Eppelsheimer, Johann; Rust, Christoph

Conference Paper

The geographic reach of knowledge spillovers: A functional regression approach with precise geo-referenced data

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2019: 30 Jahre Mauerfall - Demokratie und Marktwirtschaft - Session: Econometrics - Forecasting II, No. F18-V3

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Eppelsheimer, Johann; Rust, Christoph (2019) : The geographic reach of knowledge spillovers: A functional regression approach with precise geo-referenced data, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2019: 30 Jahre Mauerfall - Demokratie und Marktwirtschaft - Session: Econometrics - Forecasting II, No. F18-V3, ZBW - Leibniz-Informationszentrum Wirtschaft, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/203667

The Geographic Reach of Knowledge Spillovers
A Functional Regression Approach with Precise Geo-Referenced Data

Johann Eppelsheimer* and Christoph Rust†

February, 2019

Version 0.4.4

Preliminary, not to be cited or circulated.

This paper applies functional regression to precise geo-coded register data to measure productivity spillovers from high-skilled workers. We use a smoothing splines estimator to model the spatial distribution of high-skilled workers as continuous curves. Our rich panel data allows us to address spatial sorting of workers and the entanglement of spillover and supply effects with an extensive set of time-varying fixed effects. Our estimates reveal that spillovers from high-skilled workers attenuate monotonously with distance. Effects disappear after approximately 20 kilometers. Furthermore, our findings illustrate the benefits of applying functional regression to modern (spatial) economic data.

Keywords: Knowledge Spillovers, Human Capital Externalities, Functional Regression, Georeferenced Data, Education, Wages

JEL classification: D62; J24; J31; R10; R23

Acknowledgments: The authors thank Sebastian Bähr, Annette Bergemann, Gerard van den Berg, Matthias Dorner, Andreas Eberl, Peter Haller, Christian Merkl, Florian Lehmer, Joachim Möller, Malte Reichelt, Uta Schönberg, Heiko Stüber, Rolf Tschernig, Paul Verstraten, Erwin Winkler, Anthony Yezer and participants of ERSA Congress 2018, ERSA Congress 2017, NARSC Conference 2017, Statistical Week 2018, and seminars at the Institute for Employment Research (IAB) Nuremberg and the University of Regensburg for many helpful comments and suggestions. Johann Eppelsheimer acknowledges financial support from the graduate program of the IAB and the University of Erlangen-Nuremberg (GradAB).

1 Introduction

Workers interact with co-workers within and across firms. Through these interactions they share their knowledge, discuss ideas and adopt novel technologies. All these interactions potentially increase the worker's productivity and are a major source of agglomeration economies (Acemoglu, 1996; Lucas, 1988; Marshall, 1890). Extensive empirical research underpins the existence of such 'knowledge spillovers' within predefined geographical boundaries (Cornelissen et al., 2017; Ciccone and Peri, 2006; Moretti, 2004; Rauch, 1993). However, little is known about the exact spatial extent of knowledge spillovers. Knowledge spillovers should diminish with distance for

*Institute for Employment Research Nuremberg (IAB) (e-mail: johann.eppelsheimer@iab.de)
†University of Regensburg

several reasons. For instance, distance raises costs of planed social interactions, such as meetings. Further, distance lowers the likelihood of random encounters. Moreover, considering information flows within a network of individuals, the likelihood of transmitting information between individuals decreases with the number of intermediaries. Because distance should generally raise the number of intermediaries also the information flow between individuals should attenuate with distance. Despite the relevance of knowledge spillovers for policy-makers and entrepreneurs only recently available precise geo-data and methodological advances allow to measure the exact spatial reach and intensity of these effects.

Previous empirical studies provide first evidence for spatially decreasing knowledge spillovers. For instance, using cross-sectional data form the U.S. Rosenthal and Strange (2008) construct concentric rings around workers that measure the concentration of human capital within 5, 5 to 25, 25 to 50 and 50 to 100 miles. To explore the attenuation of knowledge spillovers they regress individual wages on the concentration of human capital within these rings. They find that knowledge spillovers from rings closer by are notably larger than spillovers from rings further out. Another study by Fu (2007) adopts the strategy of Rosenthal and Strange (2008) to analyze cross-sectional data from the Boston metropolitan area. More precise geo-coded data allows Fu (2007) to measure the concentration of human capital within finer rings (i.e., 0-1.5, 1.5-3, 3-6 and 6-9 miles). Fu (2007) provides evidence that knowledge spillovers may already decay after three miles. Although these studies present evidence for the spatial attenuation of knowledge spillovers, the exact attenuation of effects remains unclear because the literature is either constrained by relatively imprecise geo-data or specific data on a small area. Furthermore, empirical evidence is restricted to cross-sectional data, which complicates causal inference. Additionally, the empirical literature mostly overlooks that spillover effects from high-skilled workers are entangled with conventional labor market supply and demand effects (Katz and Murphy, 1992; Card and Lemieux, 2001; Borjas, 2003; Moretti, 2004; Ciccone and Peri, 2006).

In this paper we analyze the spatial reach and intensity of knowledge spillovers from high-skilled workers by drawing on a large and novel administrative micro panel data set that features the exact coordinates of nearly all German establishments and rich information on individual workers over more than one decade. Our aim is to estimate spillovers from high-skilled workers on individual wages.

In order to fully exploit the information that is given by the exact geocodes of the working places, we take a fresh methodological approach to measure the magnitude of knowledge spillovers with respect to distance in a continuous manner. Recently developed methodologies in Functional Data Analysis (FDA) provide a particularly suitable framework for our purposes. FDA is a branch in statistics devoted to the development of methods for random variables with a functional nature, such as curves or surfaces over a continuous domain. Typical examples are temperature curves, growth curves or the continuous evolution of stock prices over time. The main benefit of the functional view compared to a multivariate one is that data points which are located close to each other are somehow related—using this information makes FDA more efficient than standard multivariate methodologies.

While statisticians employ FDA for a wide range of applications (see Ullah and Finch, 2013 for a systematic overview, readers with 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), FDA is still applied quite rarely in economic applications. This paper, therefore, illustrates the potential of FDA in economic research with high-dimensional variables. Our model framework relies on the functional linear regression model where a scalar outcome variable (log-wage in our situation) is regressed on observations of a functional random variable (share of high-skilled workers depending on distance to the focal worker). For our purposes, we augment the classical scalar-on-function regression model to incorporate also further scalar-valued explanatory variables and use an estimation procedure, suggested by Crambes et al. (2009), that is based on smoothing splines. The smoothing splines estimator has the useful property that it allows a spline-based expansion of the function-valued spillover parameter instead of the classical representation in a function space spanned by the leading eigenfunctions of the random curve's empirical covariance operator. As a consequence, the resulting estimate does not depend on the correllation structure of the functional random variable and can be modeled much more flexibly. We estimate a spatial spillover function by evaluating the distribution of high-skilled workers every 500 meters within a range of 50 kilometers.

Two major challenges in identifying regional knowledge spillovers are confounding labor market supply and demand effects and sorting of high-skilled workers into high-wage regions. We 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 propose that an increase in the share of high-skilled workers raises (lowers) wages of high-skilled (low-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. We disentangle spillover from supply and demand effects by exploiting the different spatial nature of the two effects. While supply and demand effects are plausibly common within local labor markets (i.e., supply and demand effects originated in one part of the city uniformly affect wages in the whole 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 are able to purge spillover 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-year fixed effects in our econometric specification (i.e., a specific intercept for every labor market area in every year). Because supply and demand effects contrarily affect 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 on the firm level (Abowd et al., 1999; Card et al., 2013), we deal with sorting of high-skilled workers into high-wage regions (Acemoglu and Angrist, 2000) with 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 account for locational advantages within regions (e.g., proximity to infrastructure and facilities) and unobserved individual heterogeneity with worker-firm match fixed effects, respectively.

In this general setting, we find that spillover effects indeed diminish with distance and a positive spillover effect is measurable up to 20 kilometers for a given location. An increase of the share of highly skilled workers within that region by one percentage point raises average wages by 0.2 percent. The effect of an increase closer by (within the distance of up to 10km), however, is double as high as the effect of an incease further away (between 10km and 20km). To the best of our knowledge, we are the first to address this question in such a general context and provide generalizable results. We believe that our results will be important not only for academia but also of high relevance for professionals such as city planners.

The remainder of the paper is organized as follows. The next section explains the estimator and our identification strategy. Section 3 summarizes the data and the construction of the sample for the empirical analysis. Section 4 presents our main findings, and section 5 concludes.

2 Estimation strategy

This paper seeks to measure the spatial attenuation and reach of knowledge spillover. Therefore, we aim to describe the share of high-skilled workers around establishments as continuous curves and model a spillover function that depends on distance. In the following we explain the estimator, discuss statistical inference, and describe our representation of the share of high-skilled workers as curves. Finally, we specify our identification strategy that addresses endogenous sorting of workers and confounding labor market supply and demand effects.

2.1 The estimator

A suitable modeling approach for the function-valued spillover function is available by recently developed framework in Functional Data Analysis. In particular, we build on the smoothing spline estimator in the functional linear regression model proposed by Crambes et al. (2009) in order to estimate the functional spillover parameter. Since we also want to include additional explanatory variables, we augment this estimator to incorporate also the effect of further scalar-valued predictor variables. The classical functional linear regression model with a scalar response is given by

$$Y_i = \int_0^1 \beta(t) X_i(t)\,\mathrm{d}t + \varepsilon_i, \tag{1}$$

where Y_i is a scalar-valued dependent variable, $X_i \in L^2([a,b])$ are iid distributed random functions defined on a common domain which we set to $[0,1]$ without loss of generality. The error term ε_i is independently distributed and has mean zero and homoscedastic variance (the latter however can be relaxed). The function-valued coefficient parameter $\beta \in L^2([0,1])$ gives the influence X_i has on Y_i. Model (1) has received a lot of attention in the literature on Functional Data Analysis (see Morris, 2015, for an overview). The classical estimation of β bases on the Karhunen-Loève decomposition of the empirical covariance operator of the observed curves X_i (also known as functional principal component (FPC) estimator) and therefore the expansion of such an estimator $\widehat{\beta}$ heavily depends on the random curves' correlation structure. The approach taken here differs in the way that the basis functions are independent of the curves X_i which results in a more flexible function space for modeling β. Regularization in the situation of an FPC based estimator in practice is done by truncating the Karhunen-Loève basis and therefore

is of discrete nature while the smoothing spline approach reduces complexity by imposing a real-valued penalty on a candidate's curvatures. From an asymptotic point of view, both estimators have minimax-optimal convergence rates (Hall and Horowitz, 2007; Crambes et al., 2009), however, it turns out, that the estimator based on smoothing splines performs better in real data examples of practical relevance.

In order to account for the influence of further (scalar-valued) explanatory variables, model (1) can be augmented by writing

$$Y_i = \int_0^1 \beta(t) X_i(t)\,\mathrm{d}t + Z_i'\gamma + \varepsilon_i, \tag{2}$$

where Z_i is a k-vector of explanatory variables and the coefficient vector γ holds the corresponding marginal effects.

In order to jointly estimate the slope parameters β and γ we augment the smoothing spline estimator to incorporate also additional scalar-valued explanatory variables. Let $\mathbf{X}$ be the $n \times p$ matrix holding all the n curves $X_i(t)$ observed at the p grid values $t_1, \ldots, t_p$ and let $\mathbf{Y}$ be the n-vector holding observations of the dependent variable. The penalizes least-square estimator of β, evaluated at the grid values $t_1, \ldots, t_p$ and for given smoothing paramter $\rho \in \mathbb{R}^+$, is then given by

$$\left(\widehat{\beta}(t_1), \ldots, \widehat{\beta}(t_p)\right) = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}'\mathbf{X} + \rho\mathbf{A}\right)^{-1}\mathbf{X}'\mathbf{Y}, \tag{3}$$

where the nonstandard penalty matrix $\mathbf{A} = \mathbf{P} + p\mathbf{A}^*$ was introduced by Crambes et al. (2009) and is a combination of a classical regularization matrix $\mathbf{A}^* \in \mathbb{R}^{p\times p}$ and a nonstandard projection matrix $\mathbf{P} \in \mathbb{R}^{p\times p}$. The latter is introduced to ensure 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 $\mathbf{W} = (t_j^l)_{j,l} \in \mathbb{R}^{p\times m}$ and $2m$ is the polonomial order of the employed spline basis. We model β as an expansion of *cubic* splines, thus, we set $m = 2$. The regularization matrix $\mathbf{A}^*$ is defined as usual by

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

where $\mathbf{B}$ denotes the $p \times p$ matrix of the p basis functions, evaluated at the p grid values, and $\mathbf{b}^{(2)}(t)$ is for given value of $t \in [0,1]$ a p-vector of second derivatives for each of the p basis functions.

The estimator for β and γ in model (2) can be stated as follows: 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. The compound estimator of β and γ is then given as follows:

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}(t_1), \ldots, \widehat{\beta}(t_p), \widehat{\gamma}_1, \ldots, \widehat{\gamma}_k\right) = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}_\mathbf{Z}'\mathbf{X}_\mathbf{Z} + \rho\mathbf{A}_\mathbf{Z}\right)^{-1}\mathbf{X}_\mathbf{Z}'\mathbf{Y}, \tag{4}$$

where the extended penalty matrix $\mathbf{A_Z}$ is constructed by appending k zero columns and k zero rows to $\mathbf{A}$

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

since additional variables do not load into the roughness penalty.

The estimator, however, depends on a choice of the smoothing parameter ρ which controls the complexity of the resulting estimate of the function-valued slope parameter β. The specific value of ρ, however, has no direct interpretation with respect to the complexity of $\widehat{\beta}$. A well-established and interpretable measure for the complexity of the estimate $\widehat{\beta}$, though, is given by the so-called *effective number of degrees of freedom* (edf), defined by

$$\text{edf}(\rho) = \text{Tr}\big(\mathbf{H}_{\mathbf{Z}}^{\rho}\big), \tag{5}$$

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 *hat matrix* of model (2), also called smoother matrix in the context of smoothing spline estimation. One now can predefine a value for edf and use (5) to determine the corresponding value of the smoothing parameter ρ. In practice, when there is no prior knowledge on how complex the estimate should be, ρ is often obtained by minimizing a Generalized Cross-Validation criterion. Crambes et al. (2009) propose to use

$$\text{GCV}(\rho) = \frac{\frac{1}{n}\left\|\mathbf{Y} - \mathbf{H}_{\mathbf{Z}}^{\rho}\mathbf{Y}\right\|^2}{\left(1 - \frac{1}{n}\text{Tr}\big(\mathbf{H}_{\mathbf{Z}}^{\rho}\big)\right)^2}. \tag{6}$$

In our situation, however, it is reasonable to choose the first approach and use a predefined number of effective degrees of freedom, because theory suggests a monotonically declining effect over distance, the resulting estimate therefore should not be very demanding in terms of degrees of freedom. Later on, we will also show, that our choice is in line with the complexity chosen by GCV.

2.2 Inference

In order to construct a confidence band around the estimate of β and obtain t-statistics for the elements of γ, one can follow the classical approach lined out in Ramsay and Silverman (2005), equation (15.16) and compute an approximation of the variance-covariance matrix of the compound estimator by

$$\text{Var}\big(\widehat{\boldsymbol{\beta}}\big) = \frac{1}{n^2}\left(\frac{1}{np}\mathbf{X}_{\mathbf{Z}}'\mathbf{X}_{\mathbf{Z}} + \rho\mathbf{A}_{\mathbf{Z}}\right)^{-1}\mathbf{X}_{\mathbf{Z}}'\widehat{\boldsymbol{\Omega}}\mathbf{X}_{\mathbf{Z}}\left(\frac{1}{np}\mathbf{X}_{\mathbf{Z}}'\mathbf{X}_{\mathbf{Z}} + \rho\mathbf{A}_{\mathbf{Z}}\right)^{-1}, \tag{7}$$

with $\widehat{\boldsymbol{\Omega}}$ being an appropriate estimator for the variance-covariance matrix of the error term ε. This general formulation explicitly allows to robustify inference, for instance compute clustered standar-errors. Confidence Intervals (CI) can then be built by taking the square-root of the corresponding diagonal entry of $\text{Var}\big(\widehat{\boldsymbol{\beta}}\big)$ and multiplying it with normal quantiles for a given significance level; t-statistics can be computed as usual.

2.3 Calculation of curves

One essential part of our analysis is based on the representation of the density of high-skilled workers for each individual and distance as random curves. These functions are not directly observed and have to be calculated from geocoded workplace data. We predefined an equidistant

grid for distance values $t_1, \ldots, t_p$, to compute the value of the functions X_i at these values in the following way (based on local kernel smoother/local high-skilled density):

$$X_i(t_j) = \frac{\sum_{l=1}^{n} \mathbf{1}_{\{d_{i,l} \in [t_j - h; t_j + h] \wedge hs_l\}}}{\sum_{l=1}^{n} \mathbf{1}_{\{d_{i,l} \in [t_j - h; t_j + h]\}}}, \tag{8}$$

where $\mathbf{1}_{\{expr\}}$ is the indicator function, evaluating to one if *expr* is true end zero otherwise. $d_{i,l}$ is the euclidean distance on the earth surface between worker i and worker j, and hs_l evaluates to true if worker l is a high-skilled worker, false otherwise. Put differently, the value of the curve X_i at point t_j is given by the share of high-skilled workers on all workers in the distance window $[t_j - h, t_j + h]$, where h is a predefined bandwidth. To balance analytical precision and computational costs we chose a bandwidth of $h = 250$ meters and calculate $X_i(t_j)$ up to a distance of $t_j = 50$ kilometers.

We measure the density of high-skilled workers as shares instead of e.g., the absolute number of high-skilled workers or high-skilled workers per square meter for the following reasons. Firstly, because the geographic area covered by $[t_j - h, t_j + h]$ increases with distance t_j also the absolute number of high-skilled workers that could potentially populate that area increases with distance. Thus, using absolute numbers, the intensity of high-skilled workers would increase with distance by definition and would therefore not not give comparable values across space. Secondly, as the data shows, the proportion of inhabited land decreases with t_j. As knowledge transfers appear only in inhabited areas, using high-skilled workers per square meter would decrease the intensity of human capital with distance by construction. Thus, also high-skilled workers per square meter would not suffice to study the intensity of high-skilled workers across distance. Contrarily, the number of workers within $[t_j - h, t_j + h]$ is a reasonable unit of measurement of the *de facto* populated area, which, thinking of skyscrapers, not only covers actual land use but also intensity of land use. Therefore, we measure the intensity of human capital as high-skilled workers relative to the total number of workers (i.e., we take the share of high-skilled workers).

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, changes in the supply of high-skilled labor therefore entail a market mechanism that affects wages. In particular, due to labor market demand and supply effects an increase in the share of high-skilled workers in the labor market depresses wages of high-skilled workers and raise wages of low-skilled workers.

So far, the overall supply of high-skilled labor l_r is part of the error term of equation (2): $\varepsilon_i = u_i + l_r$. Obviously, a change in the share of high-skilled workers $X_i(t)$ at some distance t translates into a change in the overall supply of high-skilled workers within in the local labor market l_r. Thus, equation (2) yields a biased estimate of $\beta(t)$. More precisely, ignoring l_r in equation (2) exerts a uniform shift of $\beta(t)$. The direction of the shift depends on the relative

amount of high- and low-skilled workers in the labor market and the elasticity of substitution between the two groups.

To disentangle knowledge spillover from labor market supply and demand effects we exploit the different spatial nature of the two effects. On the one hand, the intensity of knowledge spillover should 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 plausibly affect the local labor market uniformly. Thus, independent of the exact location, a shift in the supply of high-skilled labor affects wages within a local labor market homogeneously. We are thus able to nullify labor market supply and demand effects by eliminating all variation that is common within local labor markets without removing intra-regional variation from knowledge spillovers.

As labor market supply and demand shifts vary over time and the direction of the shift idiosyncratically affects high- and low-skilled individuals we expand equation (2) with time-varying labor-market-area fixed effects for each skill group π_{rys} (i.e., an intercept for each labor market area and skill group in every year). Our full estimation equation is:

$$Y_{ifyro} = \int_0^1 \beta(t) X_{fy}(t)\,\mathrm{d}t + Z_i'\gamma + \theta_{if} + \tau_y + \omega_o + \pi_{rys} + u_{ifyro}. \tag{9}$$

Here Y_{ifyro} is the individual log-wage of worker i in year y and $X_{fy}(t)$ is the share of high-skilled workers, described as a continuous curve around the workplace that depends on distance t. $\beta(t)$ is the associated spillover function we aim to retrieve from the data. The model holds constant time-varying observable individual, establishment and regional characteristics Z_{iy} as well as a series of fixed effects. θ_{if} is a worker-firm match fixed effect, τ_y is a year fixed effect and ω_o is an occupation fixed effect.

Another challenge in identifying regional knowledge spillovers is endogenous sorting of workers (Acemoglu and Angrist, 2000). In our application sorting threatens identification on two levels: first, on the level of treated individuals (i.e., individuals whose wages we observe), second, on the treatment level itself (i.e., the spatial density of high-skilled workers). Inspired by Cornelissen et al. (2017) we deal with worker sorting with an exhaustive set of fixed effects.

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 create a spurious relationship between wages and the local share of high-skilled workers. Thus, 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 additionally eliminate other unobservable characteristics of workers and firms, such as personal traits and locational advantages (e.g., proximity to infrastructure).

Furthermore, high-wage areas might attract high-skilled workers, which would reverse the direction of causality in equation (9). However, as worker-firm match fixed effects (θ_{if}) nullify permanent locational advantages they also eliminate general push- and pull factors that might draw high-skilled workers into high-wage regions. Moretti (2004) additionally raises the concern that also temporal shocks in the local labor market might affect the concentration of high-skilled

workers. We address this issue with time-varying labor-market-area fixed effects (π_{rys}). Because time-varying labor-market-area fixed effects remove temporal variation in the supply of high-skilled labor they also remove supply changes due to temporal shocks.

In summary, equation (9) allows us to estimate knowledge spillover that are unrelated to labor market demand and supply effects and endogenous sorting of individuals. The remaining variation of $X_{fy}(t)$ in equation (9) stems from temporal intra-regional changes in the concentration of high-skilled workers.

3 Data and descriptive statistics

3.1 Data

For our empirical analysis we combine administrative data on almost all German firms and rich data of a representative sample of workers over a period of 15 years. Our panel data includes exact geo-coordinates of establishments and therefore allows to describe the distribution of high-skilled workers as geospatial functions around workers. Within a distance 50 kilometers we evaluate the share of high-skilled workers every 500 meters.

Two of our main meso-level data sources are the *Establishment History Panel* (BHP 7516) and *IEB GEO* from the Institute for Employment Research (IAB).[1] The Establishment History Panel comprises all German establishments with at least one employee on June 30 each year. Among others, the dataset gives establishment-level information on the number of employees and the number of employees with tertiary education. To measure the distribution of high-skilled workers we classify employees with a degree from a university or a degree from a university of applied science as high skilled.[2]

We expand the dataset with exact geo-coordinates from IEB GEO. IEB GEO is a novel data source that stores 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 gives a detailed description of the geographic landscape of workplaces. In some cases, however, firms might actually have multiple establishments within the same industry within one municipality, which they do not report. In these cases we cannot ascertain 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 geospatial functions around establishments as described in Section 2.3.

For the econometric analysis of knowledge spillovers we merge the constructed geo-spatial functions of high-skilled workers with micro-level data from the *Sample of Integrated Labour Market Biographies* (SIAB 7514).[3] The Sample of Integrated Labour Market Biographies is a 2%

[1] For a detailed description of the Establishment History Panel refer to Schmucker et al. (2016)

[2] There are are two types of universities in the German tertiary education system: traditional universities and universities of applied science (*Fachhochschulen*). Compared to traditional universities, universities of applied science focus more practical topics than traditional universities. 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.

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

random sample of social security records. To join the individual-level data to the establishment-level data we transform the spell dataset into a yearly panel with June 30 as reference date.

Because employers face legal sanctions in case of misreporting information on wages in German social security data is highly reliable in general. One limitation, however, is that around 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 Appendix A for details). Further, we improve information on education following Fitzenberger et al. (2005) and restrict the sample to full-time workers between 18 and 64. As we are only interested in effects on individuals in regular employment we exclude apprentices, interns, marginal employed workers and trainees. The final dataset consists of 3,498,536 observations from 539,179 individuals between 2000 and 2014.

We complete our dataset with additional regional information. First, we use the *de facto* standard definition of local labor-market-areas in Germany from Kosfeld and Werner (2012). The authors use factor analysis on commuter flows to identify local labor-market-areas in Germany. Their goal is to design areas with strong internal commuter links but clear detachment from other areas. They partition Germany into 141 local labor-market-areas. Second, because labor-market-areas consist of multiple counties (*Stadt- und Landkreise*, NUTS-3) we further add county-level indicators on population density, unemployment and number of hotel beds (as proxy for amenities) from the Federal Institute for Research on Building, Urban Affairs and Spatial Development to our dataset.

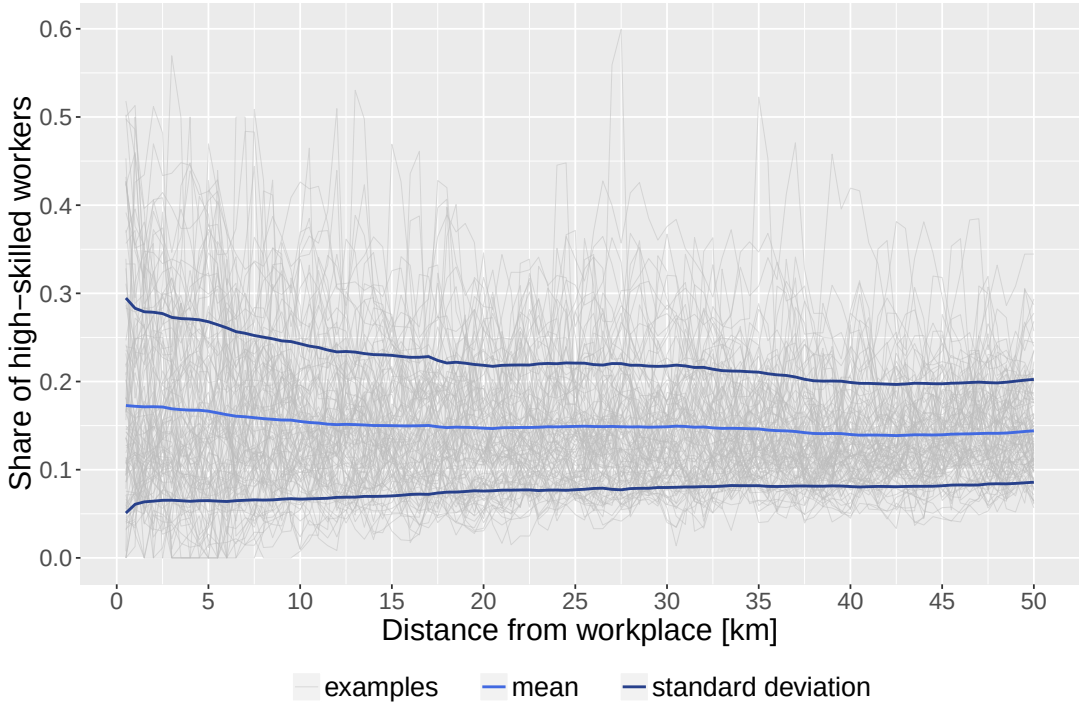
3.2 Descriptive statistics

For each workplace in our data we compute a geo-spatial function that relates the share of high-skilled workers to distance. Figure 1 illustrates the resulting curves. The light gray functions are 100 random examples and give an impression of the variability in the data. The blue curve shows the average share of high-skilled workers around establishments and the dark blue curves 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. It is apparent from the graph 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 an even mean of the intensity of human capital over the whole domain. Similarly, also the standard deviation is relatively steady in space. See Appendix B for examples of curves.

Note that the slight decline of the standard deviation is an artifact of the computation of curves. The lower value at the end of the domain is due to a larger area for which the point in the curve is computed as a mean, therefore lower variation as a result of in general more single observations (workers) used to compute one point of the curve.

To gain a first impression of the relationship between individual earnings and the spatial concentration of human capital, Figure 2 shows the correlation between log wages and the share of high-skilled workers in distance windows $[t-500, t], \forall t \in \{500, 1000, ..., 50000\}$. While the magnitude of the *ordinary* correlation has no direct interpretation, the declining trend signals

Figure 1: Geo-spatial functions of the share of high-skilled workers



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

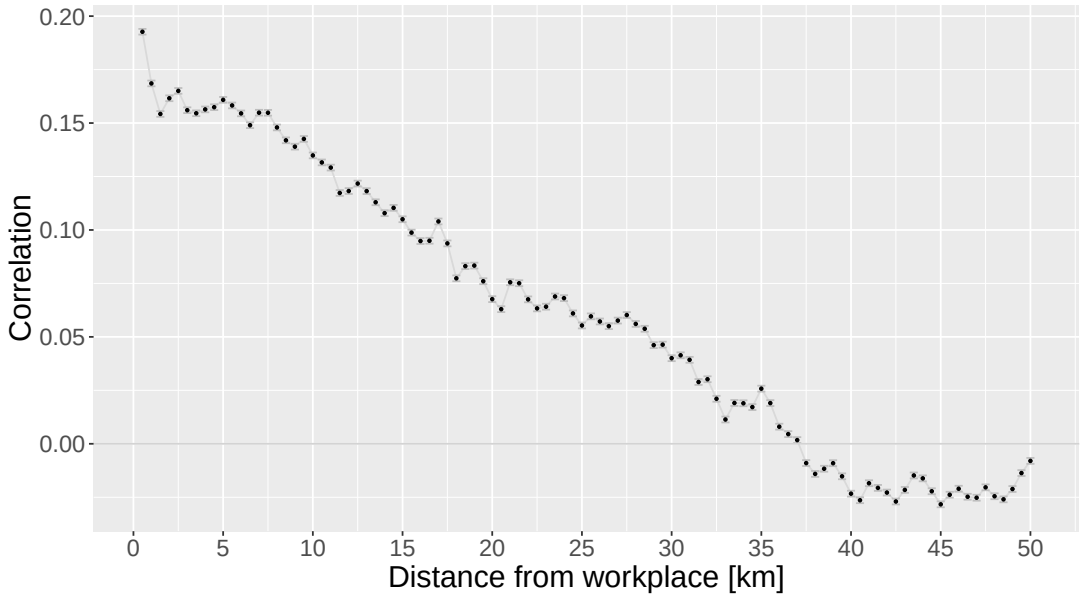
that the relationship between income and the spatial concentration of high-skilled labor decays with distance.[4]

One reason why the magnitude of the correlation coefficients has no direct interpretation is the spatial autocorrelation of values of the geo-spatial function of high-skilled workers. Figure 3 illustrates this issue. The graph shows the correlation between the share of high-skilled workers at three selected measurement points with the remaining 99 measurement points. For instance, the first panel presents the correlation of the share of high-skilled workers 0 to 0.5 kilometers away from workplaces and all other measurement points. Naturally, the correlation between measurement points attenuates with distance. Thus, close by measurement points have high correlation and distant measurement points have low correlation. While ordinary correlations ignore spatial autocorrelation, standard OLS regression is in principle able to orthogonalize covariates. However, as we will discuss in the next section, strong correlation within such a broad feature space makes OLS practically unfeasible.

Appendix C provides summary statistics on individual wages and the remaining dataset.

[4]For two reasons the magnitude of the correlation between wages and the share of high-skilled workers in some distance window has no direct interpretation. First, the bandwidth of the distance window determines the strength of the correlation. We could, for instance, shrink the correlation coefficient to arbitrarily small values by decreasing the bandwidth of the distance window. Second, the *ordinary* correlation does not partial out the relationship between wages and other distance windows than the focal one. Naturally, neighboring distance windows are (spatially auto-) correlated.

Figure 2: Correlation of individual wages and the regional share of high-skilled workers



The figure illustrates the correlation between log wages and the share of high-skilled workers in distance windows $[t-500, t], \forall t \in \{500, 1000, ..., 50000\}$. The graph suggest 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.

4 Results

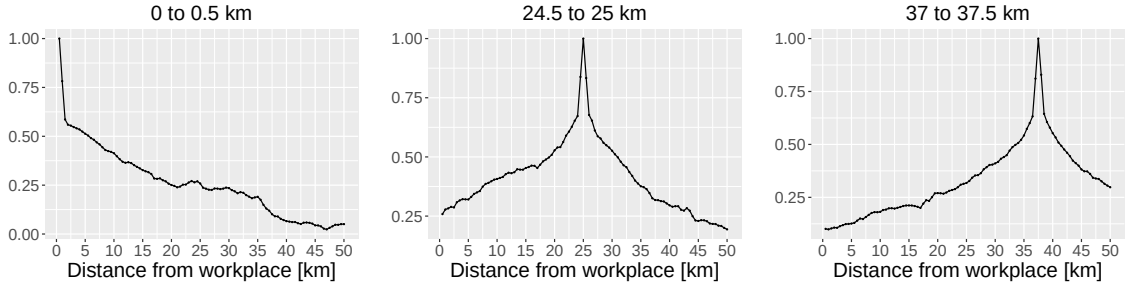
4.1 Main findings

Estimation results for the slope-valued spillover-coefficient of the final model are given in Figure 4 and Figure 5. In this model, we include fixed effects for worker-firm combinations and time-varying combinations of labor-market-area and skill levels but also occupation fixed effects (see also section 2.4). The remaining variation, thus, stems from intra-regional changes of the share of high-skilled workers. Figure 4 shows results for an unrestristricted estimate ($\rho = 0$), the result coincides with a standard OLS regression, although being scaled by the number of discretizuation points as it is an approximation of the integral by a Riemann sum. Confidence bands are computed via equation (7) where we cluster on the firm level and compute clustered standard errors according to Abadie et al. (2017).

The unpenalized estimate does not reveal a significant relationship between the share of highskilled workers at any distance and the point estimate possesses a wiggly behaviour over the whole domain. This in general indicates imprecise estimate which in our situation is a consequence of both high correlation in the curves between adjacend grid points (multicollinearity) and a weak association between the remaining variation in the curves and the outcome variable after controlling for the full set of fixed effects. As a result, the confidence band includes the null everywhere.

Things change once a penelty on the curvature of $\beta(t)$ is introduced. We choose the penalty such that the resulting curve has a flexibility comparable to a parabola (three degrees of freedom) but may stay flat over some interval. The most appropriate estimate in our view is given with

Figure 3: Spatial autocorrelation at selected measurement points



The graphs shows the spatial autocorrelation of the geo-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 selected three focal points well illustrate the general pattern of the underlying three dimensional correlation function.

degrees fo freedom taking values between 2 (linear function) and 3 (parabola). In Figure 5 we choose 2.5; the results under different choices are qualitatively equivalent but of course more flexible.

As Figure 5 illustrates, the spillover effects decay with distance. According to our estimate, the spatial reach of human capital spillovers affects wages up to a distance of 15 km. The magnitude of the total effect is in line with a classical estimate on the county level but somwhat smaller compared to results obtained by Moretti (2004). Accordingly, an evenly distributed increase of human capital by onepercentage point raised average wages by 0.2 percent. However, if the increase in human capital is closer by, the effect is obviously larger. If the increase occurs within 5km, its effect on wages will be double as high as if the increase would occur between 5km and 10 km.

If one does not include fixed effects for combinations between labour-market-area, time and skill, the estimated effect still includes conventional labor market supply/demand effects on the level of labor-market-areas and time which go in potentially different directions for the two skill levels. The resulting estimate is given in Figure 6 and the main difference compared to Figure 5 is given by a globally higher-valued function (scaling of the axis) and a measurable effect up to 30km. From a theoretical perspective, it is not clear at all, how supply/demand effects within a labor market influence average wages. Theoretically, a higher supply of high-skilled workers increases wages of low-skilled workers but decreases wages of the high skilled, (for details see Moretti, 2004). The upward-shift obtained without labour-market-area-year-skill fixed effects suggests that the total effect due to the labor market mechanism is positive which implies that the positive effect on low-skilled waged outweights the negative wage effect for high-skilled workers.

Ignoring match-specific worker-firm fixed effects (Figure 7), which also capture location specific productivity, the estimated spillover parameter is shifted upwards once again. This shift can be related to sorting of high-skilled workers into regions with location-specific advantages.

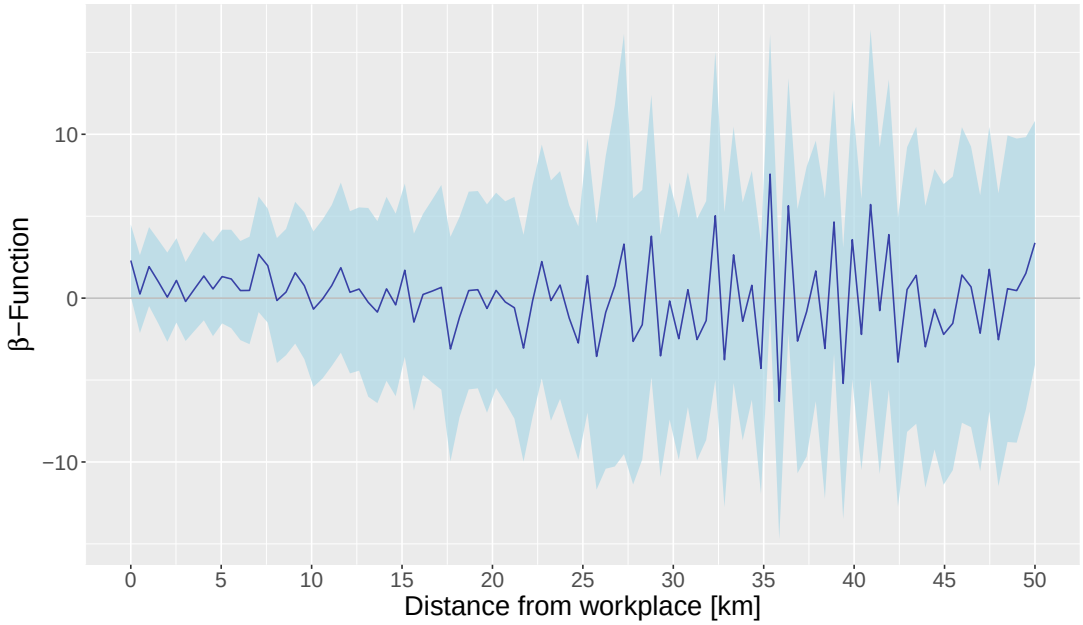


Figure 4: Unpenalized estimation result for model (9). Included fixed effects: individual–firm, labor-market-area–year–skill, year–occupation.

4.2 Simulation study

As the estimation approach chosen in this paper is not so well-established in the econometric literature, the following results of a simulation exercise are meant to evaluate the statistical properties of our estimation framework in finite samples. As we will show, the wiggly estimate of the unrestricted estimate is a direct result of a too flexible fit. Furthermore, we will show, that our inference procedure produces reliable results.

In the first set up, we evaluate the estimator's properties in a situation where the DGP is exactly the one we obtain from our final estimate (see Figure 7). To make the situation comparable to our particular case, we simulate data based on our final model (equation (9)). To make the situation comparable to our particular real application, we directly take the observed curves (measuring high-skilled density at given distance from the focal observation) and all other covariates where the respective effects on the outcome variable are taken from the estimate that belongs to Figure 7. The structure of the simulation sample (sample size, number of firms, number of workers per firm etc.), therefore, is the same as in the original sample. Simulated observations of the dependent variable are then obtained by adding iid draws from $N(0, \widehat{\sigma}_u^2)$, where $\widehat{\sigma}_u$ is the standard error of residuals obtained from the estimate of model (9). For this setup, we simulate 1000 replications to assess the estimator's statistical properties.

The results obtained from this exercise are summarized in Figure 8 which shows on top of the simulated estimates also the pointwise mean curve over all replications and the function $\beta(t)$ of the DGP used to generate the data. As it is in general the case for penalized estimation (or nonparametric), one can observe that the mean function deviates from the original one in particular at regions where the original has a more complex structure, i.e. where it possesses a stronger curvature, that is, possesses a bias. In general, however, the estimated curves well resembles the original. In particular, there is no replication that deviates substantially or has a

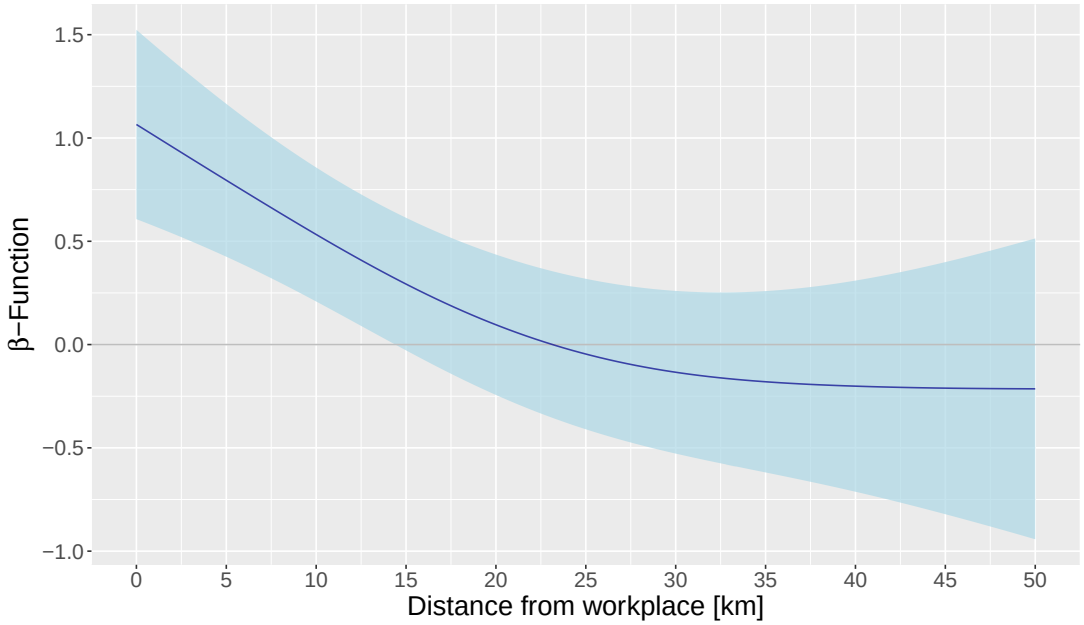


Figure 5: Penalized estimation result for model (9). Included fixed effects: individual–firm, labor-market-area–year–skill, year–occupation.

completely wrong shape. One inherent property of the estimator, therefore, is that it would never produce an estimate whith regions being statistically different from zero where the true curve is zero in a wider neighborhood. Therefore, if one believes that the function corresponding to the true spillover mechanism is monotonically decreasing and is zero beyond a certain distance, the regularized estimation is able to capture this structure well. To put this more formally, such shrinkage-type estimators aim to solve the bias-variance-tradeoff, thus, produce estimates that are optimal with respect mean prediction error which translates for linear estimators to optimal estimation with respect to mean estimation error. Consequently, our estimator would produce misleading results only if the true mechanism is much more complex (non mononotonic behaviour, which is very unlikely).

As a direct consequence of the locally biased estimate, the confidence bands fail to fully (on whole domain) cover the true curve with correct probability (i.e. at least nominal confidence level). As Figure 9 illustrates, the coverage probability is in particular low at regions, where the estimator has a large bias. This would be only a problem if this inference approach rejected a correct local null. This, indeed, turns out to be the case, but only in regions adjacent to regions where the null is not correct. The local inference procedure, thus, is less reliable in neighborhoods where the true curve has a strong curvature.

The implications for our estimation results are as follows. If the true mechanism is not too complex, the estimate is reliable and the confidence bands in general reliably indicate a significant local influence of the functional regressors on the scalar outcome variable. However, considering the exact location, where the spillover function is no longer distinguishable from zero, inference is less precise and it seems reasonable to choose a threshold somewhat smaller than indicated by our estimate. With the experience from the simulation exercise, a reasonable value for the present case is 18km. For smaller values, our estimate suggests a statistically significant influence; for a

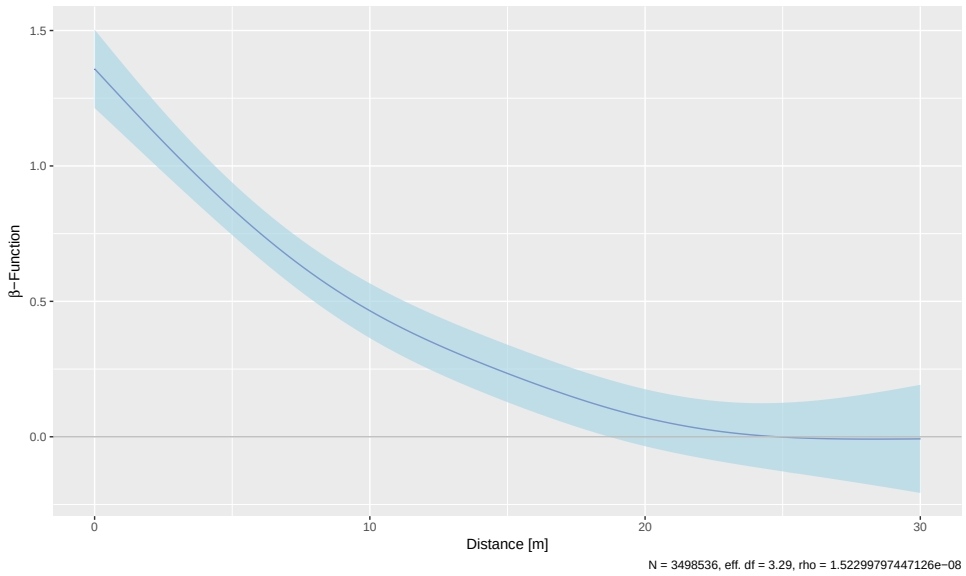


Figure 6: Penalized estimation result for model (9). Included fixed effects: individual, labor-market-area–year–skill, industry, occupation.

distance larger than 18km, a significant contribution no longer can be measured.

5 Conclusion

In this paper we estimate the spatial reach of spillovers from human capital density, an important source of long-run economic growth. By drawing on precisely geocoded register data and employing regression models for functional data, we are able to estimate the spatial extent of spillovers in a continous manner. The rich panel data allows us to control for important confounders such as spatial sorting and classical demand and supply effects.

We find economically significant spillover effects from high-skilled workers. Moreover, our estimates reveal that spillover effects linearly decay with distance and vanish after approximately 20 kilometers. In line with our expectations, spillovers from close neighbors are notably larger than spillovers from distant neighbors. An evenly distributed increase of the share of high-skilled workers of 10 percentage points (one standard deviation) raises individual wages by 2%.

References

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. 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.

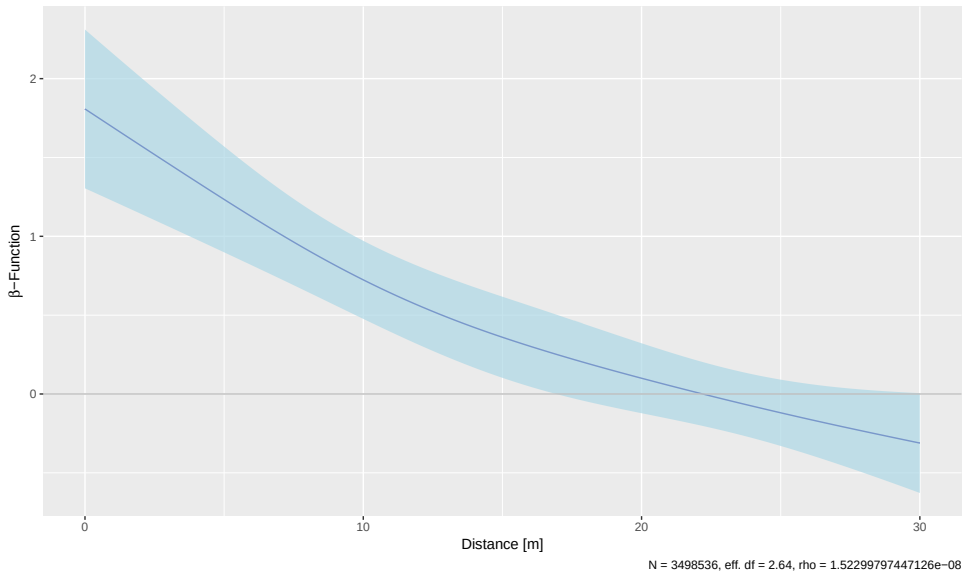


Figure 7: Penalized estimation result for model (9). Included fixed effects: individual–firm, year, occupation.

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.

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.

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.

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.

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 an application to US cities. *The Review of Economic Studies 488*(73), 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.

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

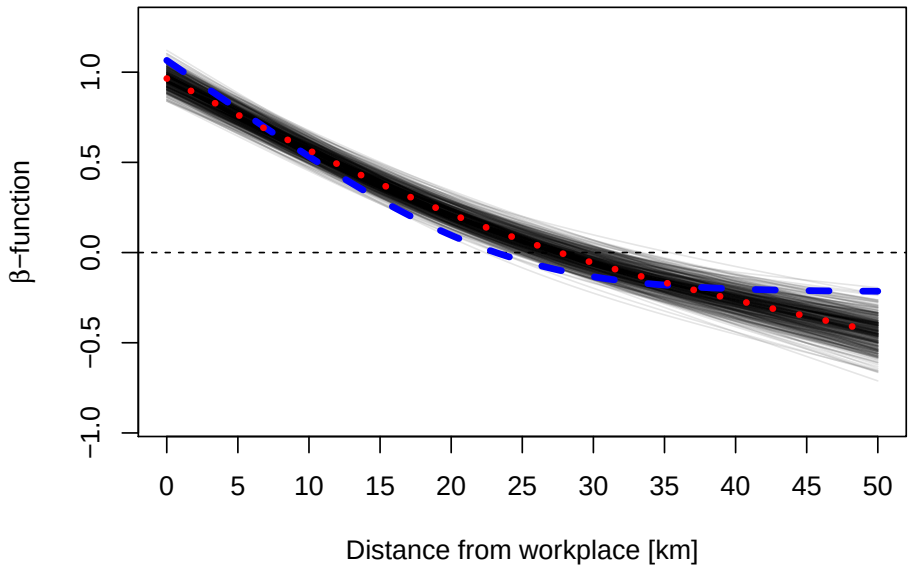


Figure 8: Simulated replications for estimates (grey lines), mean function of simulated replications (dotted, red color) and the true DGP function (dashed, blue color).

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

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

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

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

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

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.

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.

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

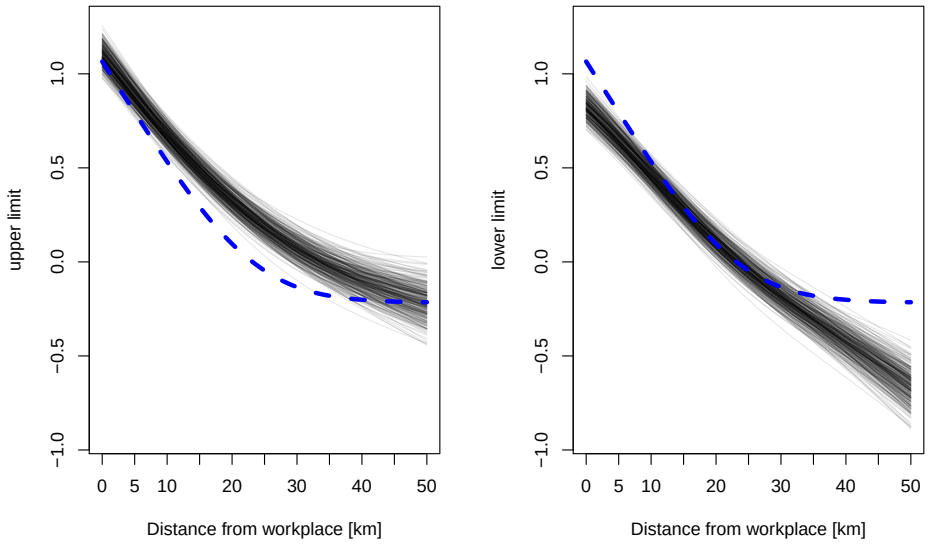


Figure 9: Simulated replications of upper (left panel) and lower (right panel) boundary of confidence intervals togehter with true DGP function (dashed, blue color).

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.

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

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

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.

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

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

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.

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

Appendix

A Imputation of wages

A common limitations 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, indicators for gender, older than 40 years and foreign. Additionally, we include interaction terms of age and age^2 with the indicator variable *older than 40*. On the county level we further include the predictors: population density, unemployment rate, number of hotel beds and share of high-skilled workers. With the parameters from the Tobit estimates ($\hat{\beta}$) we impute wages by $X\hat{\beta} + \hat{\sigma}\Phi^{-1}[k + u(1-k)]$, 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, $k = \Phi\left[(c - X\hat{\beta})/\hat{\sigma}\right]$ and c is the censoring point.

In the second step, we compute life-time 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 life-time average wages as additional predictors, we repeat the Tobit estimates. Finally, we impute censored wages by $X\hat{\beta} + \hat{\sigma}\Phi^{-1}[k + u(1-k)]$.

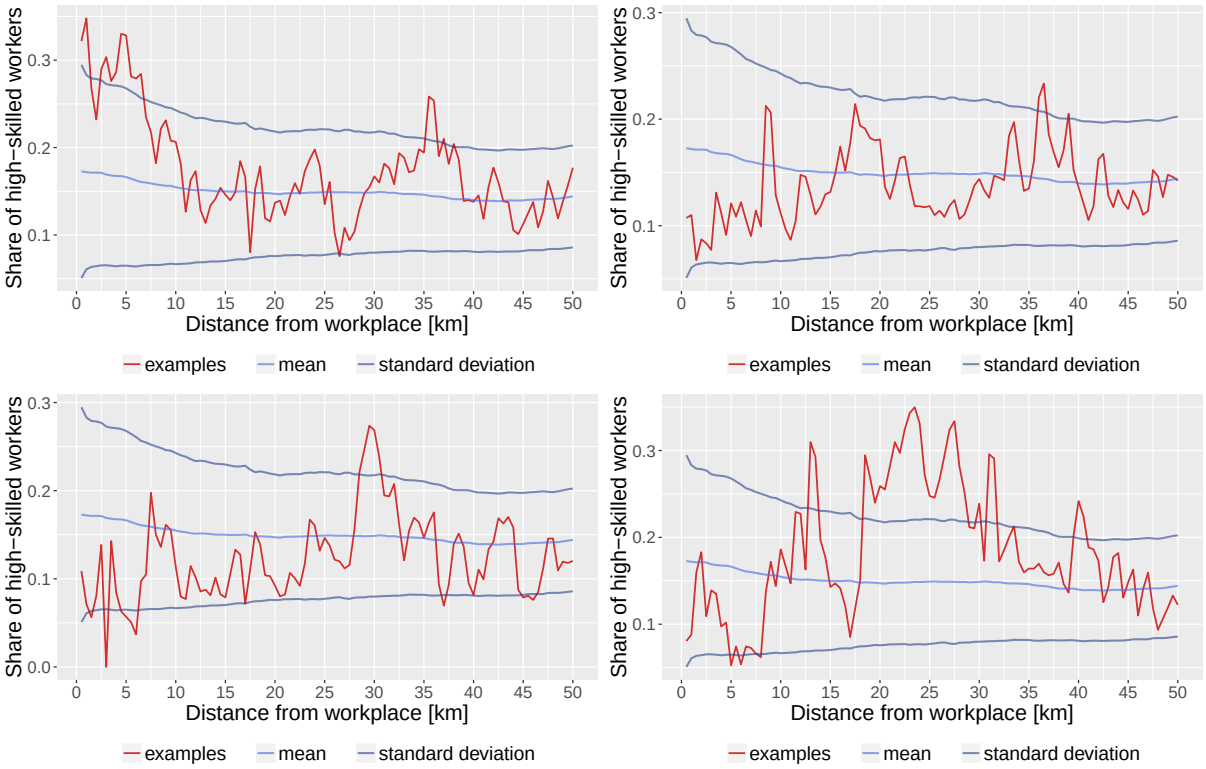
B Examples of geo-spatial functions of high-skilled workers

In the paper we describe the distribution of high-skilled workers as continuous curves. More precisely, we define geo-spatial functions that map the share of high-skilled workers to the distance from the workplace. To illustrate these functional objects Figure B.1 gives four randomly drawn examples. In each of the four graphs, red lines represent the share of high-skilled workers around an establishment. The light blue lines in the background indicate the pointwise mean and standard deviation in our dataset. For instance, in the first panel we see a high concentration of skilled labor of 30% in the close neighborhood of the workplace. Between 5 and 15 kilometers distance, the share of high-skilled workers declines to 15%. After a dip around 25 kilometers away from the workplace, the share of high-skilled workers raises again. At the end of the domain the share of high-skilled workers is around 15%.

C Summary statistics

The dataset for our econometric analysis covers 15 years and consists of 3.5 million records of 540,000 workers. Table B1 summarizes the dependent variable (log wage) and numerical control variables. In the data the mean daily wage is 111 Euros and the first and second quartile range from 68 to 129 Euros. The average individual in the dataset is 41 years old and has a work experience of 15 years. The median population density in the dataset is 119 inhabitants per square kilometer (exp(4.78)). Furthermore, 36% of the observations are from females and 7% are

Figure B.1: Examples of geo-spatial functions of the share of high-skilled workers



The figure shows the distribution of high-skilled workers around four randomly drawn workplaces (red lines). The light blue 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 geo-spatial functions that map the share of high-skilled workers to the distance from a workplace.

Table B1: 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 popoulation 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

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.

from workers with foreign nationality. The proportion of low-, medium- and high-skilled workers are 8%, 73% and 19%, respectively.