear regression model with a scalar response variable. A fundamental problem that arises in this model is related to the interpretability of an estimate of the functional coefficient. More specifically, estimation of the functional coefficient in the scalar-on-function regression is an ill-posed inversion problem, so that regularization is necessary to get an estimate. It can be shown that it is impossible to formulate a CLT for such estimators (in the topology of L^2), hence, there is no limit distribution when the sample size becomes large. As a consequence, pointwise inference about the functional coefficient is not possible, which in turn impedes its interpretation, so practitioners can not know at which regions of the function's domain the estimate differs from zero, for instance.

Each chapter tackles different questions related to the interpretability of estimation results based on the scalar-on-function regression and, in particular, intends to make this method more applicable in the context of economic data. In this context, the thesis makes empirical but also methodological contributions which are motivated by problems involving real data. On the methodological side, we contribute to the existing literature by improving the estimation procedure for the functional regression model with points of impact, an augmented scalar-on-function regression model which can be interpreted more easily. The other methodological contribution is a new testing procedure that—under some structural assumptions about the parameter function—provides pointwise interpretability of the model estimate. The empirical contribution addresses a significant question in regional economics, namely, analysing the spatial extent of so-called human capital externalities by means of functional regression.

In the first paper (Chapter 2), an improved estimation algorithm for the functional regression model with points of impact is developed. This model can be used to decompose the effect of a functional predictor on a scalar outcome variable

into the classical global component and a time-specific local component. Motivated by a problem involving real data, where the dependency between yearly trajectories of daily impressions and aggregated clicks is modelled, we propose an improved estimation algorithm which takes into account that in finite samples, there is an ambiguity between global and local effects. More specifically, the proposed procedure decouples the estimation of the model parameters and the selection of points of impact. Additionally, instead of the classical FPCA estimator, our procedure uses a smoothing spline estimator for the functional coefficient. An extensive simulation study shows a substantial improvement in terms of the precision of the estimation.

The second paper (Chapter 3) is mainly an empirical contribution to regional economics. By drawing on very precise geo-referenced data from the German labour market, we estimate the spatial extent of productivity gains from knowledge spillovers. By describing the concentration of high-skilled workers as a function of the distance to the individual's workplace, we can use the scalar-on-function model to measure the effect of these curves on the individual's wages. Using an extensive set of fixed effects, we control for confounding mechanisms, such as local classical supply and demand effects and individual sorting. To be able to interpret the estimation results, we use the pointwise sampling variability of the functional coefficient and show, in a simulation study, that, in practice, this approach is indeed useful for deciding whether an estimate contains a signal or just reflects noise. With this approach, we are able to show that changes in the concentration of human capital influence average wages up to a distance of 15 kilometers.

The third paper (Chapter 4) is motivated by the lack of a proper testing framework in the second paper. To fill this gap, a novel sequential testing procedure is developed that can—under some assumptions on the parameter function—identify parts of the function's domain where the functional predictors have significant influence on the scalar outcome variable. The procedure is constructed in such a way that the closed testing principle applies and no additional control for the family-wise error rate is necessary. The small sample performance of the testing procedure is analysed by means of a simulation study and the test is applied to data from a sports biomechanics experiment, showing that the strike index of recreational runners is mainly associated to the torque curves at the ankle join for the very beginning of the stance phase. The implementation of the method is made available to the scientific community as an R package.

Unfortunately, at the time of this writing, the author has not been able to apply the third paper's method to the second paper's research question. Due to the ongoing pandemic situation, onsite access to the Institute for Employment

Research (IAB), where the author could in principle use the highly confidential social security data, has been restricted. On the methodological side, the procedure proposed in the third paper in its current formulation is limited to very special cases. The procedure, however, can be generalized in different directions to make it applicable to even more use cases. The version included in this thesis relies in several places on limiting assumptions that become superfluous once the underlying theory is more thoroughly formulated. Moreover, an interesting exercise would be to assess whether and how the method can help to improve the prediction performance of the functional linear model. This appears to be a promising extension for future research.

Bibliography

Abadie, A., S. Athey, W. Imbens, Guido, and J. Wooldridge (2017, November). When should you adjust standard errors for clustering? Working Paper 24003, National Bureau of Economic Research.

Abowd, John, M., F. Kramarz, and N. Margolis, David (1999). High wage workers in high wage firms. *Econometrica 67*(2), 251–333.

Acemoglu, D. (1996). A microfoundation for social increasing returns in human capital accumulation. *The Quarterly Journal of Economics 111*(3), 779–804.

Acemoglu, D. (1998). Why do new technologies complement skills? Directed technical change and wage inequality. *The Quarterly Journal of Economics 113*(4), 1055–1089.

Acemoglu, D. and J. Angrist (1999, December). How large are the social returns to education? Evidence from compulsory schooling laws. Working Paper 7444, National Bureau of Economic Research.

Acemoglu, D. and J. Angrist (2000). How large are human-capital externalities? Evidence from compulsory schooling laws. In B. S. Bernake and K. Rogoff (Eds.), *NBER Macroeconomics Annual 2000*, Volume 15, pp. 9–74. MIT Press.

Ahlfeldt, Gabriel, M., J. Redding, Stephen, M. Sturm, Daniel, and N. Wolf (2015). The economics of density: Evidence from the Berlin Wall. *Econometrica 83*(6), 2127–2189.

Altman, A. R. and I. S. Davis (2012). A kinematic method for footstrike pattern detection in barefoot and shod runners. *Gait & Posture 35*(2), 298 – 300.

Andersson, M., J. P. Larsson, and J. Wernberg (2019). The economic microgeography of diversity and specialization externalities — firm-level evidence from Swedish cities. *Research Policy 48*, 1385–1398.

Antoni, M., A. Ganzer, and P. vom Berge (2016). Sample of Integrated Labour Market Biographies (SIAB) 1975-2014. Institute of Employment Research, Nuremberg. FDZ-Datenreport 04/2016.

Arzaghi, M. and J. V. Henderson (2008). Networking off Madison Avenue. *The Review of Economic Studies 75*(4), 1011–1038.

Autor, David, H., F. Katz, Lawrence, and S. Kearney, Melissa (2008). Trends in U.S. wage inequality: Revising the revisionists. *The Review of Economics and Statistics 90*(2), 300–323.

Borjas, G. J. (2003). The labor demand curve is downward sloping: Reexamining the impact of immigration on the labor market. *The Quarterly Journal of Economics 118*(4), 1335–1374.

Caldeira, J. and H. Torrent (2017). Forecasting the US term structure of interest rates using nonparametric functional data analysis. *Journal of Forecasting 36*(1), 56–73.

Card, D., J. Heining, and P. Kline (2013). Workplace heterogeneity and the rise of West German wage inequality. *The Quarterly Journal of Economics 128*(3), 967–1015.

Card, D. and T. Lemieux (2001). Can falling supply explain the rising return to college for younger men? A cohort-based analysis. *The Quarterly Journal of Economics 116*(2), 705–746.

Cardot, H., C. Crambes, A. Kneip, and P. Sarda (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics & Data Analysis 51*(10), 4832–4848.

Cardot, H., F. Ferraty, A. Mas, and P. Sarda (2003). Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics 30*, 241–255.

Cardot, H., A. Goia, and P. Sarda (2004). Testing for no effect in functional linear regression models, some computational approaches. *Communications in Statistics-Simulation and Computation 33*(1), 179–199.

Cardot, H., A. Mas, and P. Sarda (2007). CLT in functional linear regression models. *Probability Theory and Related Fields 138*(3-4), 325–361.

Cavanagh, P. R. and M. A. Lafortune (1980). Ground reaction forces in distance running. *Journal of Biomechanics 13*(5), 397 – 406.

Chiou, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics 6*(4), 1588–1614.

Choi, H. and M. Reimherr (2018). A geometric approach to confidence regions and bands for functional parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(1), 239–260.

Ciccone, A. and G. Peri (2005). Long-run substitutability between more and less educated workers: Evidence from U.S. states, 1950-1990. *The Review of Economics and Statistics 87*(4), 652–663.

Ciccone, A. and G. Peri (2006). Identifying human-capital externalities: Theory with applications. *The Review of Economic Studies 73*(2), 381–412.

Cornelissen, T., C. Dustmann, and U. Schönberg (2017). Peer effects in the workplace. *American Economic Review 107*(2), 425–456.

Crambes, C., A. Kneip, and P. Sarda (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics 37*(1), 35–72.

Dauth, W. and P. Haller (2018, Oct). Berufliches Pendeln zwischen Wohn- und Arbeitsort. IAB-Kurzbericht, Institute for Employment Research (IAB), Nuremberg.

Davis, Donald, R. and I. Dingel, Jonathan (2019). A spatial knowledge economy. *American Economic Review 109*(1), 153–70.

De Boor, C. (1978). *A practical guide to splines*, Volume 27. Springer.

De la Roca, J. and D. Puga (2017). Learning by working in big cities. *Review of Economic Studies 84*(1), 106–142.

Doty, D., K. Sruoginis, and D. Silverman (2016). IAB/PwC Internet advertising revenue report. `www.iab.com/adrevenuereport`.

Dustmann, C., J. Ludsteck, and U. Schönberg (2009). Revisiting the German wage structure. *The Quarterly Journal of Economics 124*(2), 843–881.

Eppelsheimer, J. and J. Möller (2019). Human capital spillovers and the churning phenomenon: Analysing wage effects from gross in-and outflows of high-skilled workers. *Regional Science and Urban Economics 78*, 103461.

Eppelsheimer, J. and C. Rust (2020). The spatial decay of human capital externalities - a functional regression approach with precise geo-referenced data. IAB-Discussion Paper 21, Institute for Employment Research.

Faggio, G. (2019). Relocation of public sector workers: Evaluating a place-based policy. *Journal of Urban Economics 111*, 53–75.

Faggio, G., T. Schluter, and P. vom Berge (2019). Interaction of public and private employment: Evidence from a German government move. Working Papers 19/09, Department of Economics, City University London.

Ferraty, F., P. Hall, and P. Vieu (2010). Most-predictive design points for functional data predictors. *Biometrika 97*(4), 807–824.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis - Theory and Practice*. New York: Spinger.

Fitzenberger, B., A. Osikominu, and R. Völter (2005). Imputation rules to improve the education variable in the IAB employment subsample. Institute of Employment Research, Nuremberg. FDZ-Methodenreport 03/2005.

Fraiman, R., Y. Gimenez, and M. Svarc (2016). Feature selection for functional data. *Journal of Multivariate Analysis 146*, 191–208.

Fu, S. (2007). Smart Café Cities: Testing human capital externalities in the Boston metropolitan area. *Journal of Urban Economics 61*(1), 86–111.

Geddes, B. (2014). *Advanced Google AdWords*. John Wiley & Sons.

Gellar, J. E., E. Colantuoni, D. M. Needham, and C. M. Crainiceanu (2014). Variable-domain functional regression for modeling ICU data. *Journal of the American Statistical Association 109*(508), 1425–1439.

Gibbons, S., H. G. Overman, and M. Sarvimäki (2021). The local economic impacts of regeneration projects: Evidence from UK's single regeneration budget. *Journal of Urban Economics 122*, 103315.

Glaeser, Edward, L. and C. Mare, David (2001). Cities and skills. *Journal of Labor Economics 19*(2), 316–342.

Goldsmith, J., J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich (2010). Penalized functional regression. *Journal of Computational and Graphical Statistics 20*(4), 830–851.

Goldsmith, J., C. M. Crainiceanu, B. Caffo, and D. Reich (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 61*(3), 453–469.

González-Manteiga, W., G. González-Rodríguez, A. Martínez-Calvo, and E. García-Portugués (2012). Bootstrap independence test for functional linear models. *arXiv preprint arXiv:1210.1072*.

Gromenko, O., P. Kokoszka, and J. Sojka (2017). Evaluation of the cooling trend in the ionosphere using functional regression with incomplete curves. *The Annals of Applied Statistics 11*(2), 898–918.

Hall, P. and G. Hooker (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society Series B 78*(3), 637–653.

Hall, P. and J. L. Horowitz (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics 35*(1), 70–91.

Hall, P. and M. Hosseini-Nasab (2006). On properties of functional principal components. *Journal of The Royak Statistical Society Series B 68*(1), 106–126.

Hamacher, D., K. Hollander, and A. Zech (2016). Effects of ankle instability on running gait ankle angles and its variability in young adults. *Clinical Biomechanics 33*, 73 – 78.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized additive models*, Volume 43. CRC press.

Heuermann, D. (2011). Human capital externalities in Western Germany. *Spatial Economic Analysis 6*(2), 139–165.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. Springer.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Imaizumi, M. and K. Kato (2019). A simple method to construct confidence bands in funtional linear regression. *Statistica Sinica 29*(4), 2055–2081.

Katz, Lawrence, F. and M. Murphy, Kevin (1992). Changes in relative wages, 1963-1987: Supply and demand factors. *The Quarterly Journal of Economics 107*(1), 35–78.

Kneip, A., D. Poss, and P. Sarda (2016). Functional linear regression with points of impact. *The Annals of Statistics 44*(1), 1–30.

Koeppe, R., J. Zhu, B. Nan, and X. Wang (2014). Regularized 3d functional regression for brain image data via haar wavelets. *The Annals of Applied Statistics 8*(2), 1045–1064.

Kong, D., A.-M. Staicu, and A. Maity (2016). Classical testing in functional linear models. *Journal of Nonparametric Statistics 28*(4), 813–838.

Kosfeld, R. and A. Werner (2012). Deutsche Arbeitsmarktregionen - Neuabgrenzung nach den Kreisgebietsreformen 2007-2011. *Raumforschung und Raumordnung 70*(1), 46–64.

Krusell, P., E. Ohanian, Lee, J.-V. Rios-Rull, and L. Violante, Giovanni (2000). Capital-skill complementarity and inequality: A macroeconomic analysis. *Econometrica 68*(5), 1029–1053.

Liebl, D., S. Rameseder, and C. Rust (2020). Improving estimation in functional linear regression with points of impact: Insights into google adwords. *Journal of Computational and Graphical Statistics 29*(4), 814–826.

Liebl, D., S. Willwacher, J. Hamill, and G.-P. Brüggemann (2014). Ankle plantarflexion strength in rearfoot and forefoot runners: A novel clusteranalytic approach. *Human Movement Science 35*, 104 – 120.

Lindquist, M. A. and I. W. McKeague (2009). Logistic regression with brownian-like predictors. *Journal of the American Statistical Association 104*(488), 1575–1585.

Liu, B. and H.-G. Müller (2008). *Functional data analysis for sparse auction data*, pp. 269–290. Statistical Methods in eCommerce Research (eds W. Jank and G. Shmueli) John Wiley & Sons.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics 22*, 3–42.

Manning, A. and B. Petrongolo (2017). How local are labor markets? Evidence from a spatial job search model. *American Economic Review 107*(10), 2877–2907.

Marcus, R., E. Peritz, and G. K. Ruben (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika 63*(3), 655–660.

Maronna, R. A. and V. J. Yohai (2013). Robust functional linear regression based on splines. *Computational Statistics & Data Analysis 65*, 46–55.

Marshall, A. (1890). *Principles of Economics*. London: MacMillan.

Matsui, H. and S. Konishi (2011). Variable selection for functional regression models via the L1 regularization. *Computational Statistics & Data Analysis 55*(12), 3304–3310.

McKeague, I. W. and B. Sen (2010). Fractals with point impact in functional linear regression. *The Annals of Statistics 38*(4), 2559–2586.

Mellander, C., K. Stolarick, and J. Lobo (2017). Distinguishing neighbourhood and workplace network effects on individual income: Evidence from Sweden. *Regional Studies 51*(11), 1652–1664.

Moretti, E. (2004). Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of Econometrics 121*, 175–212.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application 2*, 321–359.

Müller, H.-G. and U. Stadtmüller (2005). Generalized functional linear models. *The Annals of Statistics 33*(2), 774–805.

Muravyev, A. (2008). Human capital externalities: Evidence from the transition economy of Russia. *Economics of Transition 16*(3), 415–443.

Nelson, R. R. and E. S. Phelps (1966). Investment in humans, technological diffusion, and economic growth. *The American Economic Review 56*(1/2), 69–75.

Poß, D., D. Liebl, A. Kneip, H. Eisenbarth, T. D. Wager, and L. Feldman Barrett (2020). Super-consistent estimation of points of impact in nonparametric regression with functional predictors. *Working Paper arXiv:1905.09021*.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2. ed.). New York: Springer.

Ramsay, James, O. and B. Ramsey, James (2002). Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics 107*(1-2), 327–344.

Rauch, J. E. (1993). Productivity gains from geographic concentration of human capital: Evidence from the cities. *Journal of Urban Economics 34*(3), 380–400.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reddy, S. K. and M. Dass (2006). Modeling on-line art auction dynamics using functional data analysis. *Statistical Science 21*(2), 179–193.

Reiss, P. T., J. Goldsmith, H. L. Shang, and R. T. Ogden (2017). Methods for scalar-on-function regression. *International Statistical Review 85*(2), 228–249.

Reiss, P. T. and R. T. Ogden (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association 102*(479), 984–996.

Rosenthal, S. S. and W. C. Strange (2008). The attenuation of human capital spillovers. *Journal of Urban Economics 64*(2), 373–389.

Schmucker, A., S. Seth, J. Ludsteck, J. Eberle, and A. Ganzer (2016). The establishment history panel 1975-2014. Institute of Employment Research, Nuremberg. FDZ-Methodenreport 03/2016.

Swihart, B. J., J. Goldsmith, and C. M. Crainiceanu (2014). Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics 56*(4), 483–493.

Tekbudak, M. Y., M. Alfaro-Córdoba, A. Maity, and A.-M. Staicu (2019). A comparison of testing methods in scalar-on-function regression. *AStA Advances in Statistical Analysis 103*, 411–436.

Torrecilla, J. L., J. R. Berrendero, and A. Cuevas (2016). Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica 26*(2), 619–638.

Ullah, S. and C. F. Finch (2013, Mar). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology 13*(1), 43.

Vanrenterghem, J., E. Venables, T. Pataky, and M. A. Robinson (2012). The effect of running speed on knee mechanical loading in females during side cutting. *Journal of Biomechanics 45*(14), 2444 – 2449.

Verstraten, P. (2018, September). The scope of the external return to higher education. Discussion Paper 381, CPB Netherlands Bureau for Economic Policy Analysis.

Wang, S., W. Jank, and G. Shmueli (2008). Explaining and forecasting online auction prices and their dynamics using functional data analysis. *Journal of Business & Economic Statistics 26*(2), 144–160.

Wang, S., W. Jank, G. Shmueli, and P. Smith (2008). Modeling price dynamics in eBay auctions using differential equations. *Journal of the American Statistical Association 103*(483), 1100–1118.

Warmenhoven, J., S. Cobley, C. Draper, A. Harrison, N. Bargary, and R. Smith (2019). Considerations for the use of functional principal components analysis in sports biomechanics: examples from on-water rowing. *Sports Biomechanics 18*(3), 317–341. PMID: 29141500.

Zhang, S., W. Jank, and G. Shmueli (2010). Real-time forecasting of online auctions via functional k-nearest neighbors. *International Journal of Forecasting 26*(4), 666–683.