Survey Research Methods (2017)
Vol. 11, No. 3, pp. 267-287
doi:10.18148/srm/2017.v11i3.6729

ISSN 1864-3361
http://www.surveymethods.org

Dealing with space and place in standard survey data

Steffen Hillmert
University of Tübingen
Germany

Andreas Hartung
University of Tübingen
Germany

Katarina Weßling
University of Cologne
and University of Tübingen
Germany

Heterogeneity of local conditions and spatial dependencies are typical aspects of sociological phenomena. However, large-scale empirical data is often rather limited with regard to the spatial references that are (publicly) available to researchers. We describe several aspects of the problem and assess available options and consequences associated with limited information. Our empirical examples are popular research topics such as returns to education and gender-based and migration-related wage gaps. We base our analyses upon widely used survey data from Germany, the German Socio-Economic Panel Study (GSOEP), which contains geographical information on various levels of aggregation. Our particular interest is focused on problems of space and place in standard large-scale socio-economic surveys and what researchers need to consider when making decisions about their analytical strategy.

Keywords: Spatial analysis; survey data; GIS; multi-level model; spatial regression; territorial classification; returns to education; wage gap; labor market; GSOEP. *JEL:* R12; C83; J31

1 Introduction

There are two major geographical aspects in empirical analyses with survey data: one is related to "place", while the other is related to "space" (Logan, 2012) more specifically, the first aspect refers to *heterogeneity* across space, whereas the second aspect refers to spatial *dependencies* among the units of observation (Fotheringham, 2009b).

Against this background, many users of large-scale socio-economic survey data experience a typical contrast: On the one hand, they realize that the heterogeneity of local conditions and interdependencies among proximate units of observation are important aspects of the social phenomena they study. On the other hand, survey data often provide no information about spatial references that would allow a precise localization of the cases and make them useful for GIS-based analysis. Fortunately, data users are often provided with approximate geographical information – even though such information is typically not included in standard scientific use files. In this paper, we describe several aspects of the problem and assess the potentials of survey analyses with limited geographical information. We illustrate these aspects using the examples of monetary returns to education and gender-based and migration-related wage gaps as popular research topics.

Aspects of *place* refer to local and regional differences. They primarily raise questions of data availability: Is there explicit information about local conditions? And is there information about the location of the cases? Such information is necessary to map regional variation and to match suitable data from other sources. Matching survey data with adequate aggregate (context) data has become increasingly popular as it offers a wide range of analytical possibilities. Such matching is usually done on the basis of standard codes from specific regional classification systems. After providing an outline of conceptual challenges associated with space and place, we therefore give an overview of standard (administratively defined) spatial classification systems that are frequently used for social research in Europe as well as an overview over the geographical information that is available in major European surveys.

Aspects of *space* have primarily methodological consequences as they refer to spatial interdependencies among the units of analysis. We present a set of empirical analyses on this topic: We discuss approaches to account for spatial clustering and assess the relevance of spatial dependencies when using large-scale socio-economic surveys with restricted ge-

Contact information: Steffen Hillmert, University of Tübingen, Department of Sociology, Wilhelmstr. 36, 72074 Tübingen, Germany (email: steffen.hillmert@uni-tuebingen.de); see also http://spaceandplace.de

(a) Positive spatial autocorrelation (b) No spatial autocorrelation (c) Negative spatial autocorrelation

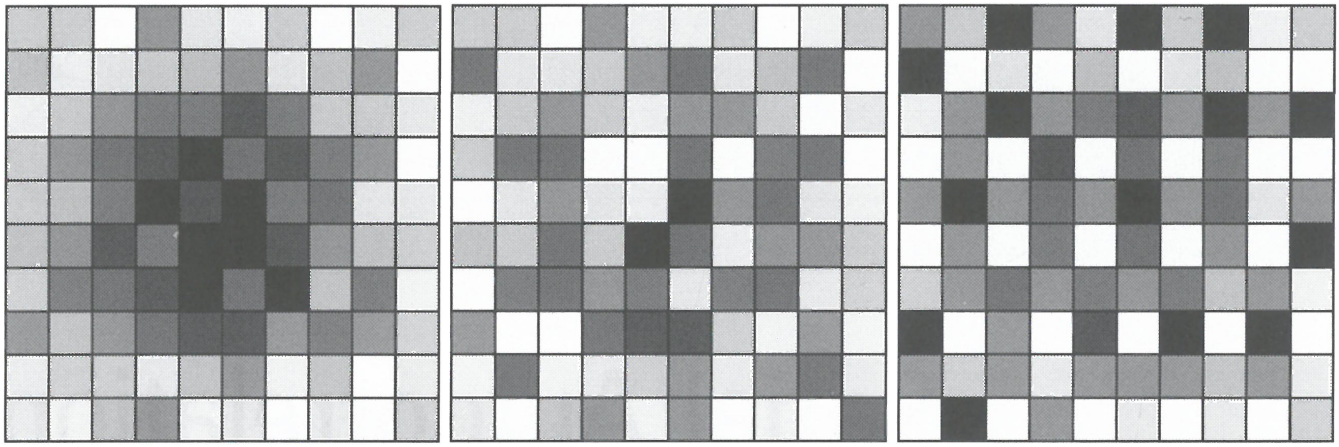


Figure 1. Spatial patterns: (a) clustered, (b) random, and (c) uniform (Fortin & Dale, 2009, p. 89).

ographical information. We begin with results from analyses using ordinary least square (OLS) regression. In the next step, multi-level techniques are applied whose basic requirement is information about which regional cluster an individual can be assigned to. Finally, we turn to spatial regression models as special techniques to account for spatial dependencies among the cases (e.g. Ward & Gleditsch, 2008). Such models require at least proxy information about the localization of the cases and are routinely used in the analysis of spatial interrelations. Our illustrative analyses are based on the German Socio-Economic Panel Study (GSOEP), but users of other standard survey datasets are faced with similar restrictions.

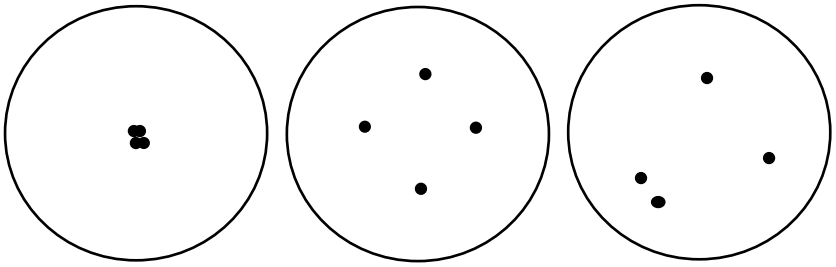
We conclude with a number of general recommendations for data users.

2 Analytical problems of spatial heterogeneity and clustering

A focus on "place" emphasizes aspects of substantive heterogeneity when comparing specific localities, i.e., the fact that life conditions typically vary geographically. Average results using unspecified (pooled) data therefore give only a very limited picture of the real situation.

Aspects of "space" refer directly to the *relative* position of observations. These observations are often not randomly or evenly distributed in geographical space, but are to a considerable degree clustered (see Figure 1). Clustering results from two potential sources. It may be a reflection of both empirical population patterns and cluster-based sampling of the survey data ("clustering by design"). Spatial clustering in surveys in terms of an uneven or non-random distribution of individual cases is not a problem per se. It is problematic if the clustering of cases corresponds with a systematic spatial variation of relevant individual or regional characteristics, a phenomenon called (positive) *spatial autocorrelation*. Spatial autocorrelation exists if values at one locality are systematically associated with values at proximate localities (Fortin & Dale, 2009). Hence, the concept of spatial autocorrelation highlights additional dimensions of possible inter-dependencies in the observed data. For example, individuals in a sample who live next to each other may be similar with regard to their income or other variables of interest. This is unproblematic as long as these differences are captured by the model parameters. However, there might be characteristics relevant for the level of individual income that vary geographically and are not observed or even considered by the researcher, such as local labor-market conditions, infrastructure, or self-selection of particular employers in certain regions. When spatially proximate cases are typically more similar than average, the basic assumption of independent sampling units in OLS regression is violated and the standard errors of estimated parameters are estimated incorrectly. Furthermore, there may be direct (causal) relationships between proximate sample units.

Various statistical techniques are available to account for spatial heterogeneity and dependencies (see the following sections) that can be implemented using standard GIS software packages, but they require information about the localization of cases. The easiest way is the use of exact *geocoded positions*, i.e., data which includes the exact geographical coordinates of the cases in terms of longitude and latitude. However, in standard large-scale socio-economic survey data, the provision of geocoded data is the exception, which is not least the result of corresponding data protection regulations. However, data users are often provided with *approximate* geographical information (regional identifiers). This means that the approximate position of a case can be inferred from its affiliation with a larger geographical unit for which a geographical code is available. The units are typically denoted according to standard systems of territorial



(a) Assignment (close) to the centroid of the spatial unit

(b) Assignment with maximum distances (between individual cases and between individual cases and the border of the unit)

(c) Random assignment within the spatial unit

Figure 2. Spatial distribution of individuals within spatial units (examples).

classification whose geo-references are publicly available in the form of geographical shapefiles. In addition, clustering identifiers are often provided in survey data sets as part of describing the (complex) sampling design. This information on the primary sampling units (PSUs) may also be used to account for spatial dependencies. Note, however, that they typically refer to only one aspect of spatial clustering ("clustering by design"). They represent a fixed level of aggregation and they may differ among various subsamples. Moreover, they typically contain no geographical information on the actual location ("place"), so that regional variation and potential spatial associations *beyond* the clusters cannot be considered. Finally, survey data may contain information on local conditions without any information to which particular regional unit this information relates and where this unit is located.

In the empirical analyses below, we will use several of the standard classification systems and compare the results for individual analyses. Prior to this, we will give a brief description of the most common territorial classification systems. In addition to our discussion of spatial dependencies, there is another good reason to become familiar with such classification systems. Many researchers are interested in describing the effects of local *context conditions* on individual behavior, e.g., analyses of the effect of regional infrastructure on education or the effect of local labor-market conditions on career mobility. Standard classification systems also allow the merging of individual-level information from surveys with adequate context data from other sources, such as aggregate official statistics.[1]

3 Making use of the geographical information in survey data

In practice, dealing with basic aspects of "place" is relatively straightforward. Regional variation in relevant survey-based information is mapped descriptively, or identifying information on spatial units is used for matching with external (context) data.

Spatial dependencies can be dealt with to various degrees of sophistication. Corresponding statistical techniques have different requirements for the accuracy of the geographical location of the observations. In the case of an OLS regression, we implicitly assume that no spatial autocorrelation is present. This means that all observations are evenly distributed across the geographical space or that their geographical positions and the distances between them do not matter for the specific research question. Problems may arise if the units of observation are in fact clustered and the proximate cases are correlated in terms of relevant characteristics in some unknown (and uncontrolled) way. In particular, standard errors may be underestimated if this correlation is relevant for the considered outcome (Moulton, 1990).

A standard option to account for clustering is the use of clustered standard errors (Rogers, 1994). This method is based on the concept of robust (Huber – White) standard errors that are used to account for heteroscedasticity of the model residuals. While the standard approach to compute robust standard errors assumes that model residuals are independently distributed, a generalized form relaxes this assumption and replaces it with the assumption of independence between clusters. Such a model allows for correlations among the observations within clusters and any heteroscedasticity in the error term, but it also assumes no correlation among observations across clusters (Primo, Jacobsmeier, & Milyo, 2007). Such simple clustering can be performed either as part of the data definition (such as `svyset` in Stata's survey commands) or as a specification of the analytical model. It can account for dependencies due to spatial clustering if these dependencies work predominantly *within* the specified geographical units. Such models do not require information about the exact localization of the regional clus-

[1] Researchers are not necessarily bound to fixed classification systems because units of such systems can be flexibly aggregated in issue-specific, suitable ways (Weßling, Hartung, & Hillmert, 2015).

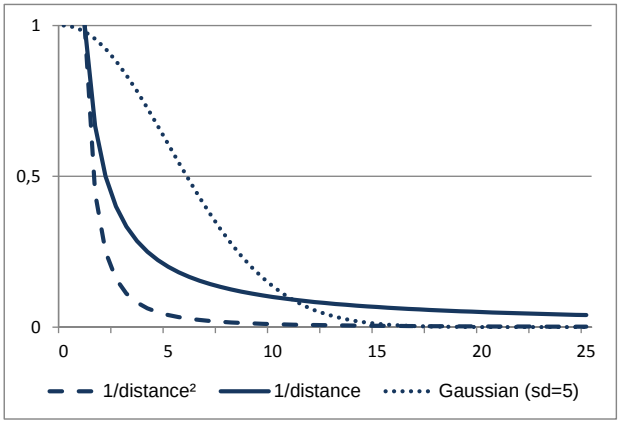


Figure 3. Illustrative distance-based weighting schemes: $\frac{1}{\text{distance}^2}$ vs. $\frac{1}{\text{distance}}$, equivalent to a continuously strong vs. a less strong decay of weight with distance, and Gaussian-type function: $\exp\left(\frac{-\text{distance}^2}{2 \cdot 5^2}\right)$.

ters.

Spatial models can account for more complex and extended spatial dependencies, but they have also higher requirements. Two central problems are spatial allocation and spatial weighting. When spatial models require the geographical position of the individuals and the exact position is not available to the data user, this information must be approximated by using regional identifiers and the geometries of available regions. The simplest way of allocating cases is to use the geographical centroids of each considered region. In this case, individuals who share the same region will be allocated to exactly the same geographical position (see Figure 2, a). More precisely, even in this case, the exact distance between cases has to be modeled as > 0 in order to avoid invalid values when computing weighting functions as inverse distances. In our example, the distance between cases is at least one meter. Alternative modes of allocation include an assignment with maximum distances (Figure 2, b) and random assignment (c).

Furthermore, spatial regression models require a function that represents the importance of geographical distance for the weighting term. The simplest way to achieve this is to use the inverse distance, which is also the default in most applications, but in principle, the choice of the specific weighting scheme should be based on theoretical arguments of distance-related relevance referring to underlying mechanisms such as density and range of social interactions, commuting or communication. Spatial weighting functions typically imply that the importance of proximate cases decreases steadily with their distance, but there are many alternatives with specific profiles (cf. Figure 3). Our empirical analyses in the sections below will show that the choice of the weighting function has consequences for the results.

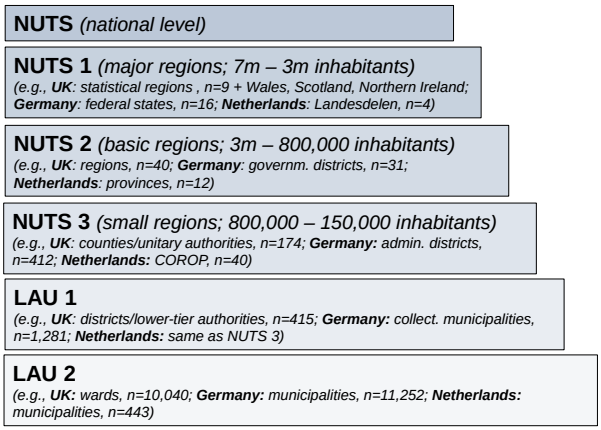


Figure 4. International administrative regional classifications (with national examples): European local and regional classification systems by Eurostat. NUTS: Nomenclature des unités territoriales statistiques; LAU: local administrative unit. Numbers of units refer to the year 2009 (Statistisches Bundesamt, 2016).

4 Standard regional classifications

Regional identifiers in survey data are used for detailed analyses on the regional level, for matching with external (context) data, and for approximating the locations of individual cases. They typically refer to standard regional classification systems. Table A3 in the appendix provides an overview of geographical information in major European surveys and access regulations.

Public administration uses several regional classifications. For an illustration of the European classification system, see Figure 4. The largest territorial units in the administrative classification are nation states; the smallest units are municipalities or their equivalents. Apart from these administration-based systems, several alternative classifications have been developed for analytical purposes. These include labor-market regions, rural and urban areas, and economic centers. These regional concepts are often based on empirical quantities such as commuter flows and economic activities. Some of the classifications follow administrative borders, while others overlap but do not directly match the official structures. For geographical illustrations of smaller levels of aggregation, ZIP (postal) code areas are suitable. They are not part of administrative classification systems as they have been conceptualized by companies such as *Deutsche Post* in Germany. Administrative structures and ZIP-code areas do not follow the same boundaries, but it is generally possible to match administrative area codes (e.g., municipalities) approximately or proportionately with ZIP codes (e.g. Statistisches Bundesamt, 2016).

Tables of conversion among classifications are provided,

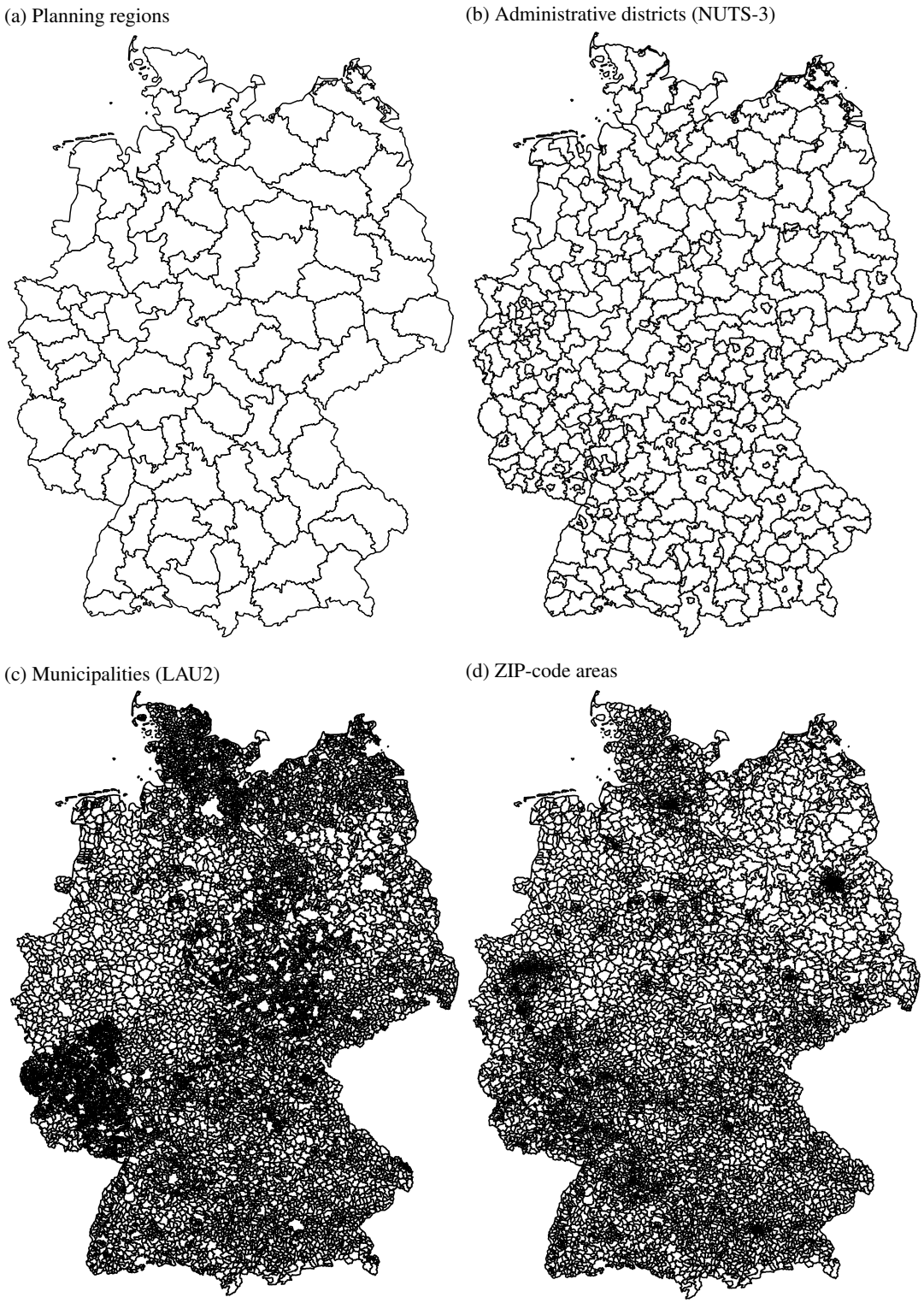


Figure 5. Regional classifications in Germany. Shapefiles: (BKG, 2016; Deutsche Post, 2003); authors' illustration.

for example, by the European Commission's Eurostat and national statistical offices (Eurostat, 2013; Statistisches Bundesamt, 2016). Moreover, commercial suppliers (e.g., *microm Consumer Marketing*, *infas360*) provide their own concepts of small-scale aggregation levels, such as living quarters or building block areas. Such data are normally not available for free.

Figure 5 maps the most important classification systems for Germany.

- *Planning regions* (PRs; *Raumordnungsregionen*) are not part of the administrative regional classification system but are based on the administrative structures. The basic units of PRs are administrative districts (NUTS-3). Also, PRs can be aggregated up to the level of federal states. The purpose of the PRs is to describe the functional classification of larger economic centers and their associated peripheries. PRs are conceptualized on the basis of commuter flows and other information. Since 2009, there have been 96 PRs in Germany. The area size of PRs varies between 325.6 km^2 and 7,775.7 km^2. The average area size is 3,720.6 km^2. A conversion of this value to a simple geometric form may allow a more intuitive interpretation: A circle with this area size would have a radius of approximately 34.4 km.
- In 2009, there were 412 *administrative districts* (*Kreise und kreisfreie Städte*, NUTS-3) in Germany. Administrative districts represent the basic analytical unit for regional statistics and are adjusted to the European regional classification system. These districts represent the NUTS-3 units conceptualized by Eurostat and therefore allow for international comparisons. The NUTS regions follow the national administrative structures. For researchers, it is important that comprehensive geographical information is available for these territorial units. NUTS-3 regions are the smallest of the NUTS units. They are suitable for specific analyses at the level of cities and coherent rural areas. The area size of districts varies between 35.7 km^2 and 2,881.8 km^2, with an average size of 869.0 km^2. This is equivalent to a circular area with a radius of 16.6 km.
- *Municipalities* represent the LAU2 level in the international classification. In 2009, there were 12,067 municipalities in Germany. Municipalities represent the smallest administrative classification units and are the basis for many statistics at the local level. Municipalities are dissimilar with respect to inhabitants and area size. The largest municipality according to the number of inhabitants is Berlin, with 3.4 million inhabitants; the smallest municipality (Holm Gröde in Schleswig-Holstein) has only nine inhabitants. The area size of municipalities varies between 0.31 km^2 and 891.8 km^2, with an average size of 31.4 km^2. This is equivalent to a circular area with a radius of 3.2 km.
- Finally, there is the map of *ZIP-code areas*. There are nine large ZIP-code areas denoted as 0 to 8. These numbers represent the first digit in the 5-digit ZIP code. The total number of ZIP codes in Germany is 28,683. However, a large number of these codes belong to mailboxes and large volume receivers; only 8,208 are relevant for private households (Deutsche Post, 2003). Compared with municipalities, ZIP-code areas are much more similar with regard to the number of inhabitants. Their average size is approximately 42 km^2, which is equivalent to a circular area with a radius of 3.7 km.

Once sampled individuals are linked to geographical units, this (approximate) spatial information can be used in statistical models to account for potential spatial interdependencies.

5 Empirical applications

5.1 A brief introduction to the substantive examples

To illustrate our discussion on spatial dependencies, we choose related, popular research topics as empirical examples: monetary returns to education and gender-related and migration-related wage gaps. We make use of the empirically well-tested relations between ethnic origin, gender, education, and economic returns and follow the established research methods in this field.

Educational attainment is among the most important individual-level determinants of earned income, occupational position, and labor-market security. It is therefore no surprise that returns to education have been a frequently analyzed topic in empirical sociological and economic research. Many analyses on the relevance of education for the labor market have been driven by the human capital approach. In this theory, education is understood as an investment of current resources – taking the opportunity costs of time as well as any direct costs into account – in exchange for future returns (Becker, 1993; Mincer, 1974). Moreover, education can be considered the most relevant information employers have on job applicants. Subsequently, education also functions as a signal on the job market (Spence, 1973). In empirical studies, earnings are almost always measured in logarithmic form. On the one hand, the distribution of log earnings – particularly on an hourly basis – is very close to a normal distribution. On the other hand, the log earnings function is best approximated by the linear schooling term and allows for an easy interpretation (Card, 1999). Returns to education represent the monetary return of an investment in education in terms of a percentage increase in income. Since the late 1960s, returns to education have been among the most researched topics in economics and the social sciences, and the positive effect of investments in human capital on earnings and employment is among the most robust findings. Findings on the overall returns to an additional year of education for Germany vary with respect to the measurement of education and income as well as with the analytical strategy used, but they have proven to be positive and stable (Ammermüller & Weber, 2005; Gebel & Pfeiffer, 2007; Harmon & Walker, 2000).

Beyond that, there are well-known income differentials with respect to individual characteristics such as gender or migration. In most countries, immigrants have, on average, lower wages compared to the native population, and females have lower wages than males. Average income differences between men and women can in large part be explained by differences in work- and education-related aspects such as working hours, employment status, level of education, and degree subject (e.g. Bishu & Alkadry, 2017; Fitzenberger & Wunderlich, 2002; Machin & Puhani, 2003). Concerning migration- and ethnicity-related differences in earnings, qualification differences can be regarded as one major reason. Discrimination is an additional explanation for differences in working outcomes between ethnicities (as well as between sexes). A large number of studies have confirmed multidimensional explanations and have disaggregated the relevance of specific explanations (e.g. Constant & Massey, 2005; Dustmann & Glitz, 2005; Nielsen, Rosholm, Smith, & Husted, 2001).

Returns to education and the effect of determinants of income are also likely to depend on location ("place"). Local or regional contexts are unequal in terms of socio-economic conditions that can influence individual outcomes, such as returns to education. We can therefore expect regional variability in outcomes. Relevant local contexts do not necessarily follow administrative boundaries. Since local labor-market conditions encourage processes of (self-)selection, there may also be disproportionate similarities and mutual influences between proximate units of observation, so that also aspects of "space" matter.

5.2 Data and calculation tools

To illustrate the alternatives and challenges of taking aspects of locality in survey data into account, we use the German Socio-Economic Panel (SOEP) (2014). The GSOEP is a survey of private households that has been carried out since 1984. The survey is conceptualized as a panel study. Respondents are household members aged 17 and older. The survey provides information on living conditions, the economic situation of individuals and households, educational careers of individuals, and a set of information on values and attitudes. The main questionnaire is a yearly, standardized instrument for individuals and households, which focuses on current living situations. Additionally, a one-time biographical questionnaire is used at the time of the first interview that records individual biographies as of the interview date. For international comparisons, there are several standardized questions in the GSOEP that are compatible with other international data sets (e.g., BHPS for the UK; cf. Frick, Jenkins, Lillard, Lipps, & Wooden, 2008). The GSOEP contains several subsamples (see Goebel, Krause, Pischner, Sieber, & Wagner, 2008). The PSUs in the GSOEP and their levels of aggregation differ among subsamples. They can be constituencies, municipalities, administrative districts or regions. Households are selected within the regional units using a random-route procedure with a random starting address and fixed selection intervals (Spieß, 2008).

For our analyses, we use individual-level data in the form of a pooled cross-sectional sample. Starting in the year 2000, i.e., the first year detailed regional information became available, we include all individual GSOEP cases between 35 and 55 years of age with valid information on our model variables. 13,952 such cases were interviewed at least once between the years 2000 and 2013. For each case, we consider the information from the first year the person was interviewed. To avoid possible bias due to clustering within households, we include only one (randomly selected) earner from a particular household. This subsampling is performed in such a way that the proportion between single-person and multi-person households remains unchanged. Due to these restrictions and missing data on the geographical location, our final sample consists of 5,832 individual cases.

The dependent variable in all of our analyses is individual gross hourly wage (in Euros) in logarithmic form. The figures have been inflation-adjusted (to the base year 2000). Besides education and migration status, independent variables on the individual level include sex, age, working hours, family status, industry or sector, and job tenure within the company. Previous research on the impact of education on earned income has confirmed these variables as highly relevant predictors (e.g. Fossen & Büttner, 2013; Strauß & Hillmert, 2011). A description of the model variables can be found in Table A1 in the appendix.

Spatial models have only recently entered the focus of survey researchers. They have been originally developed for macro-level analyses, not sample-based analyses of individuals. The consideration of non-spatial weighting in spatial modelling is still in the early stages. Most applications of spatial models do not yet incorporate individual sampling weights (Belotti, Hughes, & Piano Mortari, 2016; Mercer, Wakefield, Chen, & Lumley, 2014). For reasons of consistency, we present all of the following analyses without the use of individual sampling weights. The inclusion of individual survey weights has also made little difference for the results in our standard multivariate regression analyses.

As it is the case with comparable large-scale surveys, the GSOEP allows researchers to localize not individual respondents but larger areas in which respondents live. Different levels of territorial aggregation are available (cf. Table A3 in the appendix). The GSOEP also offers on-site access to some additional levels of aggregation (DIW, 2016). The scientific use files that can be accessed via download only contain information about the largest administrative territorial units (the federal states in which respondents live). Smaller levels of aggregation are subject to stricter data protection procedures. Municipalities and ZIP codes are only available at

the Research Data Center at the DIW (the German Institute for Economic Research) – the research institute that hosts the GSOEP. The coding of these aggregation levels allows for both the localization of territorial units and their linkage to aggregate context information. Exceptions include the classification of municipality size and the most detailed level of aggregation (neighborhood information), which only contains selected context information. A localization of specific neighborhoods would come close to a localization of individual cases. In principle, the data provider can also match the individual-level data with context information based on alternative classification systems and provided by the user. However, it needs to be ensured that the exact location of individual cases remains unidentifiable.

In the following analyses, we illustrate the regional concentration of individuals on different aggregation levels. We choose *planning regions* (PR), *administrative districts*, *municipalities* and *ZIP-code* areas to cluster individuals and apply different methodological strategies to analyze economic returns to education and wage gaps due to gender and migration status. As these levels of aggregation are ranked according to their size, a comparison of results may provide an impression of the possible consequences of the so-called *modifiable areal unit problem* (MAUP). MAUP implies that results of statistical models in which contextual information is used can be strongly affected by the level at which the contextual data is aggregated (Fotheringham & Wong, 1991; Kwan, 2012). MAUP highlights an important challenge in survey data analysis: the choice of the appropriate level of aggregation when accounting for clustering in statistical models. Both the geographical references of theoretical mechanisms and the way the empirical sample is selected should be considered (Heeringa, West, & Berglund, 2010; Hillmert, 2016).

Administrative regional units have frequently been subject to territorial reforms due to population development and regional economic situations. We use a uniform territorial state of the data and utilize all regional information as of December 31, 2009. In the prepared individual-level dataset, there are respondents in all 96 PRs. The 5,832 respondents are distributed across 402 of the 412 districts. Within the prepared dataset, we have included respondents living in 1,798 of the 12,067 municipalities and in 2,595 different ZIP-code areas.

QGIS was used for generating the maps in Figure 5, calculating spatial distributions and matching with regional context data (in shapefile format). Further calculations were performed in Stata 14 using the commands `spmat` for weighting matrices (Drukker, Peng, Prucha, & Raciborski, 2013) and `spreg` for spatial lags (Drukker, Prucha, & Raciborski, 2013). Shapefiles were converted to .dta format using `shp2dta` (Crow, 2006); results for geographically weighted regression were obtained using `gwr` (Pearce, 1998); and Figure 7 was generated using `spmap` (Pisati, 2007).

5.3 Indicators of spatial autocorrelation

The simplest and most intuitive way to get an impression of local similarity is to account for intra-class correlation of cases within defined units of clustering (Snijders & Bosker, 2012). Intra-class correlation describes how strongly cases in the same group resemble each other; this approach explicitly accounts for the grouping data structure. Intra-class correlation, however, does not consider information about the *distribution* of cases within the regions or their relations across regional borders. To account for these kinds of relationships, geo-referenced data and spatial models are indispensable. In spatial statistics, the level of autocorrelation can be quantified using different approaches. A popular concept is Moran's I (Moran, 1950). Moran's I summarizes the similarity of the values of the variables of interest at different locations as a function of the distance between cases (following Delmelle, 2009, p. 188):

$$I = \frac{N}{\sum_j \sum_k w_{jk}} \frac{\sum_{jk} w_{jk}(y_j - \bar{y})(y_k - \bar{y})}{\sum_j (y_j - \bar{y})^2} \qquad (1)$$

Denoted with y_j and y_k are the values of the variable of interest measured at the locations j and k, respectively. w_{jk} are weights based on the proximity of these two points. These weights can be defined in different ways according to theoretical considerations; usually, similarity is expected to decrease over distance (cf. Figure 3). The values of I range from -1 to $+1$. Positive values indicate positive spatial autocorrelation and negative values indicate negative spatial autocorrelation (see also Griffith & Arbia, 2010). A value of zero indicates a random spatial pattern. A transformed form of Moran's I values (z-scores) can be used for statistical hypothesis testing (Cliff & Ord, 1972).

To illustrate the degree of spatial autocorrelation in our example, we compute Moran scatter plots. The horizontal axes in the scatter plots display the z-values of the income variable, while the vertical axes display the corresponding spatial lag (of the income variable). The spatial lag variable of income is obtained by the matrix-vector-product $\mathbf{W} \times \mathbf{y}$ (with $\mathbf{y}$ the vector of the variable of interest, and $\mathbf{W}$ the square matrix holding the values of w_{jk}). For a person at location j, $\mathbf{Wy}$ represents the sum of all y-values at proximate locations k weighted with the matrix elements w_{jk} . The slope of the linear predictor in the Moran scatter plot corresponds to Moran's I; a positive slope can be interpreted as positive spatial autocorrelation (Ward & Gleditsch, 2008).

To compute a spatial weighting matrix to serve as the basis for the identification of the spatial lag, geographic locations of the cases in terms of latitude and longitude need to be specified. To get an estimate of the locations, we use the geographical centers (centroids) of the considered territorial units as approximations of the individual geographical positions and – as the default specification and in line with conventional research – the inverse distance as weighting func-

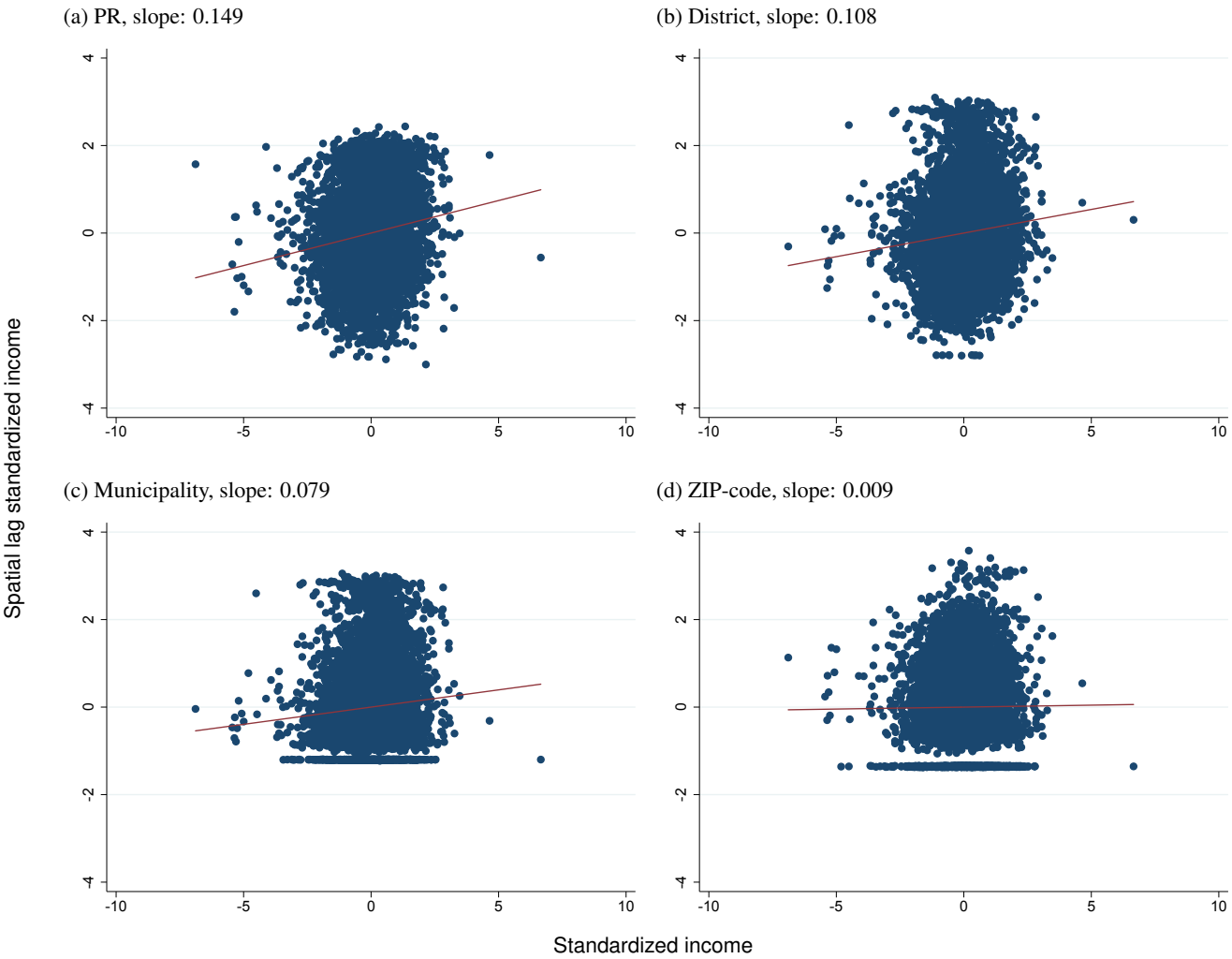


Figure 6. Linear relation between z-standardized earned income (log hourly wage) and its first-order spatial lag. Slope is equivalent to Moran's *I* (Ward & Gleditsch, 2008, p. 24). Calculations are based on inverse distance weights and approximate the individual location as geo-centers at different levels. All slopes are significantly different from 0 ($p < 0.001$) except for the level of ZIP codes. Data: GSOEP, own calculations.

tion. Geographical coordinates describing the different regional units – e.g., in the form of the geographical centroid of the region – are not part of the GSOEP. This information must be obtained with standard GIS applications using appropriate geometries in the form of shapefiles. The coordinates can then be linked with the individual-level data so that a weighting matrix between individual cases can be computed.

In Figure 6, such scatter plots have been produced using four different levels of geographical approximation and the weighting scheme $\frac{1}{\text{distance}}$ (as it was illustrated in Figure 3). As we can see, the obtained level of spatial autocorrelation is strongly dependent on the level of geographical approximation used; while there appears to be strong autocorrelation on the level of PRs, it steadily decreases when the size of considered units becomes smaller (districts, municipalities, ZIP codes).

However, a substantive interpretation of these results is difficult. Looking at the inverse distance function in Figure 3, we recognize that cases within a very close range receive very large weights. The geo-centered allocation makes the positions of individuals in the same territorial units (almost) identical, so their mutual importance is very high. The impact of these large weights is particularly consequential in larger territorial units, where relatively more cases are close to each other. Potential dependencies beyond the clusters become negligible, so the clusters resemble a simple hierarchical structure.

The general conclusion is therefore not only that estimated spatial dependencies depend on the level of aggregation on

which local and regional information is available, but also that under conditions of local approximation, the results can also be particularly sensitive to the (combined) specification of both allocation and spatial weighting.

6 Regression-based analyses (average effects)

6.1 Methods

We will now compare results on our substantive examples (determinants of earned income) using the most common types of regression-based analyses: OLS regression; OLS with clustered standard errors; multi-level (random-intercept) models; and spatial regression models.

OLS regression. We start our discussion with results from an OLS (ordinary least square) regression. This standard model has the following form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_n x_{ni} + \epsilon_i \tag{2}$$

for individual i.

OLS with clustered standard errors. Local spatial clustering is an attractive approach especially in a situation where there is information about joint cluster membership without information about the geographical positioning of these clusters. Neglecting the identifiability of territorial units in our example, such nominal clusters can be derived from the specific regional units. Some survey data sets provide adequate clustering indicators.

Multi-level analyses. A similar approach for dealing with geographically correlated error terms is to use hierarchical multi-level models with defined regions as aggregate-level (level-two) units and to compute random intercepts. Multi-level techniques allow for the determination of the impact of variables on different analytical levels. However, in this paper, we concentrate on a more specific feature of these models, their account for clustering. When computing random intercepts, we add an additional term to the model that represents the variation of the error term (or the constant) among the defined regions, i.e., the explanatory part of the regional level. Again, multi-level models may be the method of choice when there is information about local clustering (joint cluster membership) without information about the geographical positioning of these clusters. Such models still assume that local and regional mechanisms work within the boundaries of the considered local or regional units of aggregation. Furthermore, multi-level models make an implicit assumption about the distribution of the cases within the clusters; all cases within the same cluster receive exactly the same correction of the model error term. The random-intercept model has this specification (following Rabe-Hesketh & Skrondal, 2012, p. 128):

$$y_{ij} = (\beta_0 + \zeta_{0j}) + \beta_1 x_{1ij} + \ldots + \beta_n x_{nij} + \epsilon_{ij} \tag{3}$$

for individual i in region j with $\zeta_{0j} \sim N\left(0, \tau^2\right)$.

The basic requirement for estimating this model is information about which cluster an individual can be assigned to. When there are alternatives, the choice of a particular level of aggregation should follow theoretical considerations about the spatial range of relevant dependencies.

Spatial regression models. Multi-level models neglect the potential effects of equivalent geographical units that surround the unit in which an individual is located. Therefore, we now compare these models with specific techniques of geographical weighting that take non-stationarity and spatial proximities within the sample into account (cf. Anselin, 2001). On the basis of the standard OLS model, spatial dependence can be incorporated in two distinct ways. The first method is to estimate an additional parameter representing the distribution of the spatially lagged dependent variable (for each case this is the weighted average across neighbors). This approach is referred to as *spatial lag* model (or *spatial* or *simultaneous autoregressive model* – SAR). This model is appropriate when the focus of interest is on the existence and strength of spatial interactions among the units of observation. Using this technique, we get an impression of how the distribution of the dependent variable across the cases affects their values. In our example, the corresponding question of interest is how individual income spatially correlates with the level of income across proximate cases. This approach is directly related to the issue of the spatial autocorrelation of the dependent variable.

Alternatively, we can control for spatial dependence in the regression disturbance term. Models of this kind are called *spatial error* models. They are appropriate when the aim is to correct for potential bias due to spatial autocorrelation. In this regard, spatial error models pursue similar objectives as models with clustered standard errors or random-intercept multi-level models: They adjust the computed error terms for potential clustering in the data structure. In spatial models, proximity between cases is operationalized as a function of the geographical distance between them.

Spatial error and spatial lag approaches can be applied simultaneously. The notation for combined spatial lag and spatial error models (SARAR or SARSAR models) is as follows (cf. Drukker, Prucha, & Raciborski, 2013, pp. 222-223):

$$\mathbf{y} = \lambda \mathbf{W}\mathbf{y} + \mathbf{X}\beta + \mathbf{u} \tag{4}$$

$$\mathbf{u} = \rho \mathbf{M}\mathbf{u} + \boldsymbol{\epsilon} \quad . \tag{5}$$

In these equations $\mathbf{y}$ is the $N \times 1$ vector of observations on the dependent variable, $\mathbf{X}$ is the $N \times k$ matrix of observations on the independent variables; $\mathbf{W}$ and $\mathbf{M}$ are $N \times N$ spatial-weighting matrices; $\mathbf{W}\mathbf{y}$ and $\mathbf{M}\mathbf{u}$ are $N \times 1$ vectors referred to as spatial lags, and λ and ρ are the corresponding scalar parameters; $\mathbf{u}$ represents an $N \times 1$ disturbance vector; and $\boldsymbol{\epsilon}$ is an $N \times 1$ vector of independent and identically distributed innovations (see also Ord, 1975). In line with the assumptions on the spatial disturbance of $\mathbf{y}$ and $\boldsymbol{\epsilon}$, $\mathbf{W}$ and $\mathbf{M}$ can be

based on different weights, but in the following analyses we will use identical weights.

We approximate the geographical position of the individuals with available information about the regional units in which the individuals are located. We exemplify this by comparing approximation strategies of using large units, such as PRs, and relatively small units, such as ZIP codes. We already obtained rather different results when assessing the extent of spatial autocorrelation in the dependent variable using the different regional levels of approximation (cf. Figure 6).

6.2 Selected empirical results

OLS regression. The empirical results for OLS can be seen in Table 1, model "OLS". For reasons of clarity, we report in this section only coefficients and corresponding standard errors for the central independent variables years of education, sex, and migration background. The full model is presented in Table A2 in the appendix. Applying the OLS model, we observe significant coefficients for years of education as well as for sex. Years of education has a highly significant positive impact on earned income (with a coefficient of 0.083), whereas being female has a negative significant effect of -0.161. Migration background has a small negative and non-significant effect.

OLS with clustered standard errors. In Table 1, models under "OLS (clustered Std. Err.)", we define clusters on various levels of aggregation. For the sake of clarity, we restrict the analyses to two substantive alternatives: clustering on the level of PRs (regional planning units) and ZIP-code areas. A special alternative is clustering by primary sampling units (PSUs). Clustering can be achieved either as part of the complex sample description (`svy`) or as an option in the regression syntax. However, it is difficult to interpret the results of PSU-based clustering in substantive terms: First, as properties of the sample PSUs reflect only one aspect of spatial clustering ("clustering by design") and they may not correspond to a particular level of aggregation specified on substantive grounds; second, PSUs represent different levels of aggregation in different subsamples. Cluster-robust estimators for standard errors have, per definition, no impact on point estimates. In terms of standard errors, we observe slightly higher values than with OLS. However, they decrease when clustering on small-scale levels. In all cases, the estimator is obviously not more efficient than in the standard OLS. Increasing standard errors can be interpreted as indicators for "multiple information" in the original OLS regression (Moulton, 1990).

Multi-level analyses. In Table 1, models under "Random-intercept models", we see that the decision about which regional level is used as level two has crucial consequences for the results. With respect to migration background, the effect is negative and significant on the level of PRs but it becomes weaker with lower levels of geographical aggregation and statistical significance disappears. A simple interpretation of the results from these random-intercept models is that there is an effect of migration background that is suppressed in the OLS model because immigrants are systematically distributed among regions with specific economic situations and income potentials. This compositional effect is controlled by the random intercept. This interpretation would stress that there are major differences in the composition of native Germans and immigrants among regions. Immigrants are concentrated in regions with higher average incomes (in West Germany). Native Germans and immigrants are, on average, much more similar with respect to their income when compared within their immediate environments (ZIP-code units) than within large PRs. The results indicate that random-intercept models are sensitive to the modifiable areal unit problem.

The effect of sex increases (in absolute terms) with a lower level of aggregation, but it is still highly significant in all cases. Compared to this, the effect of education remains almost constant and is comparable to the result obtained from OLS. This effect is very stable; the impact of education is sensitive neither to the applied method nor to the level of regional clustering.

Computing robust standard errors does not distinguish between model error on individual and context levels. In contrast, multi-level modeling is based on the logic of variance decomposition and offers the possibility to assess the extent of heterogeneity on different levels (Snijders & Bosker, 2012). For the computed models, we obtain significant variance components for the context as well as individual levels. The explanatory power of the context increases with a smaller level of geographical aggregation. While the income differences between PRs explain around 11.3 percent of the total variance, this share increases to 13.8 percent when controlling for the affiliation with particular ZIP-code areas. Hierarchical models may also include more than one level of regional aggregation.

Spatial regression models. First, we apply a standard spatial lag and spatial error model with inverse distances as elements of the weighting matrix (**W** and **M** as defined above). Geographical positions are approximated by the geographical centroid of the region. In line with the results of the Moran scatter plots (Figure 6), wherein we used the same weighting matrix, the spatial lag term λ is significant on the level of PRs but not on the level of ZIP-code areas. λ represents the level of spatial autocorrelation of income. In contrast, the spatial error term ρ is significant regardless of the regional level that has been used for approximating the geographical position of the individual cases. This finding can be interpreted as unobserved spatial clustering that we have adjusted for by applying a spatial regression model. As we can see in Table 1, first model with "Geo-centered allocation", the consequences of this model setup for the estimated

Table 1
Model comparisons: Determinants of earned income.

	OLS and multi-level models						Spatial lag and spatial error models[a]							
		OLS (clustered Std. Err.)			Random-intercept models		Geo-centered allocation[b]		Random allocation[b,c]		Geo-centered allocation[d]		Geo-centered allocation[e]	
Cluster level:	OLS	PR	ZIP Code	PSU	PR	ZIP Code	PR	ZIP Code	PR	ZIP Code	PR	ZIP Code	PR	ZIP Code
Years of education	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***	0.08***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Sex(female)	−0.16***	−0.16***	−0.16***	−0.16***	−0.14***	−0.16***	−0.13***	−0.16***	−0.16***	−0.16***	−0.16***	−0.16***	−0.15***	−0.15***
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
Migration Background	−0.02	−0.02	−0.02	−0.02	−0.09***	−0.03	−0.08***	−0.03	−0.02	0.02	−0.05***	−0.03	−0.09***	−0.08***
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
R^2	0.42	0.42	0.42	0.42	-	-	-	-	-	-	-	-	-	-
Var($\beta_0 + \zeta_{0j}$)	-	-	-	-	0.02	0.02	-	-	-	-	-	-	-	-
Var(ϵ_{ij})	-	-	-	-	0.14	0.14	-	-	-	-	-	-	-	-
Log Likelihood	-	-	-	-	-2680	-2868	-	-	-	-	-	-	-	-
λ	-	-	-	-	-	-	0.02***	0.00	0.01	0.00	0.03***	0.00	0.00***	0.00***
ρ	-	-	-	-	-	-	0.06***	0.08***	0.11***	0.07***	0.07***	0.13***	0.01***	0.01***
GMM Criterion	-	-	-	-	-	-	0.01	0.00	0.01	0.00	0.01	0.00	0.01	0.01
N	5,832	5,832	5,832	5,832	5,832	5,832	5,832	5,832	5,832	5,832	5,832	5,832	5,832	5,832

Standard errors in parentheses. Dependent variable: gross hourly wage (log), further control variables in the model; see Table A2. Data: GSOEP, own calculations.
[a] Each with a specific distance (weights) matrix and approximation of the location (allocation). [b] Weights: 1/distance. [c] Assessed on the basis of multiple runs.
[d] Weights: $\exp(-\text{distance}^2/2)$. [e] Weights: $\exp(-\text{distance}^2/20,000)$.
* $p < .05$ ** $p < .01$ *** $p < .001$

coefficients of education and sex are minimal. However, the effect of migration background is very sensitive to the applied operationalization. There is a negative effect when using PRs but no significant effect when using the ZIP-code level as an approximation. This result appears to be a validation of the results provided by multi-level modeling (see Table 1, "Random-intercept models"). In fact, a closer inspection reveals that the specification is almost an exact reproduction of a random-intercept model or simple clustering. As mentioned before, when using an inverse distance function, cases within a very close range get very high weights. Modeled geographical locations of individuals in the same polygons are (almost) identical when using geo-centered allocation. their mutual importance for the adjustment of the spatial lag is very high whereas cases within any longer distance are effectively irrelevant.

To illustrate the relevance of allocation we vary the mode of allocation. Now cases are allocated randomly within the regional units (cf. again Figure 2 (c)). The results are shown in Table 1, under "Random alloction". Inhabitants of the same territorial unit are now more distant from each other, and there is no longer an extraordinarily high mutual interdependence between them. As a consequence, cases from neighboring regions become relatively more relevant. We find the effect of migration background to be no longer significant with this mode of allocation.

Finally, we modify the weighting distance function from an inverse distance function to a Gaussian (normal) function (with varying dispersion). For a comparison of different functional forms see Figure 3; a larger dispersion means that the weights decrease less strongly with distance. Our empirical findings on different function forms can be found in Table 1. The first four spatial regression models are based on an inverse distance function while the last four models under last two headings of "Geo-centered allocation" use Gaussian-type functions with a relatively small and a relatively large dispersion respectively. As we can see in both models, migration background again has a significant negative effect on the level of PRs. When using ZIP-code areas as the level of approximation this also applies to the last model. As already mentioned, immigrants and native Germans in the same ZIP-code areas are likely rather similar in terms of earned income. Disproportionately high weights for immediate neighbors resulting from an inverse distance function (or a Gaussian function with a small dispersion) eliminate the negative effect observable on a larger scale. By taking more distant cases more strongly into account, this effect is recovered.

7 Regression-based analyses (geographically varying effects)

In our final analyses, we return to aspects of "place" by looking at the geographical variation in the determinants of earned income. An important issue that can be attributed to specific regional positions is the spatial non-stationarity of effects (Brunsdon, Fotheringham, & Charlton, 1996). When applying a linear regression model to spatial data, we assume a stationary process, which means that the same characteristic has the same impact everywhere. If there is spatial non-stationarity, such a global model cannot appropriately explain the relationships between the sets of relevant variables. The problem is not an incorrect estimation of a global indicator, but the fact that a single global indicator does not adequately represent an effect that varies across space.

7.1 Methods

Random-slope model. One possibility that allows considering the variation of effects among clusters is a multilevel model in the form of a random-slope model. Random-slope models do not consider spatial information but account for the clustering of observations in particular (territorial) units. We can obtain an additional parameter for the considered independent variable that represents the variation of the correspondent linear coefficient among regions. The random-slope model has this specification (see also Rabe-Hesketh & Skrondal, 2012, p. 188):

$$y_{ij} = (\beta_0 + \zeta_{0j}) + (\beta_1 + \zeta_{1j})x_{1ij} + \ldots + \beta_n x_{nij} + \epsilon_{ij} \qquad (6)$$

for individual i in region j, with $\zeta_{0j} \sim N\left(0, \tau_0^2\right)$ and $\zeta_{1j} \sim N\left(0, \tau_1^2\right)$.

Geographically weighted regression. Another possibility to account for the spatial non-stationarity of effects by explicitly considering the spatial proximity of all cases is geographically weighted regression (GWR; cf. Brunsdon et al., 1996; Fotheringham, 2009a). GWR follows a logic that is slightly different from the spatial lag approach; in this model framework, there are no additional parameters that account for spatial dependency in estimated global measures, but separate linear models for every given geographical position in the dataset are calculated. For each linear regression, a separate set of coefficients is obtained. Thereby, every single local regression makes use of the entire dataset. For each of these regressions, all cases are weighted depending on their distance from a particular geographic position. Various weighting schemes can be used to specify the relevance of distance and proximity; but in most cases, a Gaussian function is used. This implies that the importance of neighboring cases decreases with a varying slope across space.

Following Fotheringham (2009a, p. 244), the GWR model can be specified as follows:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \ldots + \beta_{nj}x_{nij} + \epsilon_{ij} \qquad (7)$$

for individual i in region j, with parameter estimator

$$\beta'_j = \left(\mathbf{X}^T\mathbf{W}_j\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}_j\mathbf{Y} \qquad (8)$$

(a) Years of education (b) Sex (female) (c) Migration background

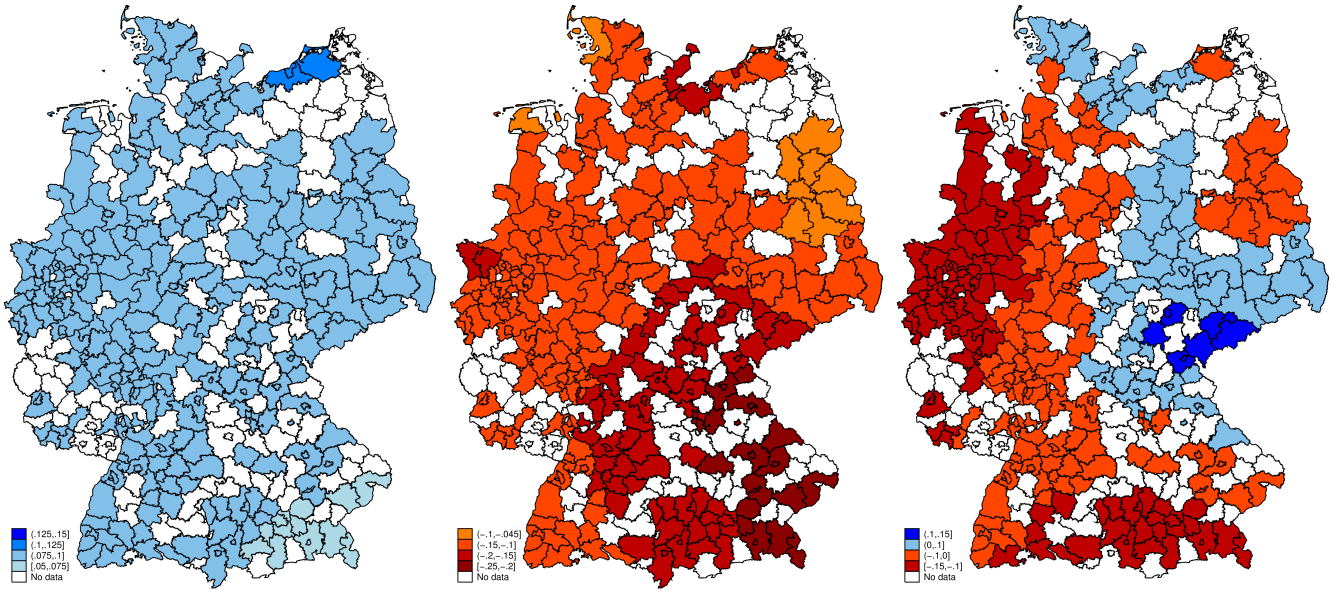


Figure 7. Geographically varying effects of education (a), sex (b) and migration (c) background on individual earnings: Results from GWR. Regional classification: admin. districts (NUTS-3). Data: GSOEP; BKG, 2016 (shapefile), own calculations.

where $\mathbf{W}_j$ is a matrix of weights specific to location j.

For the definition of the weighting function, a bandwidth – i.e., the standard deviation of the Gaussian distribution – has to be specified. The bandwidth can be set to the maximal distance between given positions in the dataset. This might not, however, be the best choice because GWR results are very sensitive to the chosen weighting function; a bandwidth that is too broad – e.g., due to outliers – will lead to a large bias in the local estimates. Therefore, most statistical packages implement a model calibration procedure whereby the bandwidth is derived by excluding all distant cases until an "optimal" function is obtained. This procedure can be carried out by minimizing a cross-validation score or via the Akaike Information Criterion (AIC). In a further step, adaptive functions can be computed, meaning that weighting functions are allowed to vary between geographical positions. GWR techniques are predominantly descriptive, but there are several approaches based upon GWR that allow statistical testing for non-stationarity of OLS model coefficients (Leung, Mei, & Zhang, 2000).

7.2 Selected empirical results

Random-slope model. When computing random-slope parameters for our example, neither the effect of years of education nor of sex or migration background vary between regions; the parameters that indicate the level of variation in the effects of sex, migration and education are not significant. However, when computing random slopes, we face the same restrictions as in the case of random-intercept models: The random-slope follows a particular distributive assumption, and proximity between individuals is only considered when they share the same region – even if the theoretically assumed connection between them is not limited by these geographical borders.

At this point we are not interested in average effects but in their heterogeneity between regions, so the coefficients are not reported.

Geographically weighted regression. To illustrate GWR, we compute examples using a Gaussian distance function with a fixed bandwidth that was obtained by a cross-validation algorithm. We use geographical centroids of administrative districts to approximate the geographical position of individuals. To obtain detailed results, researchers use the smallest level of aggregation available. However, a meaningful interpretation of GWR results demands their depiction in the form of maps. Due to strict data protection guidelines on the use of many large-scale surveys, graphical mapping of the results is not permitted if data cells are occupied by too few observations, which is often the case when using small levels of aggregation. The test for non-stationarity delivers negative results for most variables, including sex and years of education, but the effect of migration background varies significantly across space. This finding is in contrast to the conclusion made on the basis of multi-level modeling. As individuals sharing the same aggregate unit have been assigned to exactly the same position, GWR produces identical results for cases within the same aggregate unit. Using a GIS mapping tool, the results can be presented graphically. Figure 7

shows the regional distribution of the relevant coefficients.

Following the effective data protection rules, we cannot consider regional units where the number of observations is below the limit of ten individual cases. This also means that maps like this cannot be produced using a finer gradation than the level of administrative districts.

We find that (only) the effect of migration background on earnings varies markedly by region. We can also observe regional patterns: For example, immigrants in East Germany earn even more than native German employees when controlling for relevant characteristics. The immigrant population in East Germany is relatively small, so this result is consistent with established findings of a negative impact of minority group size (Granato, 2009). On the other hand, this local effect does not neutralize the overall average negative effect for immigrants that we found in a number of models. GWR does not quantify the importance of particular regions for the global estimate. Furthermore, the positive test for spatial non-stationarity does not imply that the effects at every particular location are significant. For such an examination, additional local statistics can be computed. This option is not provided by many GIS packages, but for example by GWR4 (Charlton, Fotheringham, & Brunsdon, 2006; Nakaya, 2016).

8 Summary and conclusions

In this paper, we have presented illustrative analyses of large-scale socio-economic survey data that carry underlying spatial structures. Our fundamental assumption has been that individual actions necessarily have spatial references. Individuals act in the physical world, and their relative geographical position may have an impact on the processes under consideration. Even if potential spatial interdependencies are not explicitly considered, analytical models are still based implicitly on assumptions about the geographical distribution of the cases, their relevance for the measured effects, and the quality of the models. Our example has been determinants of earned income. Comparing a number of analyses, we have found both robust effects, as in the case of education, and effects that are very sensitive to the specific operationalization, as in the case of sex and, in particular, migration background.

Which general conclusions can be drawn from these examples beyond the specificities of the particular substantive questions? Large-scale survey data do not often allow for fine-grained geographical differentiation, but researchers should in any case consider the level of geographical detail in the available data. It is obvious that analyses which are based on data that allows only the calculation of global averages are of limited value when systematic spatial variation of the particular phenomenon can be expected. One should also acknowledge, however, that local composition differences, spatial heterogeneity, and spatial dependencies are universal problems that may influence the results even if the research question does not explicitly focus on spatial phenomena. Even in these cases, including geographical information may prevent inconclusive substantive results. Another popular research interest that requires geographical information is concerned with the effects of local context conditions represented by aggregate information that is matched to survey data.

Therefore, at least proxy information about local positioning and clustering should be obtained. For determining the *local position* of cases, information should preferably be based on a classification system with a small scale of regional aggregation so that the approximation gets close to the exact position. In any case, the goal is to approximate the empirical patterns of distribution. The situation is somewhat different when deciding the aggregation level of context information and the level of regional clustering in multi-level analyses of context effects. In these cases, the challenge is to find the level of aggregation that is substantively most *relevant* for the specific topic (cf. Hillmert, 2016).

The choice of adequate analytical techniques depends on the studied phenomenon and the type of geographical information available to the researcher. *OLS* is the non-spatial standard approach for determining statistical associations in many applications. It can also be used when any geographical information is lacking. However, it assumes homogeneity of effects and the independence of observations – assumptions which are often not justified.

Multi-level models can account for dependencies due to spatial clustering if these dependencies work predominantly *within* the specified geographical units. The models do not require information about the exact localization of the regional clusters. The main challenge is choosing the optimal level of spatial clustering. In this case, dependencies are located within the clusters and spatial dependencies beyond the cluster are negligible. However, multi-level models neglect the potential effects of geographical units that surround the unit in which an individual is located. The same applies to the use of clustered standard errors. The use of small-scale aggregate units allows goods estimates of the location of cases, but the cluster may be "too small" for effectively controlling the relevant spatial dependencies. An alternative might be the inclusion of more than one aggregate level. However, a general restriction of multi-level models is that the assumption of hierarchical structures of clusters –or comparatively simple cross-classifications – does not allow considering (complex) dependencies beyond the defined clusters.

Spatial models are in many instances the most adequate way of dealing with spatial dependencies, as they allow considering multiple (matrix-like) dependencies. Still, they require a number of important decisions, particularly about distance-based weighting, and they may be rather sensitive to specifications. The applied weighting scheme should be

based on solid theoretical arguments or empirical evidence on the mutual relevance of cases in a defined range. In the likely case that only approximate information about the location of individual cases is available, the additional challenge is to make a valid assumption about the distribution of cases within spatial units, particularly if these territorial units are relatively large. This distribution should resemble the likely empirical pattern. Hence, using spatial modeling does not produce better estimates per se. Researchers should be aware of the theoretical assumptions and specific settings that may have direct implications for the results.

Many macro-level applications – particularly in the academic fields of geography or political science – focus on interdependencies between regions or countries (e.g. Beck, Gleditsch, & Beardsley, 2006). However, such analyses represent only a rather small proportion of all survey applications. More widespread potential problems may therefore apply for the larger proportion of research that is not explicitly spatial. Given the universality of clustering in the social world, spatial dependencies may still affect the validity of the results of many analyses, regardless of whether they are part of the research question or not. The aim of our paper has been to increase awareness of this set of problems among researchers, including those outside corresponding research areas such as human geography and spatial economics.

9 Acknowledgements

Research for this paper was supported by a grant from the German Research Foundation (DFG), Grant HI 767/7-2. Part of the paper was written while the first author held a BIGSSS/HWK Fellowship at the HWK Institute for Advanced Study, Delmenhorst, and the University of Bremen.

References

Ammermüller, A. & Weber, A. (2005). *Educational attainment and returns to education in Germany: an analysis by subject of degree, gender and region.* Discussion paper No. 05-17. Mannheim: Mannheimer Zentrum für Europäische Wirtschaftsforschung (ZEW).

Anselin, L. (2001). Spatial econometrics. In B. H. Baltagi (Ed.), *A companion to theoretical econometrics* (pp. 310–330). Malden, MA: Blackwell.

Beck, N., Gleditsch, K. S., & Beardsley, K. (2006). Space is more than geography: using spatial econometrics in the study of political economy. *International Studies Quarterly*, *50*(1), 27–44.

Becker, G. S. (1993). *Human capital. A theoretical and empirical analysis, with special reference to education* (3rd ed.). Chicago: The University of Chicago Press.

Belotti, F., Hughes, G., & Piano Mortari, A. (2016). Spatial panel data models using Stata. CEIS Working Paper No. 373. Retrieved from https://ssrn.com/abstract=2754703

Bishu, S. & Alkadry, A. (2017). A systematic review of the gender pay gap and factors that predict it. *Administration & Society. 49*(1), 65–140.

BKG. (2016). Administrative areas 1: 250.000, VG250 and VG250-EW. Federal Agency for Cartography and Geodesy. Retrieved from http://www.geodatenzentrum.de

Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploringspatial nonstationarity. *Geographical Analysis*, *28*(4), 281–298.

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics: volume 3* (pp. 1801–1863). Amsterdam: Elsevier.

Charlton, M., Fotheringham, A. S., & Brunsdon, C. (2006). *Geographically weighted regression.* NCRM Methods Review Papers No. 006. ESRC National Centre for Research Methods.

Cliff, A. & Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis*, *4*(3), 267–284.

Constant, A. & Massey, D. (2005). Labor market segmentation and the earnings of German guestworkers. *Population Research and Policy Review*, *24*(5), 489–512.

Crow, K. (2006). SHP2DTA: Stata module to convert shape boundary files to Stata datasets. Statistical Software Components, Boston College Department of Economics. Retrieved from https://ideas.repec.org/c/boc/bocode/s456718.html

Delmelle, E. (2009). Spatial sampling. In A. S. Fotheringham & P. Rogerson (Eds.), *The SAGE handbook of spatial analysis* (pp. 183–206). Los Angeles, London: Sage.

Deutsche Post. (2003). 10 Jahre fünfstellige Postleitzahl. Retrieved from http://www.dpdhl.com/de/presse/pressemitteilungen/2003/10_jahre_fuenfstellige_postleitzahl.html

DIW. (2016). Regional data in the SOEP data set. Retrieved from https://www.diw.de/en/diw_02.c.222519.en/regional_data.html

Drukker, D. M., Peng, H., Prucha, I. R., & Raciborski, R. (2013). Creating and managing spatial-weighting matrices with the spmat command. *The Stata Journal*, *13*(2), 242–286.

Drukker, D. M., Prucha, I. R., & Raciborski, R. (2013). Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances. *The Stata Journal*, *13*(2), 221–241.

Dustmann, C. & Glitz, A. (2005). *Immigration, jobs and wages: theory, evidence and opinion.* London: Centre for Economic Policy Research.

Eurostat. (2013). About the TERCET NUTS-postal codes matching tables. Retrieved from http://ec.europa.eu/eurostat/tercet/flatfiles.do

Fitzenberger, B. & Wunderlich, G. (2002). Gender wage differences in West Germany. A cohort analysis. *German Economic Review*, *3*(4), 379–414.

Fortin, M.-J. & Dale, M. R. T. (2009). Spatial autocorrelation. In A. S. Fotheringham & P. Rogerson (Eds.), *The SAGE handbook of spatial analysis* (pp. 89–104). Los Angeles, London: Sage.

Fossen, F. M. & Büttner, T. (2013). The returns to education for opportunity entrepreneurs, necessity entrepreneurs, and paid employees. *Economics of Education Review*, *37*, 66–84.

Fotheringham, A. S. (2009a). Geographically weighted regression. In A. S. Fotheringham & P. Rogerson (Eds.), *The SAGE handbook of spatial analysis* (pp. 243–253). Los Angeles, London: Sage.

Fotheringham, A. S. (2009b). The problem of spatial autocorrelation and local spatial patterns. *Geographical Analysis*, *41*, 398–403.

Fotheringham, A. S. & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning*, *23*(7), 1025–1044.

Frick, J. R., Jenkins, S. P., Lillard, D. R., Lipps, O., & Wooden, M. (2008). Die internationale Einbettung des Sozio-oekonomischen Panels (SOEP) im Rahmen des Cross-National Equivalent File (CNEF). *Vierteljahrshefte zur Wirtschaftsforschung*, *77*(3), 110–129.

Gebel, M. & Pfeiffer, F. (2007). *Educational expansion and its heterogeneous returns for wage workers.* Working paper No. 07-10. ZEW Mannheimer Zentrum für Europäische Wirtschaftsforschung (ZEW).

Goebel, J., Krause, P., Pischner, R., Sieber, I., & Wagner, G. G. (2008). Daten- und Datenbankstruktur der Längsschnittstudie Sozio-oekonomisches Panel (SOEP). SOEP paper No. 89. Retrieved from http://www.diw.de/documents/publikationen/73/diw_01.c.79473.de/diw_sp0089.pdf

Granato, N. (2009). Effekte der Gruppengröße auf die Arbeitsmarktintegration von Migranten. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *61*(3), 387–409.

Griffith, D. A. & Arbia, G. (2010). Detecting negative spatial autocorrelation in georeferenced random variables. *International Journal of Geographical Information Science*, *24*(3), 417–437.

Harmon, C. & Walker, I. (2000). The returns to the quantity and quality of education: evidence for men in England and Wales. *Economica*, *67*(265), 19–35.

Heeringa, S., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis.* Boca Raton: Chapman & Hall/CRC.

Hillmert, S. (2016). *Conceptualizing relevant areas for spatial context analyses.* Paper presented at Nuffield College, University of Oxford.

Kwan, M.-P. (2012). The uncertain geographic context problem. *Annals of the Association of American Geographers*, *5*(102), 958–968.

Leung, Y., Mei, C.-L., & Zhang, W.-X. (2000). Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning*, *32*(1), 9–32.

Logan, J. R. (2012). Making a place for space: spatial thinking in social science. *Annual Review of Sociology*, *38*, 507–524.

Machin, S. & Puhani, P. (2003). Subject of degree and the gender wage differential: evidence from the UK and Germany. *Economics Letters*, *79*(3), 393–400.

Mercer, L., Wakefield, J., Chen, C., & Lumley, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics*, *8*, 69–85.

Mincer, J. (1974). *Progress in human capital analyses of the distribution of earnings.* NBER Working Paper No. 53, Stanford, CA.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1–2), 17–23.

Moulton, B. R. (1990). An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Review of Economics and Statistics*, *72*(2), 334–338.

Nakaya, T. (2016). GWR4.09 User Manual. GWR4 Windows application for geographically weighted regression modelling. Retrieved from http://geoinformatics.wp.st-andrews.ac.uk/download/software/GWR4manual.pdf

Nielsen, H. S., Rosholm, M., Smith, N., & Husted, L. (2001). *Qualifications, discrimination, or assimilation? An extended framework for analysing immigrant wage gaps.* Discussion Paper No. 365. Bonn: Institute for the Study of Labor (IZA).

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, *70*(349), 120–126.

Pearce, M. S. (1998). Geographically weighted regression. *Stata Technical Bulletin*, *46*, 20–23.

Pisati, M. (2007). SPMAP: Stata module to visualize spatial data. Statistical Software Components, Boston College Department of Economics. Retrieved from https://ideas.repec.org/c/boc/bocode/s456812.html

Primo, D. M., Jacobsmeier, M. L., & Milyo, J. (2007). Estimating the impact of state policies and institutions with mixed-level data. *State Politics & Policy Quarterly*, *7*(4), 446–459.

Rabe-Hesketh, S. & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata.* Volume I: Continuous responses. College Station, TX: Stata Press.

Rogers, W. (1994). Regression standard errors in clustered samples. *Stata Technical Bulletin*, *3*(13), 19–23.

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* (2nd ed.). Los Angeles: Sage.

Socio-Economic Panel (SOEP). (2014). Data for the years 1984-2013, version 30. doi:10.5684/soep.v30.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355–374.

Spieß, M. (2008). Gewichtung und Hochrechnung mit dem SOEP. Retrieved from https://www.diw.de/documents/dokumentenarchiv/17/diw_01.c.79953.de/soep_gewichtungsvortrag2008.pdf.

Statistisches Bundesamt. (2016). Gemeindeverzeichnis-Informationssystem GV-ISys. Retrieved from https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/Gemeindeverzeichnis.html

Strauß, S. & Hillmert, S. (2011). Einkommenseinbußen durch Arbeitslosigkeit in Deutschland: Alters- und geschlechtsspezifische Differenzen im Vergleich. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *36*(4), 567–594.

Ward, M. D. & Gleditsch, K. S. (2008). *Spatial regression models.* Thousand Oaks, CA: Sage.

Weßling, K., Hartung, A., & Hillmert, S. (2015). Spatial structure counts: the relevance of regional labour-market conditions for educational transitions to vocational training. *Empirical Research in Vocational Education and Training*, *7*(12), 1–20.

Appendix
Tables

Table A1
Variable descriptions

Variable	Percentage/Mean	Std.dev.	Min	Max	N
Dependent variable					
Inflation-adjusted gross hourly wage (log)	2.574	0.524	−1.025	6.058	5,832
Independent variables					
Years of Education	12.56	2.63	7	18	5,832
Sex					5,832
Female	49.66	-	-	-	-
Male	50.34	-	-	-	-
Migration background					5,832
Native	87.52	-	-	-	-
Migration background	12.48	-	-	-	-
Age (in years)	42.75	6.29	35	55	5,832
Marital status					5,832
Married	25.31	-	-	-	-
Unmarried	74.69	-	-	-	-
Employment status					5,832
Full-time	72.89	-	-	-	-
Part-time	27.11	-	-	-	-
Duration of job tenure (in years)	10.19	9.05	0	40.8	5,832
Firm size					5,832
<20 employees	23.13	-	-	-	-
20–200 employees	30.57	-	-	-	-
200–2,000 employees	22.43	-	-	-	-
>2,000 employees	23.87	-	-	-	-
Duration of previous unemployment (in years)	0.61	1.64	0	21.2	5,832
Industry/Sector					5,832
Primary	3.60	-	-	-	-
Secondary	31.28	-	-	-	-
Tertiary	65.12	-	-	-	-

Data: GSOEP, own calculations

Table A2
Full OLS model

Independent variable	Coef.	Std. Err.
Years of education	0.830***	0.002
Sex: female (Ref. male)	−0.161***	0.013
Migration background: migrant (Ref. non-migrant)	−0.018	0.016
Age	−0.002**	0.001
Marital status: married (Ref. not married)	0.063***	0.012
Employment status: full-time (Ref. part-time)	0.063***	0.012
Firm size: < = 20 employees	−0.254***	0.018
Firm size: 21–200 employees	−0.053***	0.013
Firm size: 201–2,000 employees (Ref. > 2,000 employees)	0.019	0.014
Job tenure in years	0.012***	0.001
Previous unemployment in years	−0.051***	0.003
Sector: Primary	−0.200***	0.029
Sector: Tertiary (Ref. Secondary)	−0.064***	0.013
Constant	1.609***	0.049
R^2	0.42	
N	5,832	

Data: GSOEP, own calculations. Dependent variable: gross hourly wage (log).
* $p < .05$ ** $p < .01$ *** $p < .001$

Table A3
Geographical information in major European surveys

NUTS 1	NUTS 2	NUTS 3	LAU 1	LAU 2	Smaller-scale levels
German Socio-Economic Panel (Germany)[a]					
Federal States: SUF (1984–)	Governmental districts:*	Administrative districts: Remote Access (1985–)	Collective municipalities:*	Municipalities: Secure Data Center (1993–)	ZIP codes: Secure Data Center (2000–)
*National Educational Panel Study, Starting Cohort 6*** *(Germany)*[b]					
Federal States: SUF	Governmental districts:*	Administrative districts: Remote Access	Collective municipalities: n/a	Municipalities: n/a	ZIP codes: n/a
DNB Household Panel (Netherlands)[c]					
Lands: SUF (1993–)	Provinces: SUF (1993–)	COROP regions: On demand (1993–; $n > 20$ per unit)	COROP regions: (same as NUTS 3)	Municipalities: On demand (1993–; $n > 20$ per unit)	n/a
Swiss Household Panel (Switzerland)[d]					
(Switzerland as a whole)	Regions: SUF (1999–)	Cantons: SUF (1999–)	Districts: n/a	Municipalities: n/a	n/a
British Household Panel Study & Understanding Society (United Kingdom)[e]					
Statistical regions: SUF (1991–)	(Groups of) counties/districts/ unitary authorities: SUF (1991–)***	(Groups of) counties/districts/ unitary authorities: SUF (1991–)***	Lower-tier authorities: (1991–)****	Wards: (1991–)****	ZIP codes: Secure Data Center (1991–)
European Labor Force Survey (EU)[f]					
SUF (1983–)*****	SUF (1983–)*****	n/a	n/a	n/a	n/a
Statistics on Income and Living Conditions (EU)[g]					
SUF (2004, 2008, 2012)	SUF (2004, 2008, 2012)*****	SUF (2004, 2008, 2012)*****	n/a	n/a	n/a

Individual arrangements for matching context information on specific levels of aggregation may be available upon request.
* Can be aggregated from lower level ** Retrospective data *** Special license **** Direct access for UK bodies or on-site ***** Availability depends on country
[a] https://www.diw.de/en/diw_02.c.222519.en/regional_data.html
[b] https://www.neps-data.de/de-de/datenzentrum/%C3%BCbersichtenundhilfen/matchingvonregionaldaten.aspx
[c] https://www.dnb.nl/binaries/DNB_OS_1004_BIN_WEB_tcm46-277691.pdf
[d] http://forscenter.ch/de/our-surveys/swiss-household-panel/dokumentationfaq-2/fragebogen-als-pdf-2/ [e] https://www.understandingsociety.ac.uk/about/data-linkage#part2
[f] http://ec.europa.eu/eurostat/documents/1978984/6037342/EULFS-Database-UserGuide.pdf [g] http://ec.europa.eu/eurostat/web/income-and-living-conditions/overview