Eppelsheimer, Johann; Rust, Christoph

Working Paper

The Spatial Decay of Human Capital Externalities - A Functional Regression Approach with Precise Geo-Referenced Data

IAB-Discussion Paper, No. 21/2020

Provided in Cooperation with:
Institute for Employment Research (IAB)

Suggested Citation: Eppelsheimer, Johann; Rust, Christoph (2020) : The Spatial Decay of Human Capital Externalities - A Functional Regression Approach with Precise Geo-Referenced Data, IAB-Discussion Paper, No. 21/2020, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg

This Version is available at:
https://hdl.handle.net/10419/234279

INSTITUTE FOR EMPLOYMENT RESEARCH
The Research Institute of the Federal Employment Agency

IAB-DISCUSSION PAPER

Articles on labour market issues

21|2020 The Spatial Decay of Human Capital Externalities – A Functional Regression Approach with Precise Geo-Referenced Data

Johann Eppelsheimer and Christoph Rust

The Spatial Decay of Human Capital Externalities – A Functional Regression Approach with Precise Geo-Referenced Data

Johann Eppelsheimer (IAB),
Christoph Rust (University of Regensburg)

Mit der Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

The “IAB-Discussion Paper” is published by the research institute of the German Federal Employment Agency in order to intensify the dialogue with the scientific community. The prompt publi-cation of the latest research results via the internet intends to stimulate criticism and to ensure research quality at an early stage before printing.

Contents

List of Figures

List of Tables

Abstract

This paper analyzes human capital externalities from high-skilled workers by applying functional regression to precise geocoded register data. Functional regression enables us to describe the concentration of high-skilled workers around workplaces as continuous curves and to efficiently estimate a spillover function that depends on distance. Furthermore, our rich panel data allow us to address the sorting of workers and to disentangle human capital externalities from supply effects by using an extensive set of time-varying fixed effects. Our estimates reveal that human capital externalities attenuate with distance and disappear after 15 kilometers. Externalities from the immediate neighborhood are twice as large as those from surroundings ten kilometers away.

Zusammenfassung

Wir analysieren Humankapitalexternalitäten von Hochqualifizierten mit präzisen georeferenzierten Sozialversicherungsdaten. Functional Regression ermöglicht es uns die Konzentration von Hochqualifizierten um Arbeitsplätze herum als kontinuierliche Kurven zu beschreiben und eine von der Entfernung abhängige Spillover-Funktion zu schätzen. Unsere umfangreichen Paneldaten ermöglichen es uns außerdem räumliche Selektion von Beschäftigten zu berücksichtigen und Humankapitalexternalitäten von Angebotseffekten mittels hochdimensionaler Fixed-Effects zu trennen. Unsere Schätzungen zeigen, dass Humankapitalexternalitäten mit der Distanz abnehmen und etwa 15 Kilometer weit reichen. Humankapitalexternalitäten aus der unmittelbaren Nachbarschaft sind doppelt so hoch wie solche aus zehn Kilometern Entfernung.

JEL

C13, D62, J24, J31, R10, R12

Keywords

human capital externalities, functional regression, geodata, wages

Acknowledgments

The authors thank Martin Abraham, Annette Bergemann, Gerard van den Berg, Linda Borrs, Frank Cörvers, Wolfgang Dauth, Matthias Dorner, Gilles Duranton, Peter Haller, Moritz Kuhn, Christian Merkl, Florian Lehmer, Joachim Möller, Veronika Püschel, Uta Schönberg, Tobias Seidel, Heiko Stüber, Rolf Tschernig, Erwin Winkler, Anthony Yezer, two anonymous referees and participants in the Regional Disparity Workshop 2019, 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 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/Dingel, 2019; Acemoglu, 1996; Lucas, 1988; Marshall, 1890). Although a large body of empirical literature supports the existence of geographically bounded human capital externalities (Cornelissen/Dustmann/Schönberg, 2017; Ciccone/Peri, 2006; Moretti, 2004; Rauch, 1993) little is known about the exact spatial extent of human capital externalities. For several reasons, human capital externalities likely 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 likely benefit more from proximate than from distant neighbors.

Previous empirical research provides initial evidence for spatially decreasing human capital externalities. Using cross-sectional data from the US, Rosenthal/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/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/Murphy, 1992; Card/Lemieux, 2001; Borjas, 2003; Moretti, 2004; Ciccone/Peri, 2006).

In this paper, we attempt to address all of these issues and estimate human capital externalities based on high-resolution geodata of an entire economy. Specifically, we estimate the ex-

ternal effect from the local concentration of high-skilled workers on individual wages. External effects may arise from knowledge exchange (Marshall, 1890; Lucas, 1988) or the diffusion of new technologies (Nelson/Phelps, 1966; Acemoglu, 1998). Both channels might increase worker productivity and thus raise their wages. To estimate human capital externalities, we draw on a large and novel administrative 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 propose to use a novel estimation procedure that is capable of evaluating such detailed geodata. This allows us to estimate the spatial attenuation of human capital externalities with high precision.

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/Finch, 2013 for a systematic overview), FDA is applied quite rarely in economics (examples include Ramsay/Ramsey, 2002, Wang/Jank/Shmueli, 2008 and Caldeira/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/Kneip/Sarda (2009), that is based on smoothing splines and makes it possible to very flexibly model the function-valued spillover parameter. 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/Strange, 2008; Fu, 2007; Verstraten, 2018; Gibbons/Overman/Sarvimäki, 2017; Faggio/Schluter/vom Berge, 2019; Faggio, 2019).[2] In the semi-

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

[2] Some examples of studies that investigate the spatial patterns of agglomeration effects are: Arzaghi/Henderson (2008), who study networking effects within the advertising agency industry in Manhattan; Ahlfeldt et

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 or 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.

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 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/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/Möller (2019) and exploit the different spatial natures of the two effects. While supply and demand effects are plausibly common within local labor markets (i.e., supply and demand effects originating in one part of the city 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 are able 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/Dustmann/Schönberg (2017), who, in a related context, address worker sorting at the firm level (Abowd/Kramarz/Margolis, 1999; Card/Heining/Kline, 2013), we address sorting of high-skilled workers into high-wage regions (Acemoglu/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.

al. (2015), who examine productivity externalities in Berlin; Andersson/Larsson/Wernberg (2019), who evaluate productivity effects from industry specialization and diversity in Swedish cities; and Faggio (2019) and Faggio/Schluter/vom Berge (2019), who assess the local labor market impact of relocations of public sector jobs in the UK and Germany.

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.

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 kilometers, 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 percent. The magnitude of this effect is comparable to *classical* estimates at the aggregate level. 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 farther 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 summarizes the data. Section 4 presents our main findings, illustrates the statistical properties of the estimator in a simulation study and provides an overview of several robustness checks. Section 5 concludes the paper.

2 Estimation strategy

This paper seeks to measure the spatial attenuation and reach of human capital externalities. Therefore, our aim is 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 the identification strategy that addresses endogenous sorting of workers and confounding labor market supply and demand effects.

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. Moreover, one can measure the concentration of high-skilled workers at any distance to worker i's workplace. It is thus natural to regard the concentration of high-skilled workers with respect to the distance to worker i's workplace as a curve. We use curves 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, the model is defined as

$$Y_i = \int_0^1 \beta(z) X_i(z)\, \mathrm{d}z + \varepsilon_i, \tag{2.1}$$

where $X_i \in L^2([a,b])$ are independent and identically distributed (iid) random functions defined on a common domain, which we set to $[0,1]$ without loss of generality. The function-valued coefficient parameter $\beta \in L^2([0,1])$ describes the influence of X_i on Y_i and varies over distance z. 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 (2.1) 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. Therefore, the expansion of the so-called functional principal component (FPC) estimator depends heavily on the random curves' correlation structure. In this paper, we instead build on the smoothing spline estimator proposed by Crambes/Kneip/Sarda (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/Horowitz, 2007; Crambes/Kneip/Sarda, 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/Kneip/Sarda (2009) minimizes the penalized sum of squared residuals

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{1}{p}\sum_{j=1}^{p}\beta(z_j)X_i(z_j)\right)^2 + \rho\left(\frac{1}{p}\sum_{j=1}^{p}\pi_\beta^2(z_j) + \int_0^1 (\beta^{(m)}(z))^2\,\mathrm{d}z\right). \tag{2.2}$$

Here, $\pi_\beta(z)$ is the best approximation of $\beta(z)$ by a polynomial of degree $m-1$ and ensures uniqueness without imposing further assumptions on the random functions X_i. The penalty parameter $\rho \geq 0$ controls the flexibility of the estimated parameter function $\widehat{\beta}$. With $\rho = 0$, for instance, equation (2.2) coincides with the least-squares criterion. The minimizer of equation (2.2) is

$$\left(\widehat{\beta}(z_1), \ldots, \widehat{\beta}(z_p)\right) = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}'\mathbf{X} + \rho\mathbf{A}\right)^{-1}\mathbf{X}'\mathbf{Y}, \tag{2.3}$$

where $\mathbf{A} = \mathbf{P} + p\mathbf{A}^*$ is a penalty matrix introduced by Crambes/Kneip/Sarda (2009). This matrix 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 $\mathbf{W} = (z_j^q)_{j,q} \in \mathbb{R}^{p\times m}$, $q = 0, \ldots, m-1$. Traditional smoothing splines penalize second derivatives. Thus, we set $m = 2$, which results in an expansion of cubic natural splines with knots at $z_1, \ldots, z_p$. The regularization matrix $\mathbf{A}^*$ is defined as usual by

$$\mathbf{A}^* = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\left(\int_0^1 \mathbf{b}^{(2)}(z)\mathbf{b}^{(2)}(z)'\,\mathrm{d}z\right)(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B},$$

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 for each of the p basis functions.

To account for the influence of further explanatory variables, we expand model (2.1) with a k-vector of scalar-valued explanatory variables Z_i and a corresponding parameter vector γ:

$$Y_i = \int_0^1 \beta(z)X_i(z)\,\mathrm{d}z + Z_i'\gamma + \varepsilon_i. \tag{2.4}$$

Accordingly, we augment the smoothing spline estimator of Crambes/Kneip/Sarda (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. The compound estimator of (discretized) β and γ then is:

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}(z_1), \ldots, \widehat{\beta}(z_p), \widehat{\gamma}_1, \ldots, \widehat{\gamma}_k\right) = \frac{1}{n}\left(\frac{1}{np}\mathbf{X}_\mathbf{Z}'\mathbf{X}_\mathbf{Z} + \rho\mathbf{A}_\mathbf{Z}\right)^{-1}\mathbf{X}_\mathbf{Z}'\mathbf{Y}. \tag{2.5}$$

Because the scalar-valued explanatory variables do not load into the roughness penalty, we extend the penalty matrix $\mathbf{A}$ by appending k zero columns and k zero rows:

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

The estimator (2.5) depends on the smoothing parameter ρ that controls the complexity of the estimate of the function-valued slope parameter β. The smoothing 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):

$$\mathrm{edf}(\rho) = \mathrm{trace}\left(\mathbf{H}_\mathbf{Z}^\rho\right), \tag{2.6}$$

where $\mathbf{H}_\mathbf{Z}^\rho = (np)^{-1}\mathbf{X_Z}\left((np)^{-1}\mathbf{X}_\mathbf{Z}'\mathbf{X_Z} + \rho\mathbf{A_Z}\right)^{-1}\mathbf{X}_\mathbf{Z}'$ is the *hat* matrix of model (2.4). Given a predefined number of degrees of freedom, equation (2.6) allows us to determine ρ. In our preferred specification, we set $\mathrm{edf}(\rho) = 2.5$; the resulting estimate can thus be substantially more complex than a straight line. We experiment with different penalties in appendix A.4. Qualitatively, our results do not depend on the exact choice of the penalty term ρ.

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/Mas/Sarda, 2007: for details).

To quantify the estimation uncertainty, we 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/Silverman, 2005: equation 15.16)

$$\mathrm{Var}\big(\widehat{\boldsymbol{\beta}}|\mathbf{X_Z}, \rho\big) = \frac{1}{n^2}\left(\frac{1}{np}\mathbf{X_Z'X_Z} + \rho\mathbf{A_Z}\right)^{-1}\mathbf{X_Z'\Omega X_Z}\left(\frac{1}{np}\mathbf{X_Z'X_Z} + \rho\mathbf{A_Z}\right)^{-1}. \tag{2.7}$$

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 'meat', $\mathbf{X_Z'\Omega X_Z}$, based on clustered standard errors at the firm level (see, for instance, Abadie et al., 2017: equation 2.3).

We use the variance (2.7) 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 $\mathrm{Var}\big(\widehat{\boldsymbol{\beta}}|\mathbf{X_Z}, \rho\big)$ 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 4.2) 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.

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. To calculate these curves from geocoded data, we compute the values of the functions $X_i(z)$ for each worker i on an equidistant grid $z_1, \ldots, z_p$:

$$X_i(z_j) = \frac{n^{hs}_{[z_j\,-\,h;\,z_j)}}{n_{[z_j\,-\,h;\,z_j)}}. \tag{2.8}$$

Here, $n^{hs}_{[z_j\,-\,h;\,z_j)}$ refers to the number of high-skilled individuals for which the spheric 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, $n_{[z_j\,-\,h;\,z_j)}$ is the number of all workers (high-skilled and low-skilled) within the 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 $z_j = 500m, 1000m, \ldots, 50000m$.

There are several options for the actual measure of the concentration of high-skilled workers. We decide to measure the density of high-skilled workers by their share in all workers instead of, for instance, by their absolute numbers or high-skilled workers per square meter for several reasons. First, just as the geographic area covered by $[z_j - h, z_j)$ increases with distance z_j, the absolute number of high-skilled workers that could potentially populate that area also increases with distance. Thus, when using absolute numbers, the intensity of high-skilled workers would increase with distance almost by definition and would therefore not provide comparable values of $X_i(z)$ across space. Second, as the data show, the proportion of inhabited land decreases with z. As knowledge transfers appear only in inhabited areas, using high-skilled workers per square meter would therefore decrease the intensity of human capital with distance by construction. Thus, high-skilled workers per square meter would also not suffice to compare the concentration of high-skilled workers at varying distances. By contrast, the number of workers within the distance window $[z_j - h, z_j)$ is a reasonable unit of measurement of the *de facto* populated area, which, thinking of skyscrapers, not only covers actual land use but also the 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). Using shares is also in line with the recent literature on regional human capital externalities following Moretti (2004).

2.4 Identification

Having explained the estimator, we will now address confounding labor market demand and supply effects and the endogenous sorting of individuals. The empirical literature has established that high- and low-skilled labor are imperfect substitutes (e.g., Autor/Katz/Kearney, 2008; Ciccone/Peri, 2005; Card/Lemieux, 2001; Krusell et al., 2000). As Acemoglu/Angrist (1999), Moretti (2004) and Ciccone/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/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 plausibly uniformly affect the local labor market. Thus, independent of the exact location, a shift in the supply of high-skilled labor homogeneously affects wages within a local labor market. We are thus able to nullify labor market supply and demand effects by eliminating all variation in the data that is common within local labor markets without removing intra-regional variation from human capital externalities.

As labor market supply and demand shifts vary over time and the direction of such shifts idiosyncratically affects high- and low-skilled individuals, we expand equation (2.4) 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:

$$Y_{it} = \int_0^1 \beta(z) X_{it}(z)\,\mathrm{d}z + Z'_{it}\gamma + \theta_{if} + \tau_t + \omega_o + \pi_{rst} + u_{it}. \tag{2.9}$$

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 firm i 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/Angrist, 2000) constitutes another challenge in identifying regional human capital externalities. In our application, sorting threatens identification on two levels: first on the level of treated individuals (i.e., individuals whose wages we observe) and second on the treatment level itself (i.e., the spatial density of high-skilled workers). Regarding treated individuals, the most able workers might sort into high-skilled neighborhoods. Sorting would thus create a spurious relationship between wages and the local concentration of human capital. Regarding the treatment level, high-wage areas might attract high-skilled workers. Sorting would thus lead to reverse causality. Inspired by Cornelissen/Dustmann/Schönberg (2017), we address sorting with an extensive set of fixed effects.

Although the empirical literature finds that workers do not sort into cities based on their (unobserved) abilities (De la Roca/Puga, 2017; Glaeser/Mare, 2001), there is evidence of ability-driven sorting of workers into firms (Card/Heining/Kline, 2013; Abowd/Kramarz/Margolis, 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 the estimates, we include worker-firm match fixed effects (θ_{if}) in our model. Worker-firm match fixed effects eliminate the unobservable characteristics of workers and firms that are time-constant during the matched employment period (i.e., from the beginning to the end of the focal employment relationship). Thus, worker-firm match fixed effects prevent the data from reflecting worker ability or firm productivity.

Regarding sorting at the treatment level, high-wage regions might attract high-skilled workers, which would reverse the direction of causality in equation (2.9). Let us discuss the issue of reversed causality on two levels: the local labor market and the closer neighborhood of firms. Moretti (2004) raises the concern that local labor market conditions might affect the regional concentration of high-skilled workers. For instance, booming cities with growing wages might attract high-skilled workers. Our identification strategy overcomes such issues by removing all time-constant and time-varying variation at the local labor market level (π_{rst}). Thus, reversed causality in the local labor market area is impossible in our estimation framework. In equation (2.9), identification of human capital externalities comes from temporal variation within local labor markets. Thus, one is tempted to think that reversed causality might also threaten 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 at neighboring firms and thus not reverse the direction of causality in our framework.

As explained above, we include worker-firm match fixed effects in our estimates. An additional benefit of worker-firm match fixed effects is that they also remove neighborhood characteristics from the data that are time-constant during the matched employment period. These characteristics include locational advantages that might influence productivity, like proximity to infrastructure or market access. Our estimates of human capital externalities are thus also not biased by neighborhood characteristics that are relatively stable over time. The average length of worker and firm matches in our data is 8 years. Consequently, only small area shocks that simultaneously affect wages and the concentration of human capital in the neighborhood that have considerably short-lived effects might remain in the data. Although we believe short-lived effects that contemporaneously affect individual wages and the concentration of high-skilled workers are rare, we cannot fully exclude that our estimates might be influenced by such shocks.[3]

In summary, equation (2.9) allows us to estimate human capital externalities that are unre-

[3] An example of a highly localized shock that might influence individual wages and the concentration of human capital could be the opening of a new subway station. A new subway station might increase the market potential of shops close to the subway station and thus might raise the wages of their employees. At the same time, a subway station might increase the attractiveness of the neighborhood, and thus more high-skilled workers would be inclined to work in that neighborhood.

lated to labor market demand and supply effects and the 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 (2.9) stems from temporal intra-regional changes in the concentration of high-skilled workers.

3 Data and descriptive statistics

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 workers. 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).[4] The *Establishment History Panel* comprises all German establishments with at least one employee on June 30 of each year. The dataset provides establishment-level 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.[5]

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 2.3.

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).[6] The Sample of Integrated Labour Market Biographies is a

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

[5] 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.

[6] For a detailed description of the Sample of Integrated Labour Market Biographies, see Antoni/Ganzer/vom Berge (2016)

2 percent random sample of social security records. The dataset contains information on wages, age, work experience and education, among other data, with daily precision. To join the individual-level data to the establishment-level data, we transform the spell dataset into a yearly panel with June 30 as the reference date and link workers and firms with 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 percent of earnings are right-censored at the social security maximum. Therefore, we impute top-coded wages following Dustmann/Ludsteck/Schönberg (2009) and Card/Heining/Kline (2013) (see appendix A.1 for details). Further, we improve information on education following Fitzenberger/Osikominu/Völter (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/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/Petrongolo (2017), implying that 80 percent 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/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).

3.2 Descriptive statistics

Figure 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

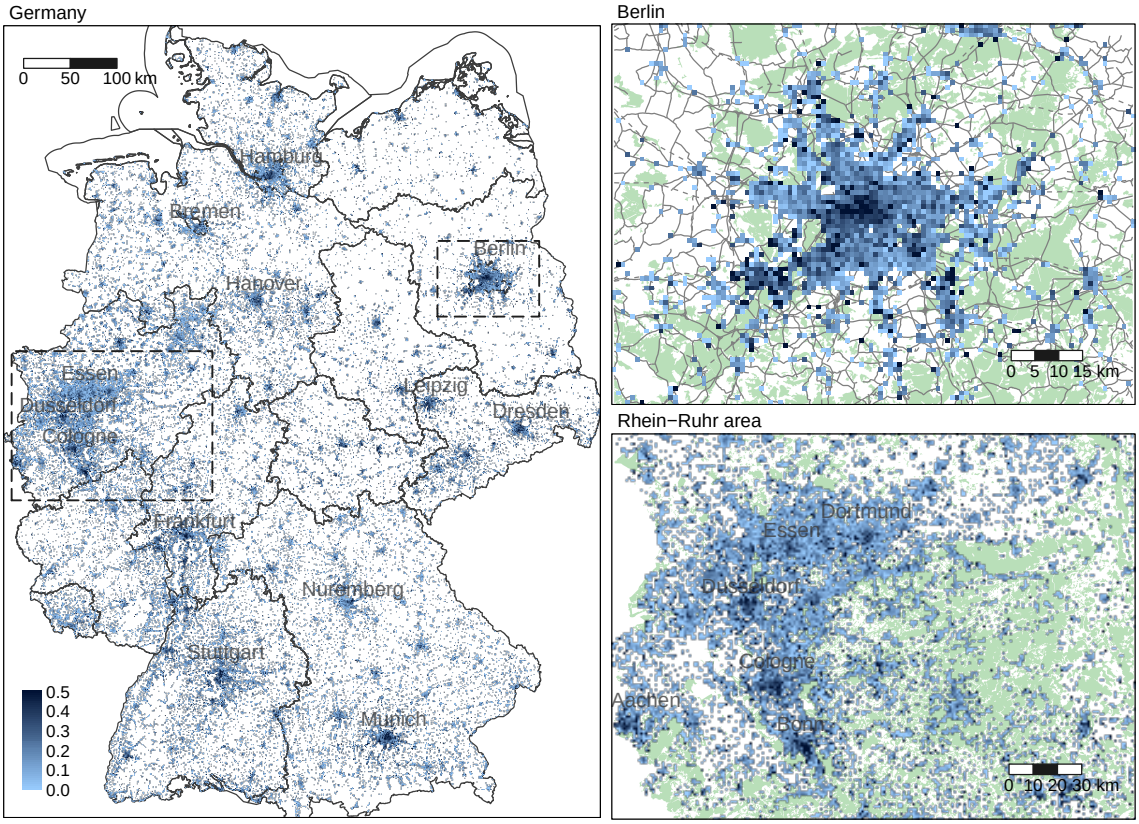
coordinates. The map illustrates the considerable diversity in the distribution of high-skilled workers in Germany. For instance, among the largest cities, there is a massive concentration of high-skilled workers in Munich, Hamburg and Berlin. By contrast, Nuremberg and Bremen exhibit significantly lower 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 percent 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 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-left 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 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 percent in the direct neighborhood of establishments and gradually declines to 14.5 percent 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 variance of the average decreases. Refer to appendix A.2 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 shows the correlation between log wages and the share of high-skilled workers within distance windows $[z_j - 500m, z_j), z_j = 500m, 1000m, ..., 50000m$. 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.[7]

[7] The magnitude of the correlation between wages and the share of high-skilled workers in some distance window has no direct interpretation for two reasons. 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

Figure 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 percent. 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.**

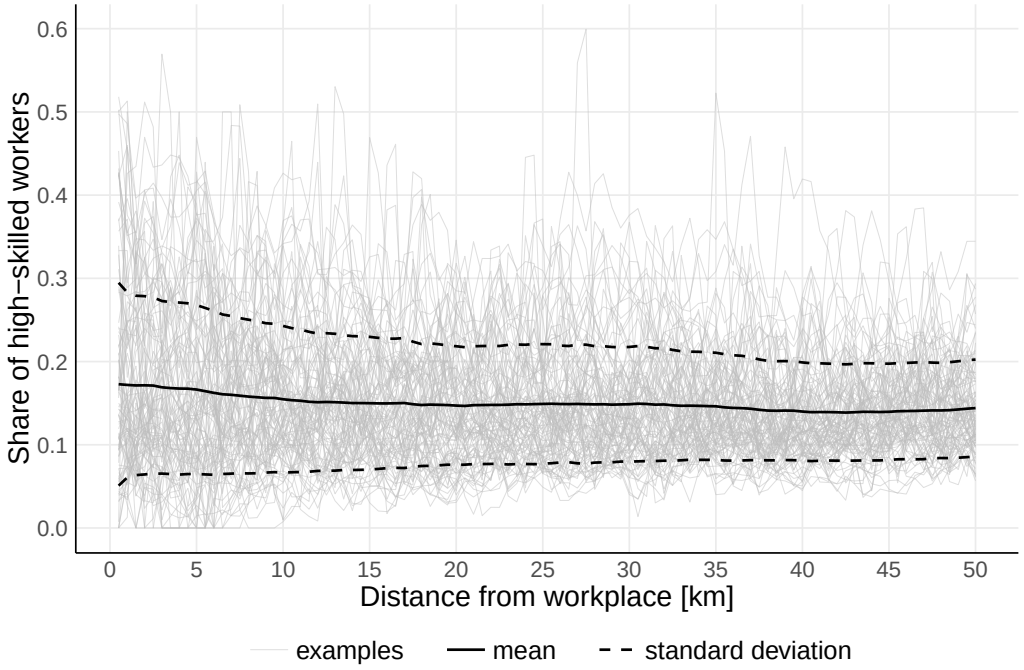
Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

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 4 illustrates this issue. The graph depicts the 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.

While ordinary correlations (figure 3) ignore spatial autocorrelation, standard OLS regression is in principle able to orthogonalize covariates. However, as discussed in the next sec-

out the relationship between wages and other distance windows than the focal one. Naturally, neighboring distance windows are (spatially auto-) correlated.

Figure 2: Spatial functions of the share of high-skilled workers

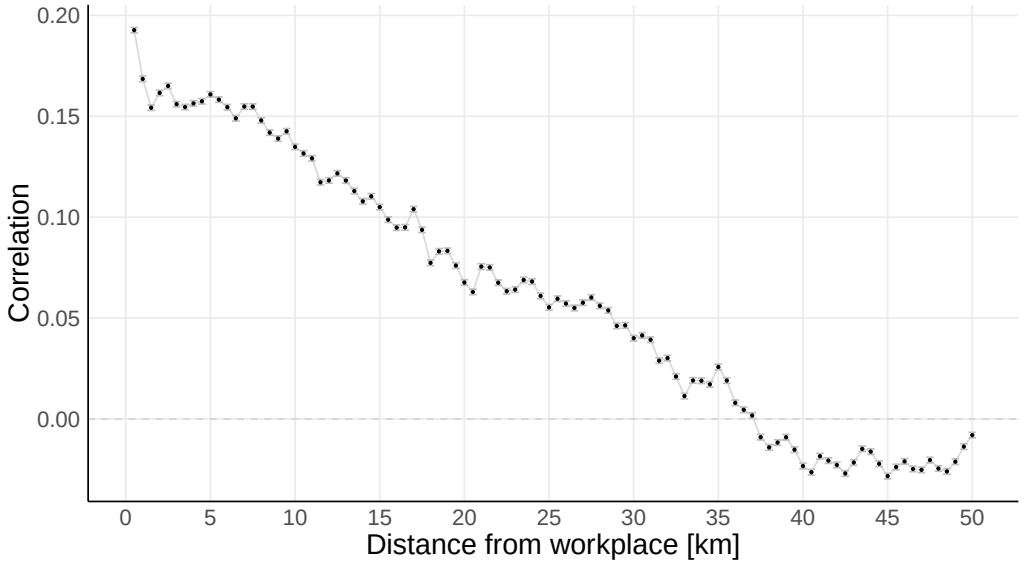


Notes: **The figure shows the pointwise mean (solid line) and standard deviation (dashed lines) of the share of high-skilled workers around workplaces. Throughout the paper, we describe the share of high-skilled workers with spatial functions that map the share of high-skilled workers to the distance 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.**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

tion, given the strong correlation between adjacent measurements, an unpenalized OLS regression does not reveal any relationship at all. For further summary statistics on individual wages and other covariates in our dataset, we refer to appendix A.3.

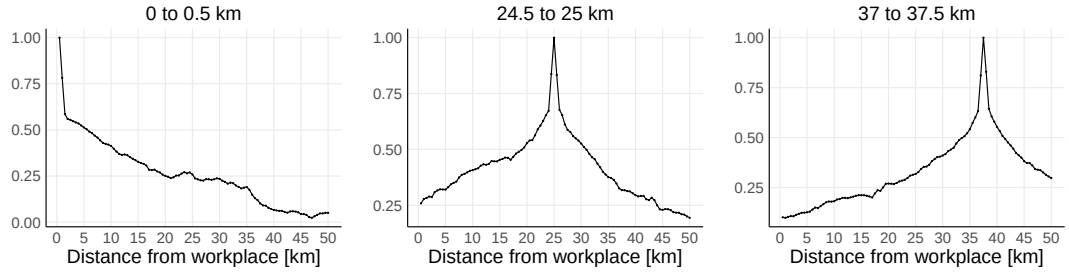
Figure 3: Correlation of individual wages and the regional share of high-skilled workers



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.**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

Figure 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.**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

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. Next, we corroborate the robustness of our estimates with a simulation study and a placebo test. Finally, we summarize several additional robustness checks.

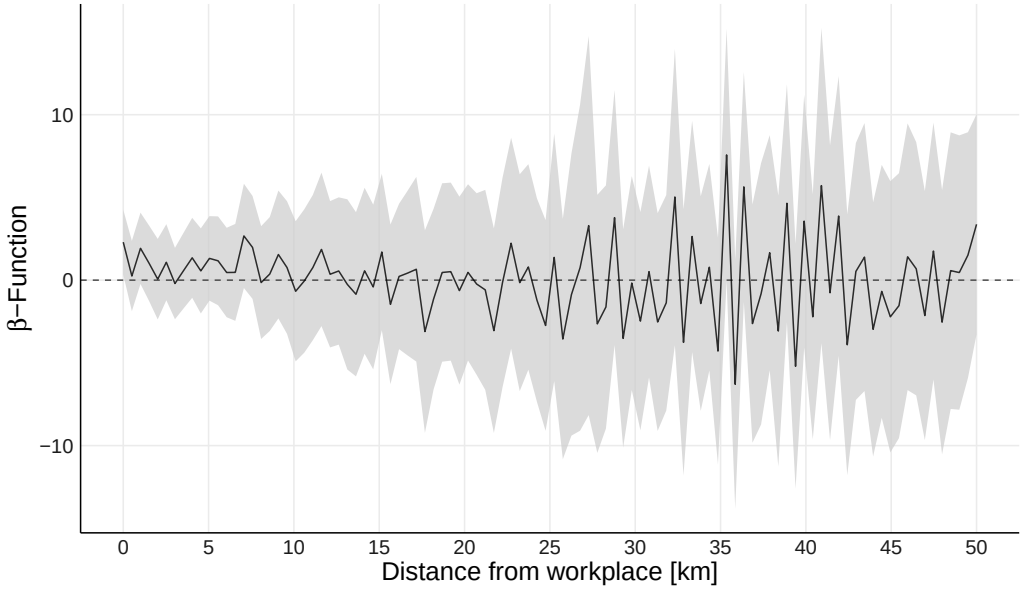
4.1 Main findings

We illustrate estimates of the spatial intensity of human capital externalities from high-skilled workers in figures 5 and 6. Figure 5 depicts an unrestricted estimate of equation (2.9) (i.e., $\rho = 0$ in equation (2.5)), which coincides with standard OLS regression. Figure 6 presents penalized estimates of equation 2.9 (i.e., $\rho > 0$). Both estimates control for labor market demand and supply effects and endogenous sorting of individuals with an extensive set of fixed effects. In addition to standard controls form 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 percent confidence band. Note that OLS estimates of equation (2.9) 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 5 shows, the unpenalized estimate of equation (2.9) 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 6 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

Figure 5: Unrestricted estimates of spatial human capital externalities from high-skilled workers



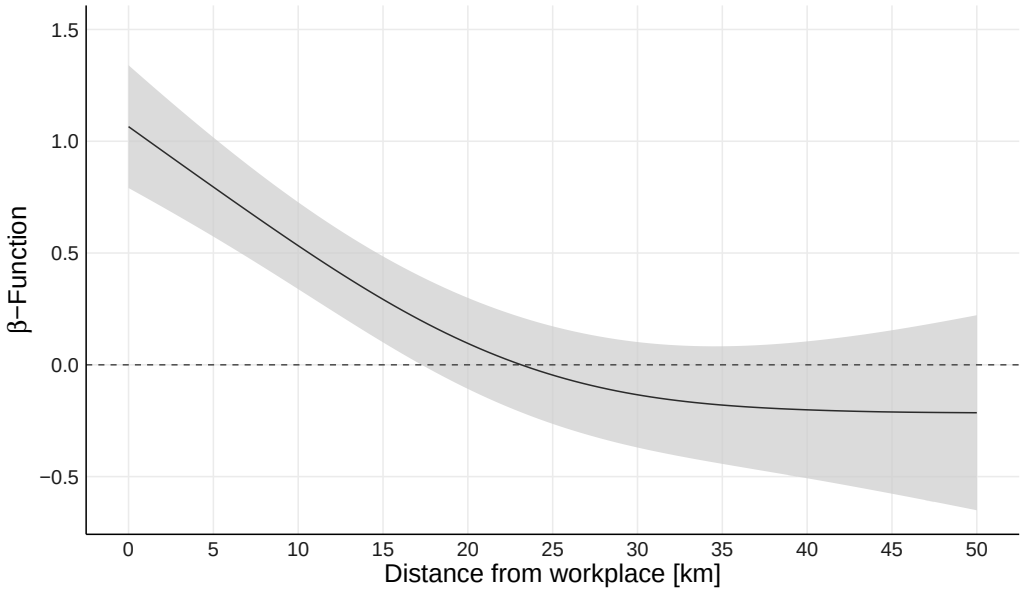
Notes: **The figure presents an unrestricted estimation of spatial human capital externalities from high-skilled workers into individual log wages (equation (2.9)). We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. The black line illustrates the estimated spillover function ($\beta(z)$), and the gray area indicates the 99 percent confidence band. The unrestricted estimator (equation (2.5), 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).** $N = 3,498,536$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

estimate can be substantially 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 A.4).

Our estimates in figure 6 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 percent ($\approx 20 \times \{0.75 \times \frac{5}{50} + \frac{1}{2}\left[(1 - 0.75) \times \frac{5}{50}\right]\}$). 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 percent ($\approx 10 \times \frac{1}{2}\left(1 \times \frac{20}{50}\right)$). Reassuringly, *classical* estimates at an aggregate

Figure 6: 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 within distance z. To compute the spatial spillover function ($\beta(z)$) we estimate equation (2.9) with the estimator (2.5). 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 gray area indicates the 99 percent 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 percent. 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 A.4 in the appendix for a complete list of parameter estimates.** $N = 3,498,536$

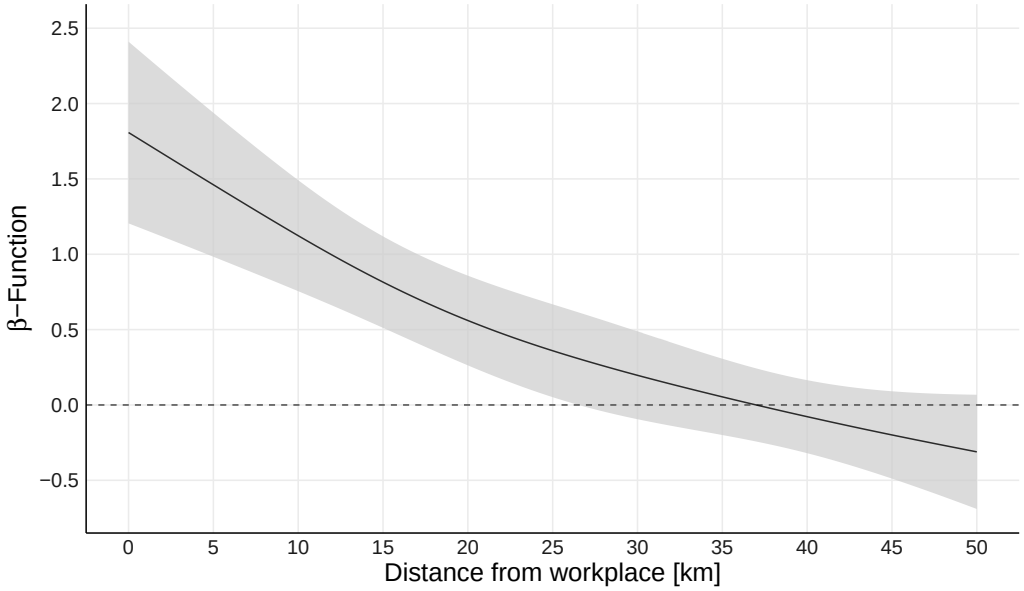
Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

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 (2.9), suggest effects of the same magnitude (see appendix A.5).

Our results are also similar to the findings of Rosenthal/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/Strange (2008).

Let us now briefly discuss the importance of removing demand and supply effects when estimating human capital externalities. Figure 7 reports estimates of our model (equation (2.9)) 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/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/Peri (2006). They also find large bias from the demand and supply effects in Mincerian estimates of human capital externalities.

Figure 7: 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 (2.9) 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 (2.5). 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 percent 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).** $N = 3,498,536$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

4.2 Simulation study

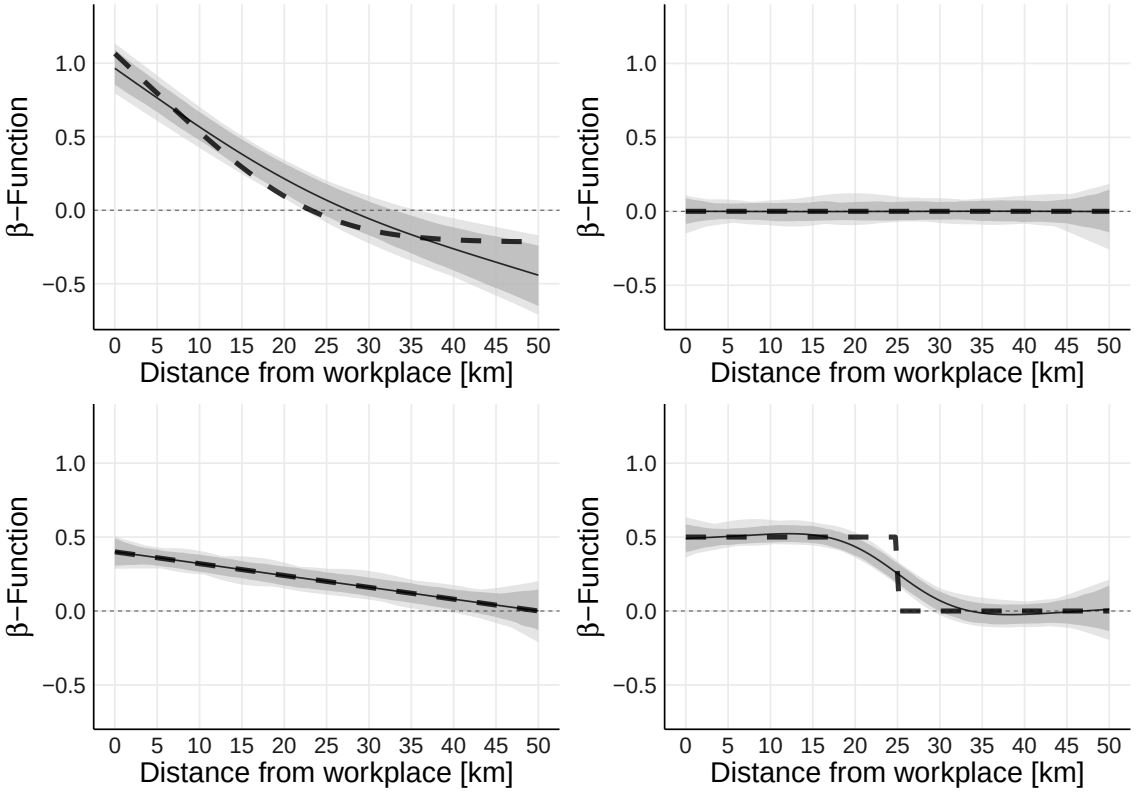
As outlined in section 2.2, drawing local inference about the function-valued parameter β is difficult. The following simulation exercise, therefore, is intended to evaluate the statistical properties of our estimation framework. The results show that our estimation framework, although yielding locally 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 for 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 6). 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.[8] 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 8 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$^{\text{th}}$ 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 panel of figure 8. 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.

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

Figure 8: 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 (2.9) 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 $\beta(z) = 0.5 \cdot \mathbb{1}(z < 0.5)$.**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

Table 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 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 (2.7) have proper coverage probability that, 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).

Table 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** $\overline{\beta}(z) = 1000^{-1} \sum_{r=1}^{1000} \widehat{\beta}_r(z)$**.**
Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

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 nonsignificant after 15 kilometers.

4.3 Placebo test: future concentration of high-skilled workers

Following Cornelissen/Dustmann/Schönberg (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 9 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.

4.4 Further robustness checks

Appendix A.6 provides details on further robustness checks. In this section, we briefly summarize the results of these exercises.

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/Strange, 2008; Fu, 2007; Verstraten, 2018; Gibbons/Overman/Sarvimäki, 2017; Faggio/Schluter/vom Berge, 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 researches 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. See appendix A.6.1 for details.

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. See appendix A.6.2 for details.

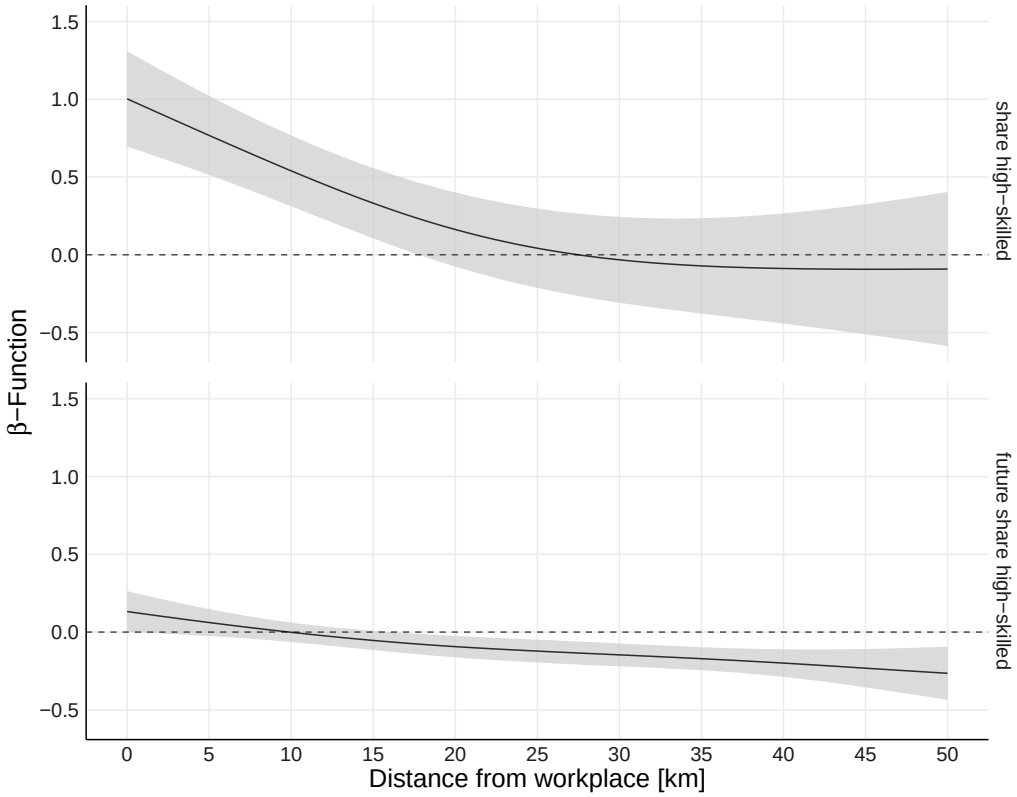
Another concern may be that global labor market shocks influence our findings through local industry or occupation clusters. If, for instance, wages and the demand for skilled labor temporarily rise within a sector and firms in this sector tend to cluster locally, our estimates would capture a spurious 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. See appendix A.6.3 for details.

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 associated estimates imply that human capital externalities are considerably stronger in urban areas than in ru-

ral 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. See appendix A.6.4 for details.

Appendix A.4 shows that our results are robust to alternative choices of the penalty parameter ρ. Finally, appendix A.5 outlines that the magnitude of the effects from our functional model is close to comparable estimates at the county level.

Figure 9: 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 (2.9), expanded with the lead of $X_{it}(z)$, with the estimator (2.5). 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 percent 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). $N = 2,959,357$**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

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 who 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 percent.

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/Kneip/Sarda (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/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.

References

Abadie, Alberto; Athey, Susan; Imbens, W., Guido; Wooldridge, Jeffrey (2017): When Should You Adjust Standard Errors for Clustering? Working Paper 24003, National Bureau of Economic Research, URL http://www.nber.org/papers/w24003.

Abowd, M., John; Kramarz, Francis; Margolis, N., David (1999): High Wage Workers in High Wage Firms. In: Econometrica, Vol. 67, No. 2, p. 251–333.

Acemoglu, Daron (1998): Why do new technologies complement skills? Directed technical change and wage inequality. In: The Quarterly Journal of Economics, Vol. 113, No. 4, p. 1055–1089.

Acemoglu, Daron (1996): A Microfoundation for Social Increasing Returns in Human Capital Accumulation. In: The Quarterly Journal of Economics, Vol. 111, No. 3, p. 779–804.

Acemoglu, Daron; Angrist, Joshua (2000): How Large Are Human-Capital Externalities? Evidence from Compulsory Schooling Laws. In: Bernake, Ben S.; Rogoff, Kenneth (Eds.) NBER Macroeconomics Annual 2000, Vol. 15, MIT Press, p. 9–74.

Acemoglu, Daron; Angrist, Joshua (1999): How Large are the Social Returns to Education? Evidence from Compulsory Schooling Laws. Working Paper 7444, National Bureau of Economic Research, URL http://www.nber.org/papers/w7444.

Ahlfeldt, M., Gabriel; Redding, J., Stephen; Sturm, M., Daniel; Wolf, Nikolaus (2015): The economics of density: Evidence from the Berlin Wall. In: Econometrica, Vol. 83, No. 6, p. 2127–2189.

Andersson, Martin; Larsson, Johan P; Wernberg, Joakim (2019): The economic microgeography of diversity and specialization externalities - firm-level evidence from Swedish cities. In: Research Policy, URL http://www.sciencedirect.com/science/article/pii/S0048733319300447.

Antoni, Manfred; Ganzer, Andreas; vom Berge, Philipp (2016): Sample of Integrated Labour Market Biographies (SIAB) 1975-2014. Institute of Employment Research, Nuremberg, URL http://doku.iab.de/fdz/reporte/2016/DR_04-16_EN.pdf, fDZ-Datenreport 04/2016.

Arzaghi, Mohammad; Henderson, J. Vernon (2008): Networking off Madison Avenue. In: The Review of Economic Studies, Vol. 75, No. 4, p. 1011–1038.

Autor, H., David; Katz, F., Lawrence; Kearney, S., Melissa (2008): Trends in U.S. wage inequality: revising the revisionists. In: The Review of Economics and Statistics, Vol. 90, No. 2, p. 300–323.

Borjas, George J. (2003): The Labor Demand Curve Is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market. In: The Quarterly Journal of Economics, Vol. 118, No. 4, p. 1335–1374.

Caldeira, JoÃo; Torrent, Hudson (2017): Forecasting the US term structure of interest rates using nonparametric functional data analysis. In: Journal of Forecasting, Vol. 36, No. 1, p. 56–73.

Card, David; Heining, Jorg; Kline, Patrick (2013): Workplace Heterogeneity and the Rise of West German Wage Inequality. In: The Quarterly Journal of Economics, Vol. 128, No. 3, p. 967–1015.

Card, David; Lemieux, Thomas (2001): Can falling supply explain the rising return to college for younger men? A cohort-based analysis. In: The Quarterly Journal of Economics, Vol. 116, No. 2, p. 705–746.

Cardot, Hervé; Mas, André; Sarda, Pascal (2007): CLT in functional linear regression models. In: Probability Theory and Related Fields, Vol. 138, No. 3-4, p. 325–361.

Ciccone, Antonio; Peri, Giovanni (2006): Identifying Human-Capital Externalities: Theory with an Application to US Cities. In: The Review of Economic Studies, Vol. 488, No. 73, p. 381–412.

Ciccone, Antonio; Peri, Giovanni (2005): Long-run substitutability between more and less educated workers: evidence from U.S. states, 1950-1990. In: The Review of Economics and Statistics, Vol. 87, No. 4, p. 652–663.

Cornelissen, Thomas; Dustmann, Christian; Schönberg, Uta (2017): Peer Effects in the Workplace. In: American Economic Review, Vol. 107, No. 2, p. 425–456.

Crambes, Christophe; Kneip, Alois; Sarda, Pascal (2009): Smoothing splines estimators for functional linear regression. In: The Annals of Statistics, Vol. 37, No. 1, p. 35–72.

Dauth, Wolfgang; Haller, Peter (2018): Berufliches Pendeln zwischen Wohn- und Arbeitsort. IAB-Kurzbericht, Institute for Employment Research (IAB), Nuremberg.

Davis, R., Donald; Dingel, I., Jonathan (2019): A spatial knowledge economy. In: American Economic Review, Vol. 109, No. 1, p. 153–70.

De la Roca, Jorge; Puga, Diego (2017): Learning by Working in Big Cities. In: Review of Economic Studies, Vol. 84, No. 1, p. 106–142.

Dustmann, Christian; Ludsteck, Johannes; Schönberg, Uta (2009): Revisiting the German Wage Structure. In: The Quarterly Journal of Economics, Vol. 124, No. 2, p. 843–881.

Eppelsheimer, Johann; Möller, Joachim (2019): Human capital spillovers and the churning phenomenon: Analysing wage effects from gross in-and outflows of high-skilled workers. In: Regional Science and Urban Economics, Vol. 78, p. 103 461.

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

Faggio, Giulia; Schluter, T; vom Berge, Philipp (2019): Interaction of public and private employment: Evidence from a German government move. In: .

Ferraty, Frédéric; Vieu, Philippe (2006): Nonparametric Functional Data Analysis - Theory and Practice. New York: Spinger.

Fitzenberger, Bernd; Osikominu, Aderonke; Völter, Robert (2005): Imputation Rules to Improve the Education Variable in the IAB Employment Subsample. Institute of Employment Research, Nuremberg, URL doku.iab.de/fdz/reporte/2005/MR_3.pdf, FDZ-Methodenreport 03/2005.

Fu, Shihe (2007): Smart Café Cities: Testing human capital externalities in the Boston metropolitan area. In: Journal of Urban Economics, Vol. 61, No. 1, p. 86–111.

Gibbons, Stephen; Overman, Henry G; Sarvimäki, Matti (2017): The local economic impacts of regeneration projects: Evidence from UK's Single Regeneration Budget. In: .

Glaeser, L., Edward; Mare, C., David (2001): Cities and Skills. In: Journal of Labor Economics, Vol. 19, No. 2, p. 316–342, URL https://doi.org/10.1086/319563.

Hall, Peter; Horowitz, Joel L (2007): Methodology and convergence rates for functional linear regression. In: The Annals of Statistics, Vol. 35, No. 1, p. 70–91.

Horváth, Lajos; Kokoszka, Piotr (2012): Inference for Functional Data with Applications. New York: Springer.

Hsing, Tailen; Eubank, Randall (2015): Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Chichester: John Wiley & Sons.

Katz, F., Lawrence; Murphy, M., Kevin (1992): Changes in Relative Wages, 1963-1987: Supply and Demand Factors. In: The Quarterly Journal of Economics, Vol. 107, No. 1, p. 35–78.

Kosfeld, Reinhold; Werner, Alexander (2012): Deutsche Arbeitsmarktregionen - Neuabgrenzung nach den Kreisgebietsreformen 2007-2011. In: Raumforschung und Raumordnung, Vol. 70, No. 1, p. 46–64.

Krusell, Per; Ohanian, E., Lee; Rios-Rull, Jose-Victor; Violante, L., Giovanni (2000): Capital-skill complementarity and inequality: a macroeconomic analysis. In: Econometrica, Vol. 68, No. 5, p. 1029–1053.

Lucas, Robert E. (1988): On the mechanics of economic development. In: Journal of Monetary Economics, Vol. 22, p. 3–42.

Manning, Alan; Petrongolo, Barbara (2017): How local are labor markets? Evidence from a spatial job search model. In: American Economic Review, Vol. 107, No. 10, p. 2877–2907.

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

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

Morris, Jeffrey S (2015): Functional Regression. In: Annual Review of Statistics and Its Application, Vol. 2, p. 321–359.

Nelson, Richard R.; Phelps, Edmund S. (1966): Investment in humans, technological diffusion, and economic growth. In: The American Economic Review, Vol. 56, No. 1/2, p. 69–75.

Ramsay, J.O.; Silverman, B.W. (2005): Functional Data Analysis. New York: Springer, 2. ed..

Ramsay, O., James; Ramsey, B., James (2002): Functional data analysis of the dynamics of the monthly index of nondurable goods production. In: Journal of Econometrics, Vol. 107, No. 1-2, p. 327–344.

Rauch, James E. (1993): Productivity Gains from Geographic Concentration of Human Capital: Evidence from the Cities. In: Journal of Urban Economics, Vol. 34, No. 3, p. 380–400.

Rosenthal, Stuart S.; Strange, William C. (2008): The attenuation of human capital spillovers. In: Journal of Urban Economics, Vol. 64, No. 2, p. 373–389.

Schmucker, Alexandra; Seth, Stefan; Ludsteck, Johannes; Eberle, Johanna; Ganzer, Andreas (2016): The Establishment History Panel 1975-2014. Institute of Employment Research, Nuremberg, URL http://doku.iab.de/fdz/reporte/2016/DR_03-16_EN.pdf, FDZ-Methodenreport 03/2016.

Ullah, Shahid; Finch, Caroline F. (2013): Applications of functional data analysis: A systematic review. In: BMC Medical Research Methodology, Vol. 13, No. 1, p. 43, URL https://doi.org/10.1186/1471-2288-13-43.

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

Wang, Shanshan; Jank, Wolfgang; Shmueli, Galit (2008): Explaining and forecasting online auction prices and their dynamics using functional data analysis. In: Journal of Business & Economic Statistics, Vol. 26, No. 2, p. 144–160.

Appendix

A.1 Imputation of wages

A common limitation of social security data is the right-censoring of earnings. To address this issue, we follow Dustmann/Ludsteck/Schönberg (2009) and Card/Heining/Kline (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.

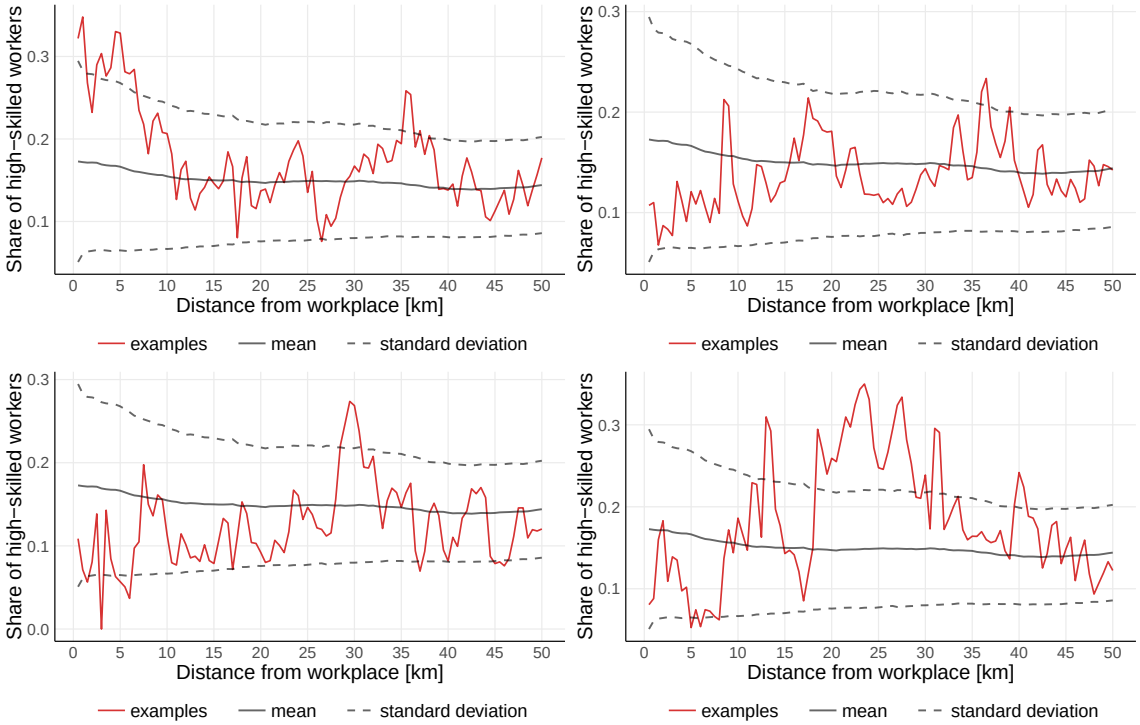
In the second step, we compute the lifetime average wages of each worker and firm, excluding the focal period. For workers and firms with only one observation, we assign the sample mean. With the period-specific lifetime average wages as additional predictors, we repeat the Tobit estimates. Finally,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

A.2 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 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 percent in the near neighborhood of the workplace. Between 5 and 15 kilometers' distance, the share of high-skilled workers declines to 15 percent. 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 percent. The remaining three panels illustrate different patterns.

Figure 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.**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

A.3 Summary statistics

The dataset used in our econometric analysis covers 15 years and consists of 3.5 million records of 540,000 workers. Table A.1 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 15 years of work experience. The median population density in the dataset is 119 inhabitants per square kilometer ($\exp(4.78)$). Furthermore, 36 percent of the observations are from females and 7 percent are from workers with foreign nationality. The proportions of low-, medium- and high-skilled workers are 8, 73 and 19 percent, respectively.

Table 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.**

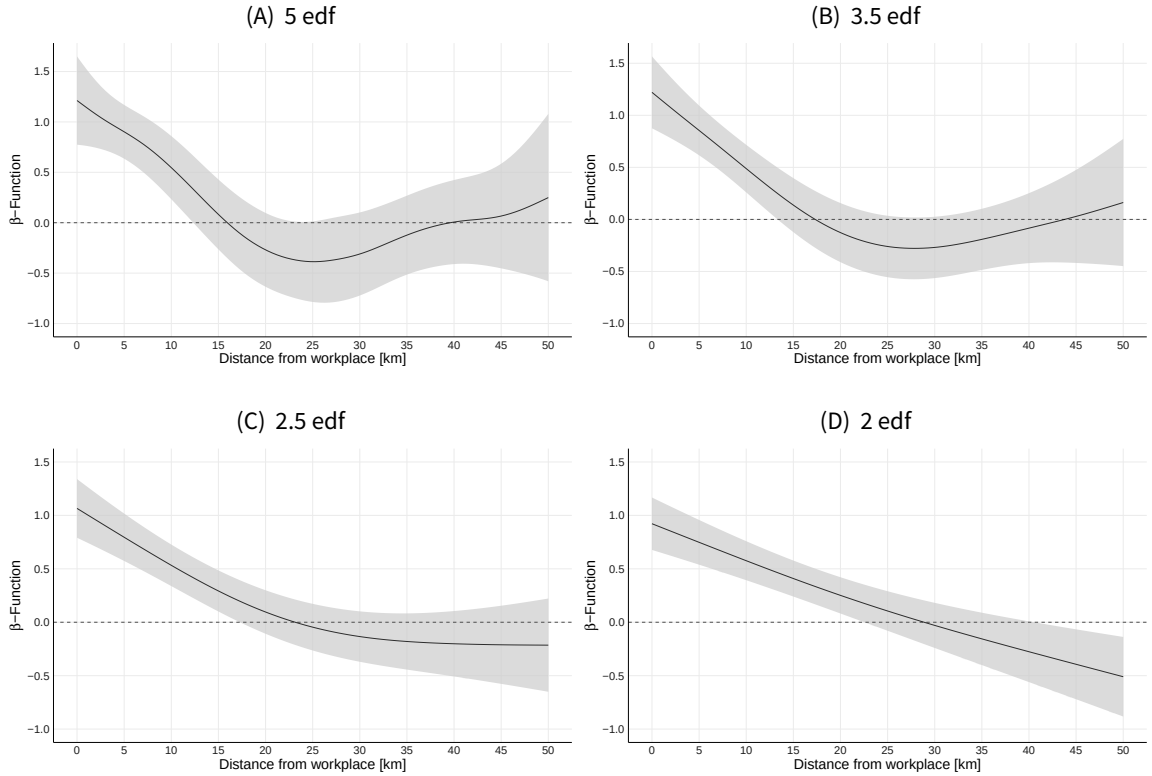
Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

A.4 Estimates with different penalties

In our preferred specification, we estimate equation (2.9) with the estimator (2.5) 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 A.2 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 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 nonsignificant after 13 to 23 kilometers. The confidence bands of the four estimates overlap over 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

Figure A.2: 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 (2.9) with the estimator (2.5). 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 percent 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).** $N = 3,498,536$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

spillover function from panel D as overly inflexible.

A.5 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/Kneip/Sarda (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 (equation (2.9)):

$$Y_{it} = \alpha x_{it} + Z'_{it}\gamma + \theta_{if} + \tau_t + \omega_o + \pi_{rst} + u_{it}. \tag{5.1}$$

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 (5.1), we use the same dataset as in the paper and cluster standard errors at the county-level. Table A.2, 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 percent. 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 percent. 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.

Table A.2: 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.** $N = 3,498,536$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

A.6 Robustness

A.6.1 Semi-parametric OLS estimates with broader rings

When estimating the spatial attenuation of economic effects, the literature follows a semi-parametric approach (e.g., Rosenthal/Strange, 2008; Fu, 2007; Verstraten, 2018; Gibbons/Overman/Sarvimäki, 2017; Faggio/Schluter/vom Berge, 2019; Faggio, 2019). In such models, econometricians estimate linear models in which the main explanatory variable is measured in several geographically concentric rings or circles around observations. The bandwidth of the rings or circles are usually of varying size. As a robustness exercise, we apply such a procedure to our application.

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 generated 1,000 replications of the DPG (2.1) using predictors resembling the first and second moments of our real data application. The functional coefficient β_0 corresponds to the dashed line of Figure A.3. We then computed averages of the simulated curves with respect to larger intervals of the domain.[9] 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 (5.2).

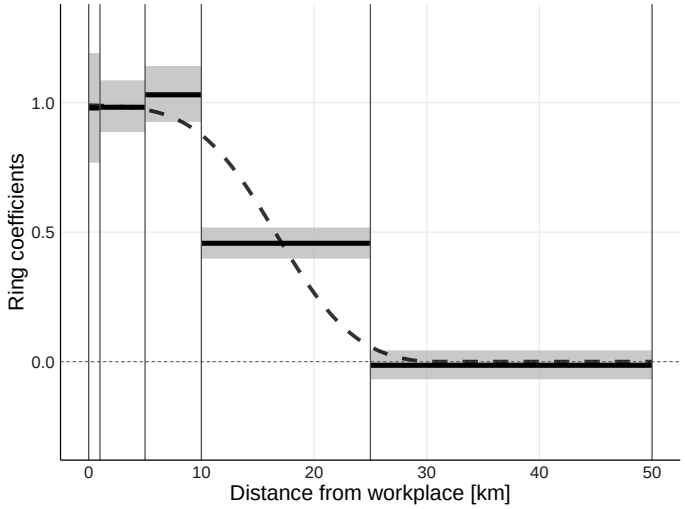
In Figure A.3, we illustrate the results of the simulation study. The coefficient function of the DGP is depicted by the dashed line, and the vertical solid lines indicate boundaries of the rings used in our specification. The grey areas illustrate the first and 99$^{\text{th}}$ percentiles of all replications, and the vertical black lines represent the mean over all replications. In general, the results show that the approximation via a Riemann sum also works quite well, but the outcome heavily depends on how the rings are defined. In addition, such an estimation framework does not allow learning from the data how the coefficient function behaves inside the intervals.

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

$$\begin{aligned} Y_{it} = {} & \alpha_1 x_{1\text{km},it} + \alpha_2 x_{5\text{km},it} + \alpha_3 x_{10\text{km},it} + \alpha_4 x_{25\text{km},it} + \alpha_5 x_{50\text{km},it} \\ & + Z_{it}'\gamma + \theta_{if} + \tau_t + \omega_o + \pi_{rst} + u_{it}. \end{aligned} \tag{5.2}$$

[9] 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 observations obtain a greater weight than 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 A.3: 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 (5.2) on data generated by DGP (2.1) but with the same predictors used in the Monte-Carlo exercise described in section 4.2.**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

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, α_z is the spillover coefficient we seek to estimate. In line with our main model, we control for the 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 A.3 summarizes the results. Column 2 of table A.3 shows 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 a distance of 25 kilometers.

Due to different bandwidths, we cannot directly compare the magnitude of the raw estimates. To illustrate the issue, 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 these individuals. The parameter estimate on the second ring expresses the effects of an one-percentage-point increase at a one to five kilometer 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 work-

Table A.3: 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 $\times$ year $\times$ 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.** $N = 3,498,536$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

ers 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 divide the raw estimates by their underlying bandwidth in column 4. The corresponding 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, column 4 shows that human capital externalities decay with distance. Also similar to our main findings, human capital externalities lose their economic significance between 10 to 25 kilometers of 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-point increase in the share of high-skilled workers within five kilometers leads to wage gains of 1.75 percent. According to our semi-parametric estimates with broader rings, the same increase in the share of high-skilled workers raises wages by 2 percent. The difference between the two estimates is minor. In summary, the semi-parametric estimates buttress our main findings.

A.6.2 Non-border regions

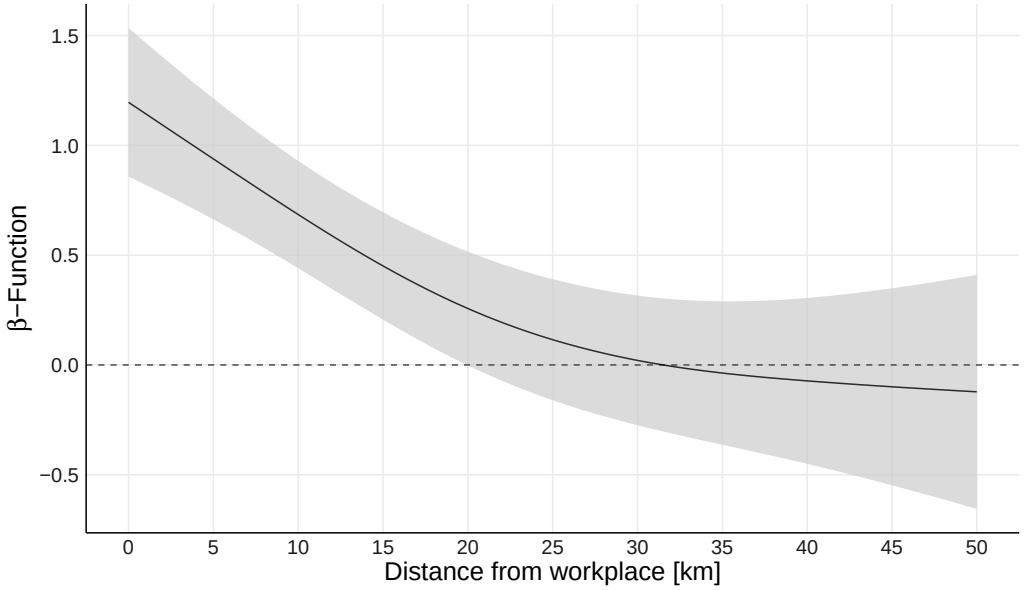
Because we have no data on workers outside of Germany, measurements of the distribution of high-skilled workers in border regions are partly truncated. For instance, establishments in the city center of Passau are only two kilometers from the Austrian border. Therefore, past two kilometers' distance, 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 A.4 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.

A.6.3 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

Figure A.4: Spatial human capital externalities from high-skilled workers (without border regions)



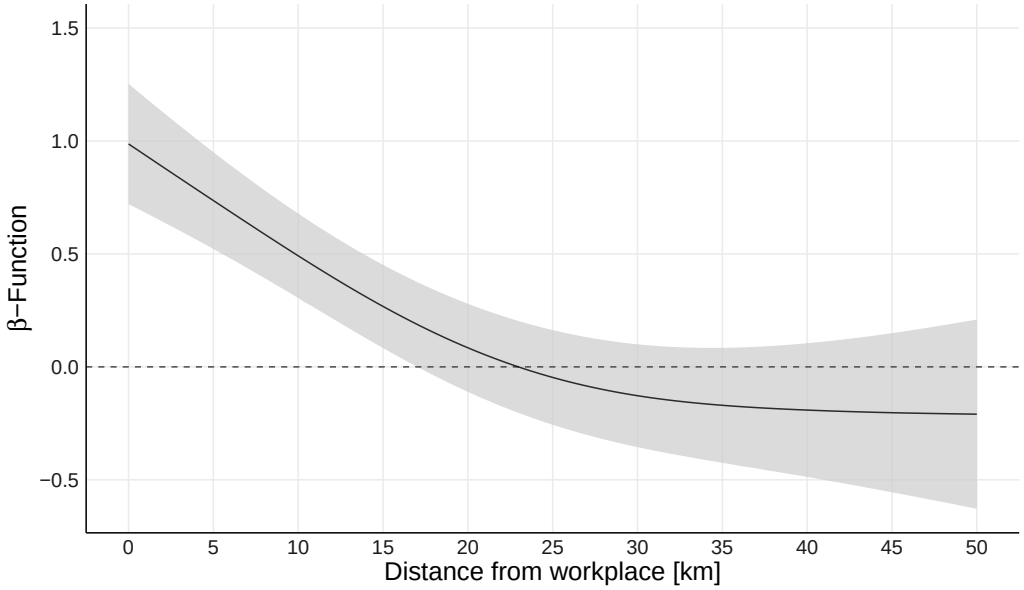
Notes: **The figure shows spatial human capital externalities from high-skilled workers into individual log wages in regions that are at least 50 kilometers from the German border. We measure the concentration of high-skilled workers as the share of high-skilled workers within distance z. To compute the spatial spillover function ($\beta(z)$), we estimate equation (2.9) with the estimator (2.5). 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 gray area indicates the 99 percent 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).** $N = 2,489,083$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

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.

To assess whether industry or occupation trends in the global labor market affect our results, we augment our estimation equation (equation (2.9)) 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 A.5 shows the resulting spillover function. The curve is almost identical to that from our main specification (figure 5). We therefore conclude that trends at the industry or occupational level do not influence our results.

Figure A.5: 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 equation (2.9) with the estimator (2.5). 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 percent 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). $N = 3,498,536$**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

A.6.4 Effects in urban and rural areas

Plausibly, marginal travel costs for physical distance differ between 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 A.6 and figure A.7 illustrate the estimates of human capital externalities 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 the wages of workers in cities by 2.5 percent. The same increase in the share of high-skilled workers raises wages of the average worker by only 1.75 percent.

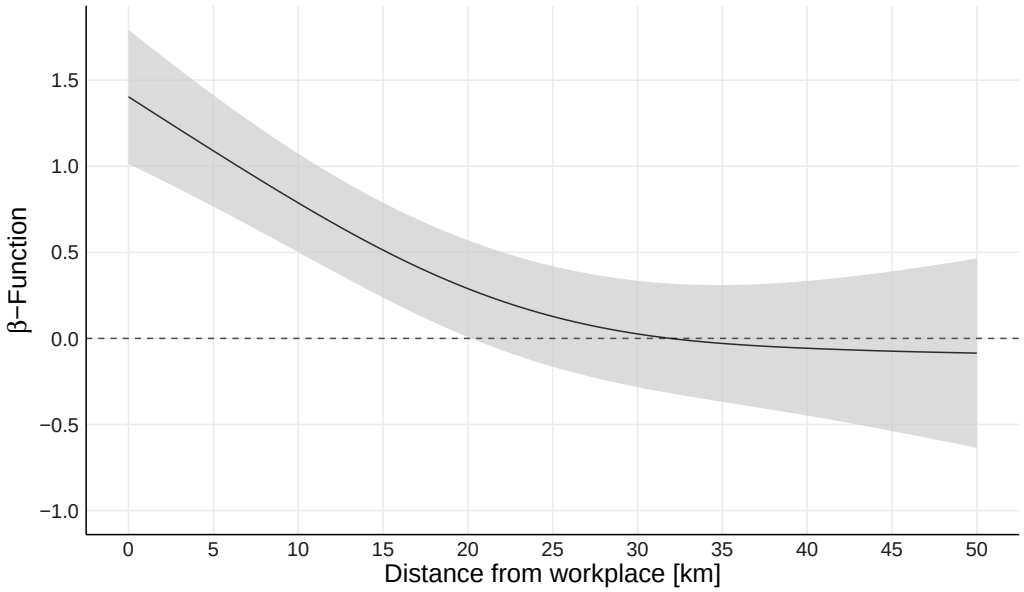
Contrarily, as figure A.7 indicates, estimates of human capital externalities in rural areas are nonsignificant. These results suggest that workers in rural areas do not benefit from human capital externalities. Our identification strategy relies on a extensive set of fixed effects that remove all variation in the data that comes from the labor market area and time-invariant individual and establishment characteristics. Thus, we only measure human capital externalities from 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 and 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 the number in urban areas, we cannot rule out that nonsignificant results in the rural sample might be due to efficiency issues. Figure A.8 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 A.8 are therefore less demanding because they use not only time-variant variation in the data but also variation between workplaces. Allowing between variation leads to significant estimates of human capital externalities. However, estimates 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.

A.7 Estimates of spatial human capital externalities: full table

Table A.4 presents parameter estimates from our preferred specification and accompanies figure 6. 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 nonsignificant 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 (2.9)), the parameter estimates for county-level variables are statistically nonsignificant.

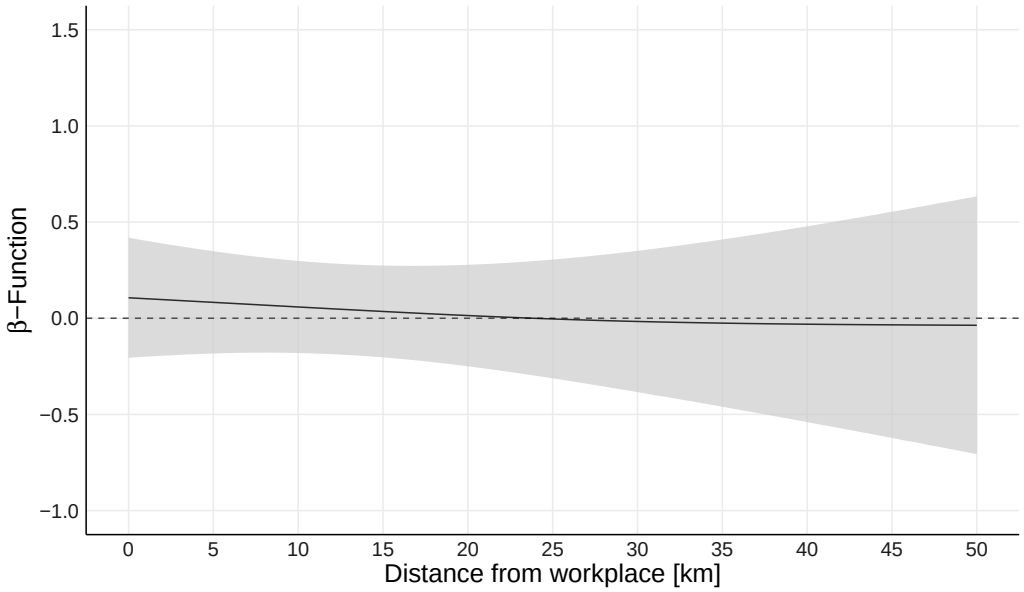
Figure A.6: Spatial human capital externalities from high-skilled workers (urban areas)



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 equation (2.9) with the estimator (2.5). 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 gray area indicates the 99 percent 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, 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).** $N = 2.601.624$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

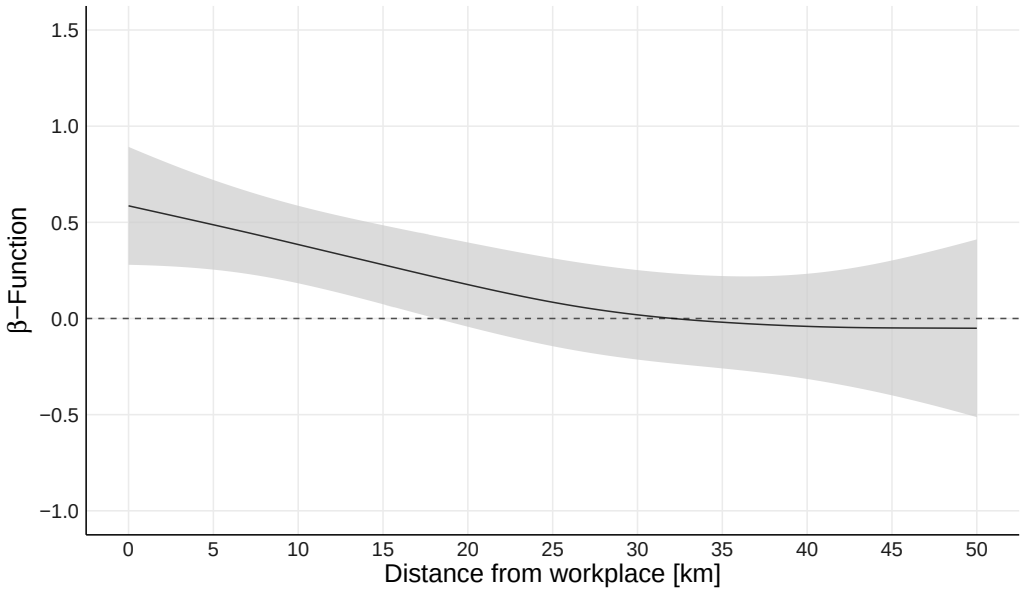
Figure A.7: Spatial human capital externalities from high-skilled workers (rural areas)



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 equation (2.9) with the estimator (2.5). 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 gray area indicates the 99 percent 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, 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).** $N = 896.912$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

Figure A.8: 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 equation** (2.9) **with the estimator** (2.5)**. 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 gray area indicates the 99 percent 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). $N = 896.912$**

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

Table A.4: 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 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 (2.9) with the estimator (2.5). 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 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.** $N = 3,498,536$

Source: **Own calculations, IAB-SIAB (7514 v1), IAB-BHP (7516 v1), IAB-IEB-GEO (v01.00.00.1504)**

Imprint

IAB-Discussion Paper 21|2020

Publication Date
22 July 2020

Publisher
Institute for Employment Research
of the Federal Employment Agency
Regensburger Straße 104
90478 Nürnberg
Germany

Download
http://doku.iab.de/discussionpapers/2020/dp2120.pdf

All publications in the series "IAB-Discusssion Paper" can be downloaded from
https://www.iab.de/en/publikationen/discussionpaper.aspx

Website
www.iab.de/en

Corresponding author
Johann Eppelsheimer
johann.eppelsheimer@iab.de