Essays in Regional and Migration Economics

Von der Mercator School of Management, Fakultät für Betriebswirtschaftslehre, der

Universität Duisburg-Essen

zur Erlangung des akademischen Grades

eines Doktors der Wirtschaftswissenschaft (Dr. rer. oec.)

genehmigte Dissertation

von

Philipp Markus

aus

Iserlohn

Referent: Prof. Dr. Tobias Seidel

Korreferentin: Prof. Dr. Nadine Riedel

Tag der mündlichen Prüfung: 9. August 2023

Acknowledgment

First, I want to express my gratitude to my first supervisor Tobias Seidel who helped me in many different ways with my first steps in academic research. This thesis would not have been possible without his dedication as an advisor. I also thank Nadine Riedel, Arnaud Chevalier, and Martin Karlsson for countless discussions and helpful comments about my research projects. Especially, I want to express my gratitude to Martin and Therese Nilsson for making my research stay in Lund (Sweden) possible.

I also thank Ana Rodríguez-González, Martin Fischer, Matthew Collins, Matthias Westphal, Karolin Süß, Maren Kaliske, Malte Borghorst, Siegfried To, Theresa Markefke, Eyayaw Beze, Fabian Bald, Lu Wei, Nina Furbach, Avtandil Abashishvili, Johannes Gallé, Lea Nassal, Lorenz Gschwent, Jan Wickerath, Alena Chalupka, all members of the Research Training Group (RTG) Regional Disparities & Economic Policy, the Ruhr Graduate School in Economics (RGS Econ) and the colleagues in Duisburg and Lund for helpful comments and fruitful discussions. I also thank Christian Göb for his excellent research assistance.

I thank the Centre for Economic Demography (CED) in Lund for generously providing the data. Financial support from the German Research Foundation (DFG) via the RTG Regional Disparities & Economic Policy and the RGS Econ is gratefully acknowledged.

Finally, I would like to point out how grateful I am for the constant nonscientific support that I experienced in recent years. The contribution of the emotional support of my family and my girlfriend Nicola as well as from many colleagues and friends like Nicole, Jan-Patrick, Theresa, Tobi, Steve, Luisa, Jana, Lisa, Rene, Alena, Sabi, Franzi, Jen, and the 15th cohort of the RGS Econ cannot be overstated.

Contents

List of Figures

List of Tables

Introduction

The place of living plays a role in all important choices throughout the life circle. Access to educational institutions might influence how much schooling we get, we find friends and partners among people around us, local labor market conditions affect our job careers, and things like air quality or the supply of health care in the neighborhood determine how long we live when we are old. It is not exaggerated to say that the characteristics of the region we decide to live in impact overall well-being in countless ways.

The relevance of differences between locations is one of the main reasons why regional and urban economics exists as a field of economics. There, it is common to assume utility to depend on the level of consumption determined by local wages and prices but also on so-called amenities. Amenities are characteristics of a location that, in contrast to regular goods, can not be traded on a market. This can include everything from natural conditions like clean air over public goods provided by the local government, e.g. security, to the size and quality of social ties to friends and family members, depending on what individuals value having in their neighborhood. Researchers have developed concepts to measure the overall importance of these local amenities as well as the willingness to pay for specific characteristics. However, given the broad concept of amenities, there is a large variation in considered indicators which, of course, has substantial implications for the results.

Chapter 1 of my thesis provides a better idea of which characteristics of a location matter by proposing a statistical learning approach to identify the most relevant measures of amenities. This involves three steps. First, I collect over 100 indicators of German counties motivated by previous theoretical and applied academic literature as well as public region and city rankings, including novel proxies of social participation. My dataset

gives future scientists a comprehensive overview of the availability and sources of spatial amenity data in Germany. Second, I use the information on local wages and rent prices to derive a cardinal measure of regional differences in the total value of amenities in German counties, based on the logic that households in regions with a relatively low income net of housing costs have to be compensated by a high level of utility from location characteristics. The resulting ranking allows for comparing regions in terms of livability in a more objective way than public rankings with arbitrary indicators selection and weighting. Third, I apply statistical methods to the set of gathered indicators to identify the best predictors of attractiveness differences between locations. My results indicate that there is no entirely irrelevant category of measures. However, the prediction accuracy does not necessarily increase when including a large number of indicators, giving future researchers a statistical justification to be parsimonious with the number of considered amenities. Namely, the turnout at national elections, the poverty rate, measures of innovativeness, and proxies of gender equality in the labor market are identified as the most relevant indicators to describe spatial differences in livability.

Since the results of Chapter 1 represent purely statistical relationships the remaining Chapters 2 and 3 focus on establishing a causal effect of location characteristics on human behavior. Yet, the interconnectedness of spatial mobility and the place of living with many different aspects of life implies various challenges for identifying causal relationships. Among other things, a causal interaction could exist in both directions. The theoretical framework in Chapter 1 implies that households base migration decisions on a region's suitability for living. However, endogenous location characteristics, for example, the provision of public goods like health care, might in turn depend on the population structure which is heavily dependent on (past) migration flows. Moreover, the first Chapter aims to provide an idea about the importance of location characteristics for a representative household. As suggested already above, there is a notable heterogeneity in the perceived importance of certain conditions. The availability of general practitioners and pharmacies might be significantly more relevant for elderly residents, while the night club density or the access to university education is of crucial importance for young adults. Even if university students exclusively sort

into cities with a higher education institution it is challenging to identify the importance of the provision of tertiary education. This is especially problematic if the presence of universities is correlated with other measures that have a higher priority for the whole population or if the population share of students is negligible.

In Chapter 2, I identify a causal relationship between the characteristics of a region and the decision of individuals where to live with a natural experiment to overcome the highlighted issues of reversed causality, heterogeneity, and correlation with other factors. I exploit the exogenous variation induced by a geographic higher education expansion reform in Sweden to estimate the causal effect of changes in access to tertiary education on the migration decision of high school graduates. Using a two-way-fixed-effect estimator, I find that a new university has stronger effects on short-term mobility than on college participation rates. While local high school graduates were 6.6% more likely to attend college, the propensity to move away in the four years after finishing secondary education decreased by more than 10%. The results confirm that conditions of the place of residence can have a substantial influence on important decisions in life. Some individuals would not have obtained a college degree if they had finished high school without a higher education institution close by. But the negative effect on mobility can have an impact on seemingly unrelated but meaningful decisions at later stages of life as well, as I show in my last Chapter.

In Chapter 3, I show how the choice of a spouse partially depends on the place of residence and mobility. Building on the findings of Chapter 2, I estimate that the higher education expansion reform increases the local high school graduates' likelihood of marrying someone born in the same municipality by 11.62%. In contrast, I find no evidence for an impact on educational homogamy, i.e. marrying someone with the same level of education. This is especially surprising as I show that marital sorting in Sweden is stronger at the upper end of the educational distribution. Nevertheless, the positive impact on college uptake documented in Chapter 2 does not lead to a higher chance of coupling with other college-educated. My results suggest that the correlation between the degree of educational homogamy and years of schooling is partially driven by the larger mobility of highly educated individuals rather than the level of education itself. Or, in the

context of the higher education expansion reform, the lower rate of mobility mitigated the effect on (presumably positive) qualification effect on educational homogamy.

CHAPTER 1

The Determinants of Quality of Life: Measuring Local Amenities

Chapter Abstract

Amenities play an essential role in the perception of regions' liveability in the public and politics as well as in urban economics. However, the selection of measures is arbitrary in the academic literature and popular city rankings. In this paper, I collect a comprehensive list of over 100 amenity indicators and test their explanatory power for quality of life differences between German counties with random forest and Lasso regression methods. My results suggest that a parsimonious model with not more than ten indicators from six categories has a reasonably good prediction quality. The statistically most relevant characteristics are turnout at national elections, the poverty rate, measures of innovativeness, and proxies of gender equality in the labor market.

1.1 Introduction

To understand people's location choices, the economic literature usually assumes the utility of living in a specific region to depend on local wages, the cost of living in this region, and the livability of that location. The latter is typically termed as *quality of life* (QoL) driven by *local amenities*. Local amenities are location-specific non-tradeable goods that have no market and therefore no explicit price like clean air or the proximity to a nice beach. Although results indicate that these amenities play a significant role in the question of why people move between regions and the welfare consequences, the literature does not offer a deep understanding of what is summed up under these umbrella terms. While academic literature from various fields as well as popular region and city rankings suggest a wide range of possibly relevant local amenities, a systematic and general approach is missing.[1] There is neither a commonly agreed list of relevant amenities nor the question of which amenities are relatively more important than others answered.

Spatial equilibrium models based on Rosen (1979) and Roback (1982) are often used to explain people's location choices by assuming a utility function as mentioned above. Under the additional assumption of utility equalization over space in equilibrium due to perfect mobility of individuals, the Rosen-Roback framework is able to provide implicit prices for the sum of every region's amenities. Still, amenities remain a (structural) residual summarizing the effect of all unobservables in almost all prominent applications of this framework in spatial general equilibrium models.[2] This makes it hard to get a better understanding of the mechanisms involving local amenities. Not only the economic interpretation of this factor lacks inevitable information to drive concrete policy recommendations from it. Also, there is the technical problem that having data for a measure of amenities is necessary to endogenize local amenities in such a model as in Diamond

[1] Popular international city rankings include the Mercer Quality of Living City Ranking, The Global Livability Index by the Economist, or the Prognos Zukunftsatlas for German regions.

[2] See for example Allen and Arkolakis (2014); Ahlfeldt et al. (2015) and Ahlfeldt et al. (2020). Redding and Rossi-Hansberg (2017) provides an overview on spatial general equilibrium models.

(2016). But even if amenities are assumed to be exogenous, checking the validity of the model's results often involves an overidentification test on the structural amenity residual. However, there is no extensive statistically validated evidence of what indicators should be used in that case. Therefore, identifying the (statistical) determinants of QoL is not only interesting in itself but also vital to get a better understanding of the role of local amenities in spatial economics in the future.

To overcome that problem, I develop a methodical approach to choosing relevant measures of local amenities based on statistical methods from a set of over 100 indicators. My first contribution is the collection of the data itself. Guided by academic literature as well as lists of potential determinants from public city and region rankings, I collect and generate a large number and variety of indicators that potentially measure the level of local amenities. This includes indicators for natural/environmental factors like weather, the proximity to coasts or air quality, a wide range of public goods like infrastructure, education, health care, or security, characteristics of the housing market that are not (fully) incorporated in house prices or rents, features of the local labor market and the local economy that are not (fully) incorporated in wages, demographic indicators including migration and fertility, leisure and consumption opportunities like the number of restaurants, theaters, etc., and proxies of social capital, for example, members in sports clubs or internal social connectivity. I also discuss the advantages and disadvantages of various indicators and provide a visualization in maps for a good understanding of the indicators. By describing how each indicator was obtained, I provide a comprehensive overview of data sources for local amenities in the German context.

My second contribution is the calculation of a cardinal measure of quality of life for all 401 German counties using the hedonic Rosen-Roback framework. Given the above-mentioned assumptions of utility equalization across counties as well as the absence of migration costs, the total value of all local amenities has to offset the regional differences in nominal wages deduced by local costs of living. Following the approach of Albouy (2008) in the tradition of Blomquist et al. (1988), Gyourko and Tracy (1991) and others, I calculate the so-called compensating differential for all counties for a representative German household using national averages of the expenditure share on hous-

ing, the share of labor income of households' gross income, and (effective marginal) tax rates as well as data on regional wages and housing and rent prices. This results in a QoL county ranking where the difference between regions can be interpreted in a meaningful way. Including federal taxation and non-labor income is an extension to the approach of Buettner and Ebertz (2009), who derived a Rosen-Roback-based QoL ranking of German counties using survey data in 2009. My ranking shows a reasonably high level of correlation with popular rankings like the Prognos Zukunftsatlas (2019), the ZDF Deutschland Studie (2018), and the IW Städteranking (2017). In line with many rankings, Salzgitter ranks last, while counties in the Munich region are at the top of my list. According to my calculation, Munich city offers a 12.88% higher QoL than the national average. In contrast, an average household living in Salzgitter experiences a 10.75% lower QoL compared to the national average.

The third contribution and the main goal of the paper is the selection of indicators that explain the variation in QoL derived in the step before reasonably well. Regressing over 100 amenity indicators on my QoL measure using OLS is not feasible for many reasons, mainly because of multicollinearity and the danger of overfitting. Instead, I use random forest regression to derive the relative importance of single indicators as well as groups of indicators. Additionally, I use a least absolute shrinkage and selection operator (Lasso) regression to filter out less relevant or highly correlated amenity indicators. The first approach does this without any upfront imposed structure, including every indicator I collected. In an alternative approach, I drop indicators using economic reasoning before applying Lasso and random forest. To investigate broader groups of indicators, I use principal component analysis (PCA) to estimate an index of each respective category, e.g. a natural environment index. After deriving indices for all clusters, random forest can be applied to compare the relative importance of each group.

I find that the most relevant groups of indicators are socio-political and economic measures. The turnout at federal elections, a social indicator measuring social participation, is identified as the most important predictor. Economic measures like poverty (the share of recipients of unemployment benefits of type SGB II) and proxies of innovativeness (number of new firms per resident or the share employed in the knowledge-intensive sector), as well

as gender equality measures (gender pay and gender employment gap), are also performing well in predicting QoL differences. However, characteristics of the built environment, natural factors, public good provision, and leisure and consumption opportunities are all part of the most relevant variables, where the latter category is the least important. I also show that there is not necessarily a trade-off between model sparsity and explanatory power when dealing with a large number of indicators. Reducing the number of indicators can result in an increase in prediction accuracy, both when using economic reasoning and the results of my statistical learning approach. For example, the prediction error does not increase when using only the ten most important predictors identified by Lasso and random forest regression instead of the full set of 109 variables. Nevertheless, my results suggest that there is no entirely irrelevant category. Hence, I recommend including at least one indicator of each of the six broad amenity groups defined by Lambiri et al. (2007).

The remainder of the paper is structured as follows. The next section goes through the QoL literature both in economics and related fields to identify potential determinants of QoL. Section 1.3 describes the empirical strategy, including the model to derive the QoL measure itself as well as the statistical learning methods to identify its determinants. In the subsequent section, I present the data collected both for my QoL differentials and the amenity indicators. In section 1.5, I derive the cardinal QoL measure and discuss the resulting county ranking. This measure is then used in section 1.6 to identify the most important amenity indicators. The last section concludes.

1.2 Quality of Life in the literature

The interest in *quality of life* (QoL) is older than the field of economics. And since then, economics was, of course, not the only discipline that tried to grasp a better understanding of what determines QoL. The broad and inter-disciplinary character of this concept is already visible in the terminology. While urban economists mostly use the term quality of life, you can also find concepts of *well-being*, *happiness*, or *satisfaction (with life)*. Thorndike (1939), one of the first to comprehensively assess QoL systematically, used indicators of what we call amenities today to rank 300 cities by their "goodness of life".

In the 1960s, when researchers started to question GDP as a welfare indicator, a whole social indicator movement formed to find ways of measuring QoL at the national but also at the local level. Smith (1973) summarized this development in his book "The Geography of Social Well-Being in the United States: An Introduction to Territorial Social Indicators". Campbell et al. (1976) had a notable impact on sociology with their book "The Quality of American Life: Perceptions, Evaluations, and Satisfactions". Rosen (1979) and Roback (1982) introduced the hedonic price model that this paper is based on in economics, also using the term quality of life. Kahneman et al. (1999) provided a more psychological perspective in their seminal book "Well-Being: Foundations of Hedonic Psychology". All of these concepts are related and mean similar things. But in some cases, like in questionnaires, subtle differences can become very important. It should be noted that I use these terms more or less interchangeably here while having, typical for urban economics, a regional interpretation in mind. In contrast, for example, Kahneman et al. (1999) focused on subjective well-being and the individual perception of the world. The link between the two concepts is that residents of a location experience various characteristics as (dis-) amenities that directly impact their individual well-being. How (dis-) amenities are perceived and valued and how individuals interact with them is certainly very heterogeneous. In fact, there is a whole literature on how QoL itself but also the interaction with amenities depends on individual attributes.[3] However, this paper aims to estimate a single measure of QoL per location, relying on the assumption of one (uniform) representative household. That is also reflected in my indicators. In order to present a method that can be applied to other settings, I focus on "objective" indicators that can be reliably collected on a spatial scale like the concentration of NO_2 particles in the air or crime rates rather than "subjective" indicators, which almost always require surveys, like the **perception** of air quality or general life satisfaction.[4]
This brief assessment of QoL terminology already shows: the research on

[3]Additionally to Kahneman et al. (1999), Hsieh (2003); Beeson (1991); Black et al. (2009); Lee (2010) are notable examples.

[4]See Marans and Stimson (2011), chapter 1, for an overview. Some studies define behavioral indicators, like election turnout, to be a distinct third group of indicators. Since these indicators can be measured objectively on a regional level, a differentiation is redundant for this paper. Anyway, some researchers like Costanza et al. (2008) argue that subjective and objective indicators often measure the same.

QoL is not only old and rooted in many different disciplines. It is so diverse and extensive that it has become its own field with its own conferences and journals. Therefore, I can not claim to give a complete overview of the whole universe of QoL research. Instead, the two goals of the remainder of this section are first, to present methods and models to measure QoL that have been used so far and second, to identify (categories of) indicators to measure QoL. Hence, I will refer to reviews and summaries whenever possible and will showcase only some studies on more specific topics exemplarily when collecting indicators for my empirical approach.

1.2.1 Measuring QoL

There are several lines of the literature that touch on the topic of QoL. Some of them deal with QoL only indirectly. A large literature in urban economics attempts to find determinants of urban growth and, related, the urban wage premium (see Ahlfeldt and Pietrostefani (2019) and Duranton and Puga (2020) for reviews). This group of studies focuses on the question of what characteristics of a location attract workers, firms, or economic activity in general. Since QoL is not the center of attention in this body of literature, only relatively few indicators are used to proxy QoL, if any. Therefore, I focus on papers that have measuring QoL as their main purpose.

Overall, there are three different approaches to estimating QoL. First, researchers and various national and international organizations published city or region rankings based on a list of indicators and corresponding weights. One of the first examples from the academic literature is the Places Rated Almanac by Boyer et al. (1985).[5] Popular international rankings like the Mercer "Quality of living city ranking" or the "world's most liveable cities" ranking by The Economist work similarly, as well as the "Prognos Zukunftsatlas", the "ZDF Deutschland-Studie", and the "IW Städteranking" in the German context. This ad-hoc approach comes with a lot of problems, most notably the subjectivity in the choice of indicators and their weights. Some projects aimed to solve the latter problem of arbitrary weights by using survey data to derive the relative importance of indicators (see for example Okulicz-Kozaryn and Valente (2019) using quantitative survey data or Biagi

[5] Savageau (2007) is a more recent example.

et al. (2018) for a qualitative approach). Nevertheless, one can argue that popular rankings have the advantage of being verified by a public audience. If the Mercer "Quality of living city ranking" would produce "non-intuitive" results year by year, the perceived credibility and popularity can be assumed to suffer over time. That does not mean that these rankings are more reliable, but I assume that they reflect a public perception of relative attractiveness to some extent. For that reason, I will compare my QoL to the German region rankings mentioned above.

The second, arguably more objective approach is the hedonic price method based on Rosen (1979) and Roback (1982). The underlying idea of the so-called Rosen-Roback framework is that the value of location characteristics like amenities that have no explicit market price is embodied in house prices and (local) wages. Households choose where to live by picking a location-specific bundle of house prices, income, and amenities that suits their preferences best, i.e. that maximizes their utility. If a household lives in an area that offers a low level of QoL (e.g. because of poor air quality), it has to be compensated by higher wages and/or a lower cost of living. For that reason, differences in amenity values are also called the compensating differential in that context. Without that compensation, the household would move to another region that yields higher utility, driving up wages by reducing labor supply and decreasing house prices through a lower level of demand in the abandoned location. By assuming the absence of migration costs, the economy is in a competitive spatial equilibrium if no household has an incentive to move, i.e. when utility is equalized over space. The assumption of free mobility leads to perfect spatial arbitrage since no household can be better of by moving. Then, the implicit price of amenities is fully "capitalized" in house prices and wages. Hence, given data on regional house prices and local wages, an implicit price of all location characteristics can be derived. By summing up indicators of amenities weighted by these implicit prices one can obtain a total value of QoL for each location. The last step is similar to the ad-hoc approach presented above, except that the weights are derived objectively. However, the choice of which indicators are included in the model is still made by the authors. That choice can be driven by the focus of the paper or more practical reasons like data availability. Because of that, the QoL literature lacks consensus and consistency about

relevant indicators as Lambiri et al. (2007) argue in their review.

Following the seminal work of Rosen (1979), many papers have used the framework to rank cities by QoL while refining the model. Blomquist et al. (1988) added agglomeration effects by allowing productivity and therefore wages to depend on city size. Gyourko and Tracy (1991) were the first to include a public sector. They argue that local public services like schools are not capitalized in wages or housing prices if the service is fully financed by local taxes. The latest extension was presented by Albouy (2008) who emphasizes the role of federal taxation and non-labor income. Progressive federal taxes reduce higher nominal wages more than lower ones and therefore diminish the importance of (nominal) wage differentials. In addition, non-labor income, for example from capital, that is not place-depended reduces the role of wage differentials further. Notable examples of German QoL rankings based on the Rosen-Roback framework are the studies of Buettner and Ebertz (2009) and Hiller and Lerbs (2015). Buettner and Ebertz (2009) rank German counties based on a survey on a wide range of social and political issues carried out in 2004 and 2005. Hiller and Lerbs (2015) extend the work of Buettner and Ebertz (2009) to labor market regions to deal with spatial correlation.

As a third way, a whole strand of the literature has embodied the idea of hedonic price regressions in discrete choice models in the tradition of McFadden (1974). These models estimate a household's probability of choosing a location to live based on wages, housing costs, and a set of location-specific amenities (Cragg and Kahn, 1997; Bayer et al., 2007; Lee et al., 2021). More recently, a quickly growing literature embodies the discrete choice framework into quantitative spatial models (see Redding and Rossi-Hansberg (2017) for an overview). A notable attempt to estimate regional QoL using a quantitative spatial model is the work of Diamond (2016) which incorporates the idea of amenities depending endogenously on city size. As the canonical hedonic models, the quantitative spatial models usually assume housing and labor markets to be in a long-run spatial equilibrium where no household has the incentive to move from one location to another in the absence of migration costs. Some related studies relaxed this assumption by using data on bilateral migration flows in a traditional discrete choice framework (Douglas, 1997; Wall, 2001; Nakajima and Tabuchi, 2011). Quantitative spatial

models have been extended as well to incorporate migration costs. Ahlfeldt et al. (2020) for example shows that QoL differentials are substantially larger compared to those of the canonical hedonic price framework when amenities do not fully capitalize in labor and housing markets. These approaches are especially data-intensive since they require information on bilateral migration flows over time. Depending on the spatial resolution, this data is often hard to get. A more detailed discussion about my choice of the theoretical model is presented in section 1.3.

1.2.2 Amenity indicators

This subsection aims to identify relevant categories of local amenities that have been used in the past literature to investigate QoL and similar concepts. In addition, I want to collect numerous indicators for each group that has been used to measure regional attractiveness. Even more than the methods, the used indicators are countless and very diverse. Lambiri et al. (2007) provide an excellent overview of QoL indicators used in the urban economic literature.[6] Based on the previous work, they propose to group indicators into six categories: (1) natural environment, (2) built environment, (3) socio-political environment, (4) local economic environment, (5) cultural and leisure environment, and (6) public policy environment.

Column 2 of Table 1.1 groups all indicators listed by Lambiri et al. (2007) into their proposed six categories. The list is supplemented by indicators used by Buettner and Ebertz (2009) and Hiller and Lerbs (2015) as well as by the above-mentioned public county rankings "ZDF Deutschland-Studie 2018"[7] and "Prognos Zukunftsatlas 2019"[8] for the German context. In addition, I defined subcategories for my empirical approach later, which also provides a better overview. Of course, some indicators can not be linked to one of the six categories uniquely. Land use variables, such as the share

[6] I refer the interested reader to the works of Gyourko and Tracy (1991), Blomquist (2006), Marans and Stimson (2011) and Sollis et al. (2022) for additional summaries also including other disciplines.

[7] `deutschland-studie.zdf.de`, last accessed on 10.11.2022. The homepage went offline before the time of publishing this paper. Information on results and indicators can be shared upon request.

[8] `https://www.prognos.com/de/projekt/zukunftsatlas-2019`, last accessed on 09.01.2023.

covered by forest, could either belong to the natural environment or public policy environment since local authorities can decide to some extent how much of the available area should be dedicated to the forest. I decided to define them as part of the natural environment, as the exogeneity to local authorities is less important here than how a representative household perceives characteristics. Average living space is assumed to be more dependent on purchasing power rather than features of the local housing stock. Therefore, it is categorized as an economic indicator. In general, it should be mentioned that variables of local housing and labor markets are not treated as amenities in this paper. Instead, they are used as controls when deriving a house price index and the average income a region offers to account for structural differences between local housing and labor markets. The approach will be explained in more detail in section 1.3. Given their focus, the studies of Glaeser et al. (2001) and Florida (2002) use a large number of different recreation/leisure measures (like golf courses, bowling lanes, family theme parks, automobile race tracks, live performance venues, etc.). As my focus is more general, I decided to summarize these indicators as "various recreation facilities" for the sake of the readability of the table. There are some more categories of indicators that received increased attention from dedicated papers. The field of ecological economics has a long tradition of evaluating the implicit price of the natural environment (see Schaeffer and Dissart, 2018, for a summary) while papers like Shapiro (2006) focus on the role of human capital and the composition of the local population. As already mentioned above, column 2 of Table 1.1 includes the most common indicators, but not necessarily all indicators that have been used in the past.

When comparing the objective indicators collected from quantitative research with subjective indicators that are more common in other disciplines, the low number of social and community indicators is striking. Sollis et al. (2022) name several subjective "well-being areas" that fit into the socio-political environment defined by Lambiri et al. (2007) (that is Family relationships; Other relationships; Community and belongingness; Intimate relationships; and Treated with dignity and respect). Nonetheless, election turnout and the number of marriages are the only objective indicators that represent that category separate from general demographic characteristics of the size and composition of the local population. In section 1.4.3, I describe

in detail how I fill that gap when collecting my own indicators for this study.

Table 1.1: Comparison of QoL indicators

Subcategory	Indicators in previous studies	Own indicators	Source
	(1) Natural environment		
Climate	Average temperature	Average temperature	DWD
	No. of rainy/cloudy days		
	Precipitation	Precipitation	DWD
	No. of sunny/clear days		
	Sun hours	Sun hours	DWD
	No. of heating degree days		
	No. of cooling/freezing degree days		
	Total snowfall		
	Humidity		
	Average wind speed		
	Seasonal temperature variation		
Air quality	Particulate (PM_{10}) emission	Particulate (PM_{10}) emission	UBA
		Particulate ($PM_{2.5}$) emission	UBA
	Sulfur dioxide (SO_2) emission		
	Nitrous oxide (NO_x) emission	Nitrogen dioxide (NO_2) emission	UBA
	Methane (CH_4) emission		
		Days with high Ozone (O_3) $\mu g/m^3$	UBA
	Inversion		
	Visibility (in miles)		
	No. of days with high air quality		
Natural environment	Distance from coast / lake	Distance from coast	BKG, own calculation
		Average slope	BKG, own calculation
	Inland water area	% of bodies of water	INKAR, RDB
	National parks	% of natural spaces	INKAR, RDB
	% of forest	% of forest	INKAR, RDB
	% of built-up area		
	& of open spaces	& of open spaces	INKAR
		& of agricultural spaces	INKAR, RDB
Location	Various location dummies	East / West dummy^{++}	own calculation
	(2) Built environment		
Housing market	% of housing with water/electricity	% of new flats with renewable energy	INKAR
	No. of rooms	No. rooms per flat^{++}	Immoscout24
		Living space per flat^{++}	Immoscout24
	No. of bathrooms		
	Age of housing	Age of housing^{++}	Immoscout24
Urbanization	Total population	Total population^{+}	INKAR
	Population density	Population density^{+}	INKAR
		Habitat density^{+}	INKAR
	Degree of urbanization	Degree of urbanization^{+}	INKAR
	(3) Socio-political environment		

Continued on next page

Table 1.1 – *Continued from previous page*

Subcategory	Indicators in previous studies	Own indicators	Source
Demographics	Diversity	Female share of council	INKAR
	Population grown rate	Population grown rate^{+}	INKAR
		Migration balance^{+}	INKAR
	Fertility rate	Fertility rate^{+}	INKAR
	No. of marriages	No. of marriages	INKAR
		No. of divorces	INKAR
	% of population aged 22-29	% of population aged 18-29^{+}	INKAR
		Average age $^{++}$	INKAR
	% of students	No. of students^{+}	INKAR
		No. of apprentices^{+}	INKAR
Social participation	Election turnout	National election turnout	INKAR
		% of residents in sportsclubs	DOSB, own calculation
		Social Connectedness Index	Facebook
Noise Pollution	Noise Pollution		
	(4) Economic environment		
Labor market	Unemployment rate	Unemployment rate	INKAR, RDB
		Long-term unemployment rate	INKAR
		Vacancies^{+}	INKAR
	Poverty rate	Poverty rate	INKAR
		Children poverty rate	INKAR
		% of residents in dept	INKAR
	% of out-commuters	% of out-commuters^{+}	INKAR, RDB
		% of in-commuters^{+}	INKAR, RDB
	Rate of unionization		
	% of highly qualified worker	% of highly qualified worker^{++}	INKAR
	% of low qualified worker		
		% employed in creative sector	INKAR
		% employed in knowledge intensive sector^{+}	INKAR
	% gender wage gap	gender wage gap	INKAR
		gender employment gap	INKAR
		Average working hours^{++}	INKAR
Firms	% primary sector production	% employed in primary sector^{++}	INKAR, RDB
	% secondary sector production	% employed in secondary sector^{++}	INKAR, RDB
		% employed in tertiary sector^{++}	INKAR, RDB
	% of large firms	% of firms by size^{++}	INKAR
	No. of new firms	No. of new firms	RDB
		No. of firm closures	RDB
	No. of patents	Investment	RDB
Fiscal variables	Local tax rates	Local tax rates	Statistikportal
	Local public dept	Local public dept	INKAR
		Staff of local administration	INKAR
Cost of living	Non-land cost of living	CPI (w/o housing)$^{++}$	see below
	Expenditure share for food		
	Average living space	Average living space^{++}	INKAR
	(5) Cultural & leisure environment		

Continued on next page

Table 1.1 – *Continued from previous page*

Subcategory	Indicators in previous studies	Own indicators	Source
Recreation & entertainment	No. of professional sports teams		
	No. of museums	No. of museums	OSM
	No. of restaurants	No. of restaurants	OSM
	Green space area	% recreation areas	INKAR
	No. of further recreational facilities	No. of theaters	OSM
		No. of shops	OSM
Culture	Media intensity (radio, television)		
	Public library acquisitions		
	No. of tourist overnight stays	No. of tourist overnight stays^{+}	INKAR, RDB
	No. of hotel rooms	No. of hotel beds^{+}	INKAR, RDB
	(6) Public policy environment		
Health	No. of physicians	No. of physicians	INKAR, OSM
		Access to GPs	INKAR
	No. of hospital beds	No. of hospital beds	INKAR
		Staff in nursing homes	INKAR
		No. of nursing home places	INKAR
		Staff in nursing services	INKAR
		Access to pharmacies	INKAR
	Infant mortality	Infant mortality^{+}	INKAR
	Life expectancy	Life expectancy^{+}	INKAR
	Obesity rate		
	No. of smoker		
Public safety	Total crime rate	Total crime rate	PKS
	Violent crime rate	Violent crime rate	PKS
		Street crime rate	PKS
	Property crime rate		
	% crime victims		
	Quality of fire service		
	Road traffic casualties	Road traffic casualties	INKAR, RDB
Education	Student/teacher ratio	Student/teacher ratio	KBS
	% of children in secondary school		
	High school drop out rate	High school drop out rate	KBS, INKAR
		% of graduates by level of degree^{++}	INKAR
	Childcare places	No. of children in childcare	INKAR
		No. of kindergartens	OSM
		No. of schools	KBS, OSM
		Access to primary schools	INKAR
Infrastructure & mobility	No of telephones	Coverage with high speed internet	INKAR
	Access to high speed railway	Access to high speed railway	INKAR
	Access to highways	Access to highways	INKAR
	Access to airport	Access to airport	INKAR
		Access to public transport	INKAR
		Access to bigger city^{++}	INKAR
		Access to grocery stores	INKAR
		No. of cars	INKAR, RDB

Continued on next page

Table 1.1 – *Continued from previous page*

Subcategory	Indicators in previous studies	Own indicators	Source
N	88	109	

Notes: All indicators which measure a total amount ("no. of ...") are per capita, except for total population. + indicates indirect measures where a more direct measure is available. Indicators with ++ are controls. INKAR = Indikatoren und Karten zur Raum- und Stadtentwicklung by Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR); RDB = Regionaldatenbank Deutschland by Statististische Ämter des Bundes und der Länder; UBA = Umweltbundesamt; KBD = Kommunale Bildungsdatenbank; DOSB = Deutscher Olympischer Sportbund; DWD = Deutscher Wetterdienst; BKG = Bundesamt für Kartographie und Geodäsie; PKS = Polizeiliche Kriminalstatistik; CPI = Comsumer price index by Weinand and von Auer (2020). The Statistikportal is a new online database by Statististische Ämter des Bundes und der Länder which is in the process of building up by the time of writing this paper. More details on sources and definitions of single indicators can be found in section 1.A in the Appendix.

1.3 Empirical Strategy

To evaluate the relative importance of different amenity indicators it is necessary to have a cardinal measure of regional QoL. Then, the set of indicators can be regressed on QoL to obtain information on the explanatory power of each indicator in explaining the variation in the overall attractiveness of regions. Therefore, deriving an arguably objective (cardinal) measure of QoL on the level of German counties is the first step which will be described in the next subsection. After that, I discuss several statistical dimension-reducing methods to identify the most relevant indicators both in general but also within each of the above-defined categories in section 1.3.2.

1.3.1 Estimating local QoL

I summarized the three methods to construct a QoL ranking in section 1.2.1 above. The first way using arbitrarily weighted sums of location characteristics is not suitable for the goal of this paper. The choice of included indicators as well as their relative importance is not based on an objective assessment but based on the author(s) preference/opinion. Even when building on survey data, the inferred relative preferences over location characteristics are not reproducible in other contexts like other countries where no comparable

survey was conducted.
The third method involving discrete choice models is beyond the scope of this paper. Usually, data on bilateral migration flow is required to infer a distribution of location-specific amenity preferences. This data is often not publicly available, especially on lower levels of spatial disaggregation. Therefore, I use the hedonic price method based on Rosen (1979) and Roback (1982). It should be noted that the Rosen-Roback framework allows estimating implicit prices for each amenity directly which already provides an insight into the relative importance of amenities. However, this approach requires adding every single indicator to both, the hedonic wage and the hedonic house price regression. When including a large number of indicators, these regressions suffer from multicollinearity and overfitting.[9] Therefore, I choose to follow the approach which allows estimating total QoL by using the combined total difference in wages and house prices (cf. Gyourko et al., 1999). The derived QoL is then based on the whole bundle of amenities in each location without estimating prices for every single indicator. The approach of Albouy (2008) which includes taxation and non-labor income is the most advanced version of the method and suits the purpose of the paper to estimate reasonable QoL differentials in an adequate way.

As already described above, the approach builds on the assumption of utility equalization across space in a spatial equilibrium. Intuitively, every region has to offer every household the same level of utility in equilibrium. Then, if everyone is indifferent about where to live, there are no incentives to reallocate. In contrast, if this condition was violated, individuals would move to increase utility in the absence of migration costs, which can not be the case in a spatial equilibrium. The assumption of free mobility is common in the existing literature as explained in section 1.2.1. On the one hand, it comes with several questionable implications which will be discussed below in more detail. On the other hand, this assumption enables me to calculate an objective measure for regional QoL directly from regional wage- and cost-of-living-differentials. Ahlfeldt et al. (2020) show that QoL differentials are substantially larger when allowing for migration costs. Since I only use relative QoL differentials between counties, an unbiased up-scaling

[9] In fact, if the number of indicators is close to or even higher than the number of regions, the regression equation is not identifiable anymore.

of QoL values does not change my results. Only if migration costs of a representative household depend systematically on the location of residence, the assumption of free mobility would bias my QoL ranking. However, it is a discussion on its own on how to distinguish between migration costs and location preferences. If low rates of mobility represent the (relative) preference to stay, I consider that as part of the QoL a location offers. For example, the social cost of moving might be high because of strong social ties, something which is considered an amenity in my paper. I conclude that relying on a model setup similar to Albouy (2008) with free mobility is sufficient.

The model assumes the country to be populated by a large number of households, such that every region can have a sufficiently large number of inhabitants. The preferences of these households can be represented by a utility function of the form $U = U(\mathbf{y}, Q)$, where $\mathbf{y}$ denotes a vector of consumable market goods and Q denotes the QoL.[10] The utility is increasing in both of these arguments. Note that so far, no spatial dimension is inherent in the utility function. The utility derived from consuming goods as well as experiencing QoL does not depend on the location of individuals. However, I assume regions denoted by i to differ in terms of their locational characteristics or amenities. The QoL a region offers depends on a vector of amenities ($\mathbf{A_i}$), such that $Q = Q(\mathbf{A_i})$.

Each household inelastically offers one unit of labor earning gross labor income w_i which varies across regions.[11] In addition to labor income, households generate income from other sources such as capital income. This non-labor income I is independent of the place of residence, as for example, the return from assets does not change when moving from one region to another. Thus, total gross income $m_i = w_i + I$ consists of a location depended and a location-independent part. Income is taxed with a federal tax rate τ. With a total net income of $(1 - \tau)m_i$, households can purchase goods $\mathbf{y}$ given local prices $\mathbf{p_i}$. The representative household's optimization problem

[10] The vector of consumable market goods can include services as well as housing. Therefore, not all of these goods are necessarily tradeable between regions as housing for example is not.

[11] The assumption of inelastic labor supply has no first-order effects on QOL estimates as shown by Roback (1980).

is therefore

$$\max_{i,\mathbf{y}} U(\mathbf{y}, Q(\mathbf{A_i})) \text{ s.t. } \mathbf{p_i y} = (1-\tau)m_i. \tag{1.1}$$

By duality, this maximization problem can be expressed equivalently as an expenditure minimization problem:

$$E(\mathbf{p_i}, w_i, \tau, u; Q_i) = \min_{i,\mathbf{y}} \{\mathbf{p_i y} - (1-\tau)(w_i + I) : U(\mathbf{y}, Q(\mathbf{A_i})) \geq u\}. \tag{1.2}$$

In the absence of migration costs, the utility has to be equalized across all regions, meaning that every household enjoys the same level of utility, $\bar{u}$. In equilibrium, every household has to reach this level of utility by spending the whole net income on goods (including housing):

$$E(\mathbf{p_i}, w_i, \tau, \bar{u}; Q_i) = 0 \; \forall \; i. \tag{1.3}$$

An alternative interpretation of that condition is, that no household has to receive additional compensation (e.g. by a lump-sump transfer) to live in a specific region. The next step is to totally differentiate Eq. (1.3) around national averages to arrive at

$$\frac{\partial E}{\partial \mathbf{p_i}} \mathbf{dp_i} + \frac{\partial E}{\partial w_i} dw_i + \frac{\partial E}{\partial Q} dQ_i = 0. \tag{1.4}$$

This trick allows for deriving one unified formula for every region, where local price, wage, and QoL levels are expressed as (marginal) deviations from the national average $(\bar{\mathbf{p}}, \bar{w}, \bar{Q})$.[12] By applying Shepard's Lemma, it becomes more salient that Eq. (1.4) implicitly incorporates the marginal tax rate τ' and the marginal willingness-to-pay for QoL $\frac{\partial E}{\partial Q}$: $\mathbf{y dp_i} + (-1+\tau')dw_i + \frac{\partial E}{\partial Q} dQ_i = 0$.[13] Rearranging and expanding by national averages turns total absolute deviations into relative differences. Defining $\hat{w}_i \equiv dw_i/\bar{w}, \hat{\mathbf{p}}_i \equiv \mathbf{dp}_i/\bar{\mathbf{p}}$ and normalizing $\hat{Q}_i \equiv -(\frac{\partial E}{\partial Q}) dQ_i/\bar{m}$ produces

$$\hat{Q}_i = \frac{\mathbf{y}\bar{\mathbf{p}}}{\bar{m}} \hat{\mathbf{p}}_\mathbf{i} - (1-\tau') \frac{\bar{w}}{\bar{m}} \hat{w}_i, \tag{1.5}$$

[12] One could also differentiate around a reference region are any other point. However, taking the total differential implicitly assumes that all changes considered are small changes. To make this assumption as mild as possible, it makes sense to minimize the deviations by taking the national average as the reference point.

[13] A more detailed derivation can be found in section 1.B in the Appendix.

where $\bar{m}$ denotes the national average of total gross income. $\frac{\mathbf{y}\bar{\mathbf{p}}}{\bar{m}}$ is a vector of (national average) expenditure shares on goods, including housing while $\frac{\bar{w}}{\bar{m}}$ represents the (national average) share of gross income received from labor. $\frac{\mathbf{y}\bar{\mathbf{p}}}{\bar{m}}\hat{\mathbf{p}}_\mathbf{i}$ measures how high the effective cost-of-living in location i is compared to the national average and $(1-\tau')\frac{\bar{w}}{\bar{m}}\hat{w}_i$ represents similarly the percentage deviation of total income netting out federal taxes relative to the national average. Hence, $\hat{Q}_i$ equalized differences in the cost of living and income between regions where QoL is higher if cost-of-living is high or net income is low. With suitable parameters as well as income and price differentials described in the next section, Eq. (1.5) is used to calculate a relative QoL measure for all 401 German counties in section 1.5.

1.3.2 Identifying relevant indicators

Building on the relative QoL ranking derived with Eq. (1.5), the next step is to identify the most relevant determinants of QoL. It is important to emphasize that I do not aim to draw any causal conclusions. Since indicators often interact with each other, it is very challenging to estimate the causal impact of specific indicators on local well-being without some kind of natural experiment. The goal here is to find indicators that can explain or predict regional variation in QoL in purely statistical terms. Previous studies presented in section 1.2 used implicit prices derived within the Rosen-Roback framework or principal component analysis (PCA) to derive a data-driven amenity index (Diamond, 2016). The former approach is not feasible with a large number of indicators due to potential multicollinearity and the threat of overfitting. The latter way is preferential for summarizing many indicators into one index but is neither suitable for identifying (most) relevant indicators nor for comparing single amenities indicators in terms of relative importance.

My main statistical learning method is the supervised machine learning technique random forest. Random forest is especially suitable for identifying the best predictors and their relative importance when the (true) relation between variables is not necessarily linear (Hastie et al., 2015a).[14] In addition, Hastie et al. (2015a) argue that forest algorithms are relatively good in

[14] Basuchoudhary et al. (2017) compare different machine learning methods with regard to prediction performance for economic growth. They conclude that the (boosted) random forest algorithm is among the best methods for prediction.

handling missing values in the independent variables and in dealing with irrelevant inputs, two characteristics that are important in the context of this paper. Random forest is an ensemble tree-based learning algorithm first proposed by Breiman (2001) and involves stratifying or segmenting the predictor space (set of possible values for all predictors) into a number of simple regions. In a first step, the algorithm creates a large number of so-called trees, which is a number of consecutive splitting rules. Given the response or dependent variable Y, the algorithm minimizes the residuals by splitting (segmenting) the sample into two distinct regions, e.g. $< X_j$ and $\geq X_j$, where X_j is the best predictor of a randomly chosen subset of indicators. All observations in one region get the same prediction of Y. This process is repeated (for each region of the previous split) with newly drawn subsets of predictors to grow trees. In a second step, numerous trees consisting of several splitting rules each are combined to obtain one single prediction rule (ensemble tree). Generally, due to the flexibility in the functional form, forest regressions are prone to overfitting (Hastie et al., 2015b). To overcome this, I additionally use *bootstrap aggregating* (bagging) to increase predictive accuracy. With bagging, each tree is fitted on a bootstrap sample instead of the original sample, which increases out-of-sample performance. Additionally, bagging allows calculating a valid error measure using observations that were intentionally left out of the bootstrap sample before (out-of-bag error). I choose the two tuning parameters, i.e. the number of iterations and the number of randomly chosen variables at each split, such that the out-of-bag prediction error is minimized. After running the optimized random forest regression, I can calculate the total amount of the residual sum of squares (RSS) reduction due to splits for each (used) predictor. That allows me to derive a measure of the relative importance of each indicator for predicting QoL differentials.

To compare the statistical relevance of the six main categories proposed by Lambiri et al. (2007), rather than single indicators, I use the above-mentioned dimension reduction method principal component analysis (PCA) to derive indices for each environment as in Diamond (2016). PCA reduces the dimensionality of a data set by generating a new, smaller set of inputs that retains most of the sample's information in form of variation. In contrast to variable selection models, PCA generally does not drop irrelevant indicators

and, more importantly, does not aim to provide predictors for any kind of dependent variable (e.g. QoL differentials). Therefore, PCA is also called a unsupervised shrinking method. Instead, PCA derives so-called principal components (PCs) only based on the correlations between the original inputs. Each PC is a linear least squares fit of the sample and linear independent of all other PCs. In a sample of N observations with a vector $\mathbf{x} = (x_1, x_2, ..., x_P)$ with a total of P inputs (i.e. indicators), the first PC is defined as

$$z_1 \equiv \mathbf{a_1}^T \mathbf{x} = \sum_{i=1}^{p} a_{i1} x_i. \tag{1.6}$$

It is derived by choosing the vector of coefficients $\mathbf{a_1} = (a_{11}, a_{21}, ..., a_{p1})$ such that it maximizes the variation of the first PC z_1. Further components $k > 1$ require to fulfill the additional constraint of being linear independent of all previous components, i.e. $cov[z_k, z_l] = 0$ for $k > l \geq 1$.[15] The derived indices of the six main categories of indicators are then used as inputs in a random forest regression.

In the second part of the analysis, I use single indicators instead of indices. As one goal of this paper is to provide a guideline on how many indicators to include, I use the supervised shrinkage method least absolute shrinkage and selection operator (Lasso) to reduce the number of predictors before applying the random forest algorithm. Lasso regressions drop variables that have low explanatory power and are especially useful for variable selection (Tibshirani, 1996). Technically, this is done by adding a constraint on the absolute size of coefficient estimates to the regular ordinary least squares regression. The Lasso coefficients β_λ minimize the RSS and the so-called shrinkage penalty in the form of

$$\sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} \beta_j x_{ij}) + \lambda \sum_{j=1}^{P} |\beta_j|, \tag{1.7}$$

where P is the total number of indicators, also called predictors or inputs, and N is the total number of observations. λ is the Lagrangian multiplier acting as a tuning parameter. If λ is chosen to be large the penalty is strong

[15] A further, rather technical constraint is the normalization of loadings: $\mathbf{a_k}^T \mathbf{a_k} = 1$.

and only a few most important predictors are included in the estimation model. If λ is small (close to 0), the model is more similar to the standard OLS regression with only very few, if any, excluded variables. The choice of the tuning parameter embodies the trade-off between lower levels of variance and sparsity (large λ) versus prediction accuracy (small λ). To measure prediction accuracy, the sample is split into two parts. One part (here 75% of the sample) called the training sample is used to fit the model, i.e. to estimate coefficients. The other part (here 25% of the sample) is intentionally left out for estimation. Instead, the estimated coefficients based on the training data are used to predict the dependent variable of the remaining, so-called test or validation sample. That allows for calculating an out-of-sample mean squared error (MSE). This process is also known as cross-validation (CV). One version of CV, K-fold cross-validation, is often suggested to minimize the prediction error (Hastie et al., 2015b).[16] However, as cross-validation tends to over-select predictors, I use an extension of CV-based parameter selection introduced by Zou (2006) called adaptive Lasso (see Bühlmann and Van De Geer (2011) and Chetverikov et al. (2021) for more information on over-selection of the traditional CV Lasso). Adaptive Lasso selects a subset of relevant predictors using cross-validation in the first step. In the second step, each input j gets a penalty loading of $w_j = 1/|\hat{\beta}_j|$ where $\hat{\beta}_j$ is the estimated coefficient of step one. Since indicators with a small coefficient from the first step receive a stronger penalty in the following step, adaptive Lasso is more likely to drop more variables with relatively low predictive power. This makes variable selection consistent even under a high degree of correlation between predictors in the (true) model and predictors outside of the model, which is very likely to be the case here (see also Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006). As shrinking methods are not scale equivalent, each indicator x_i is standardized by

$$\hat{x}_i = \frac{x_i}{\sqrt{1/n \sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{1.8}$$

[16] K-fold cross-validation partitions the sample into K equal-sized subsets (folds) randomly. Then, the model is fit on a training data set consisting of K-1 folds. The resulting coefficients are used to calculate the prediction error in the validation data set, the hold-out fold. After repeating this procedure for all K folds as the validation set, all prediction errors are combined into one MSE depending on λ.

before estimation (Hastie et al., 2015a).[17] The selected indicators are then used as predictors by the random forest algorithm. I compare the predictive performance with a specification without upfront variable selection. Alternatively, I manually select the two indicators of each category that show the highest correlation with the QoL differentials and run the random forest regression with only 12 inputs. The methods described above are used in section 1.6 after introducing the data in the next section.

1.4 Data

1.4.1 Income and price differentials

As visible in equation Eq. (1.5), the two main determinants of QoL are income and cost of living differentials. $\hat{w}_i$ and $\hat{p}_i$ are supposed to measure the change in wages and prices, respectively, when a representative household moves from one location to another.

Data on gross median income from full-time labor was collected from the online database INKAR (2021). To incorporate structural differences between local labor markets with regard to education, I calculate the median gross income by aggregating the location-specific income of three educational groups (no professional qualification; professional (non-academic) qualification; and academic qualification) weighted by national averages of the group's respective share of the labor force. This is important to measure the expected change in income when a representative household, given its level of education, moves from one location to another. In addition, the special historic background of Germany following the Second World War with a long period of separation until the reunification in 1990 is still measurable in income.[18] Therefore, I net out structural differences between labor markets in "new" and "old" states by regressing the log of income on an east-dummy

$$\ln w_i = east\beta + e_i \tag{1.9}$$

[17] Note that the standardization is also applied on the categorical dummy variable *east*. This is important to make all indicators comparable at the cost of intuitive interpretability (see Tibshirani, 1997).

[18] See section 1.D in the Appendix.

and using the residuals $\hat{e}_i$ as my measure of gross income differentials. This procedure ensures that differences in wages that originate from historical differences are not considered as differences in QoL. Wages in the east are assumed to be lower not because of higher levels of QoL but because of an adjustment process that has not been completed. Note that other structural differences between local labor markets like the sector composition are not controlled for, since this would cause endogeneity problems when using aggregated data. Instead, the share of workers with a college degree and other labor market indicators will be included as controls when identifying important amenities in section 1.6 to make sure the effect will not be picked up by other, potentially correlated measures.[19] A drawback is that this approach does not allow for treating the so-called control indicators, including the share of high-skilled workers as an amenity indicator itself. However, the college share is a rather indirect measure of attractive features of a location. As already mentioned above, these indirect measures are only used when lacking more direct indicators. Lee (2010) for example argues that high-ability workers sort into big cities due to the taste for consumption variety which I measure directly by the number of shops and restaurants per inhabitant in this paper. The estimated wage differentials are visualized on the left-hand side of Figure 1.1.

To measure the cost of living differentials optimally, one would construct price indices for all consumption goods and weight them by national expenditure shares. Weinand and von Auer (2020) use consumer price index micro data and detailed expenditure weights collected by the Federal Statistical Office and statistical offices of the states in May 2016 and 2010, respectively, to construct a general regional consumer price index for Germany.[20] Their results reveal that more than 89% of total price variation can be explained by differences in housing costs, derived with a hedonic price regression on rent data. Adding services accounts for an additional 7% of explained variance only.[21] I conclude that housing differentials are sufficient to mea-

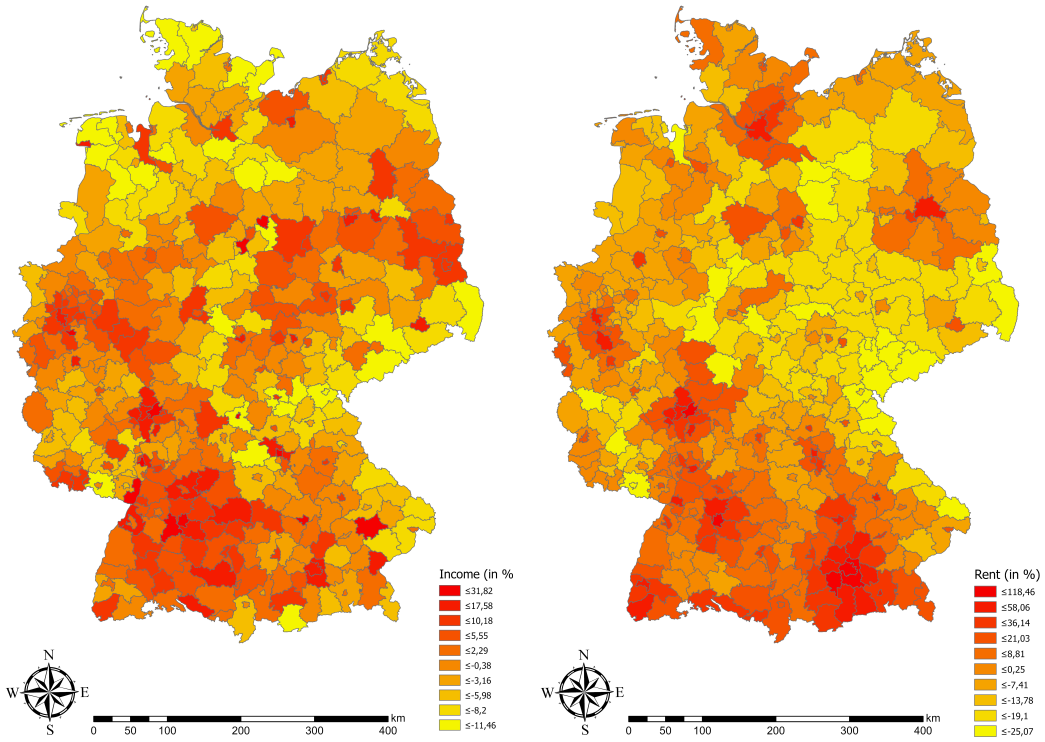
[19] All control indicators are marked with ++ in column 3 of Table 1.1.

[20] They complement information on rents from *Bundesinstitut für Bau-, Stadt- und Raumforschung* (BBSR) for 21 counties.

[21] In addition to the low level of explanatory power, price variations in services are difficult to measure due to the lack of data on quality differences. Easily tradeable goods like food do not require hedonic price regressions since systematic differences in quality between

Figure 1.1: Income and cost of living differentials

(a) Gross income (b) Rents



Notes: (a): Gross median labor income (deviation from the national mean in %) controlling for east-west differences (equation Eq. (1.9)). (b): Rent index (deviation from the national mean in %) from hedonic rent regression Eq. (1.10).

sure the cost of living differences across German counties. To do so, I use property micro data from the online market platform ImmobilienScout24 (ImmoScout24) covering more than 1.2 million renting proposals for apartments in 2017.[22][23] The data were accessed via the FDZ-Ruhr. Following Albouy (2008), I regress the log of apartment rent per square meter $\ln p_i^j$ on apartment-level controls Y_i^j like the number of rooms, type, age, equipment,

counties in Germany can be expected to be negligible. Though, services and housing are harder or even impossible to trade between locations and require controlling for potential spatial heterogeneity, which is impossible for many services due to data availability.

[22] A detailed description of the data is provided by Boelmann and Schaffner (2019).

[23] Platforms like ImmobilienScout24 only report quoted rents from advertisements. Faller et al. (2009) estimate market rents to be 7-8 % lower than quoted rents taken from the first advertisement in Germany. However, since there are no strong systematic regional differences, quoted rents can be used to calculate regional price indices.

etc., and county fixed-effects ρ_i in the form of

$$\ln p_i^j = Y_i^j \beta + \rho_i + e_i^j. \tag{1.10}$$

The estimates of the county fixed-effects ρ_i were standardized to a mean of zero to measure county i's deviation from the national average of housing costs. The resulting rent index is mapped on the right-hand side of Figure 1.1. In line with the results of Albouy (2008) for the US, it is visible that the variation in rents is larger than in income. While rents are 118% more expensive in the city of Munich and 32% cheaper in Vogtlandkreis compared to the national average, gross labor income only ranges from -17% (City of Hof) to 32% (Ludwigshafen am Rhein). To provide a better understanding, the two differentials are also plotted against each other in Figure 1.C.13 in the Appendix.

In contrast to rents, purchase prices of land or houses should reflect the present value of a stream of expected future rents. Therefore, a city that is expected to grow shows higher levels of (relative) purchase prices than rents. However, the housing cost index derived here is not meant to incorporate dynamic effects, which means that the cost of living index derived from rents is preferred.[24] Appendix section 1.D presents alternative cost-of-living measures including the consumer price index of Weinand and von Auer (2020), a purchasing house price index, and a rent price index based on easily accessible rent data from INKAR (2021).

1.4.2 Parameters

I use information from the *Einkommens- und Verbrauchsstichprobe* (EVS) 2018 to calculate the model parameter on housing expenditure share and labor income share (Statistisches Bundesamt (Destatis), 2020).[25] On average, a German household spends 739 EURO per month on housing, where rents are imputed based on housing prices for homeowners. With a total gross income of 4846 EURO, the expenditure share $\frac{\mathbf{y}\bar{\mathbf{p}}}{\bar{m}}$ is roughly 0.1525. As 3122

[24] In addition, a majority of 54% of flats and houses were occupied by renters in 2017 (Statistisches Bundesamt (Destatis), 2017).

[25] The EVS is a survey with roughly 80,000 households participating that is conducted every five years by the official federal statistical office in cooperation with the statistical offices of the states.

EURO is earned in self-employment or dependent employment, the labor income share $\frac{\bar{w}}{\bar{m}}$ amounts to 0.6442.

The marginal federal tax rate should incorporate all labor income-dependent tax payments since they vary between regions. Therefore, I use the "effective" marginal tax rate calculated by Peichl et al. (2013). They consider not only taxation of income but also social security contributions, indirect taxes, and payroll taxes. For an average household with a monthly gross labor income of 3122 EURO, the effective marginal tax rate is 0.49 compared to a marginal income tax rate of 0.29. Plugging the parameters into equation Eq. (1.5) yields

$$\hat{Q}_i = 0.1525\hat{\mathbf{p}}_\mathbf{i} - 0.3285\hat{w}_i. \tag{1.11}$$

This puts twice as much weight on wage differentials relative to variations in the cost of housing. Compared to Albouy (2008) for the US, who derived a wage/cost of living ratio of roughly 1.5, my results suggest that variations in income are more and in housing cost less important for QoL in Germany than in the US. However, my results are fairly similar to the relative weight of Albouy (2008) compared to Blomquist et al. (1988) (3.61), Gyourko and Tracy (1991) (4.82), and others who did not incorporate federal taxation and non-labor income.

1.4.3 Amenity indicators

As noted above, the choice of included indicators when estimating QoL is often very arbitrary and poorly justified, despite being relevant for the results (Lambiri et al., 2007). This section aims to provide a way of picking indicators on the German county level for the year 2017 that is as objective as possible. One of the biggest limitations in the previous literature as well as for me is data availability. Data availability may depend on the spatial level of aggregation, the country, or the time. For example, survey data as used by Buettner and Ebertz (2009) is often taken from non-repeated questionnaires, making reproducing updated results impossible. To provide a more generalized way of how to pick measures of local attractiveness, I follow three heuristics for choosing my set of indicators: (I) Indicators that have been commonly used in the academic literature, as well as public rankings, should be included. Therefore, I try to find comparable measures

of local characteristics to those listed in column 2 of Table 1.1. If there are several potential candidates I included all in the data collection phase before reducing the number of indicators in section 1.6. (II) Every category (sub-category) should be represented by at least one indicator. That makes studies more comparable, even if the same indicators are unavailable. Also, single-focus papers, for example on the value of air quality, might overestimate the importance of clean air via the omitted variable bias when other categories are ignored.[26] (III) Indicators should be easily accessible also at later points in time. That excludes above mentioned non-repeated surveys and assigns more relevance to publicly available data.

Given these heuristics and building on the previous literature summarized in section 1.2 above I collected over 100 potential determinants of QoL. Column 3 of Table 1.1 assigns these indicators into the six categories defined by Lambiri et al. (2007) and compares them to the list of established indicators. A more detailed description of each indicator can be found in section 1.A in the Appendix. Most indicators are available in the online database of *Bundesinstitut für Bau-, Stadt- und Raumforschung* (BBSR) called INKAR (2021) and the *Regionaldatenbank Deutschland* run by *Statistische Ämter des Bundes und der Länder*.[27] Other important sources include the *Bundesamt für Kartographie und Geodäsie* (BKG) for geographical characteristics and distance calculations, the *Deutscher Wetterdienst* (DWD) for data on climate and weather, the *Umweltbundesamt* (UBA) for air quality measures, the *Polizeiliche Kriminalstatistik* (PKS) of the *Bundeskriminalamt* for information on crime, and the *Kommunale Bildungsdatenbank* (KBD) offered by the *Statistische Ämter des Bundes und der Länder* for educational data. As highlighted in section 1.2.2, there has been a lack of social indicators in previous QoL literature. Building on the seminal work by Putnam (2000), there have been several approaches to measure social capital (see Rupasingha et al. (2006) and Joint Economic Committee (2018) for the most prominent ones on the regional level). In general, social indicators aim to capture to

[26] It should be noted that the classification in sub-categories does not play a role in my analysis. After collecting all indicators, it only serves the purpose of organizing the data.

[27] INKAR stands for *Indikatoren und Karten zur Raum- und Stadtentwicklung*, which can be translated to "indicators and maps for area and city development". The INKAR database can be found under `https://www.inkar.de/`. The Regionaldatenbank is the combined successor of the Regionalstatistik and Genesis databases. It is available here: `https://www.regionalstatistik.de/genesis/online/`.

what extent residents feel socially integrated into their local community. All of the mentioned approaches consider, besides the already included election turnout, membership in (local) organizations. To fill the gap of local social indicators in Germany, I collect data on sports team membership from the *Deutsche Olympische Sportbund* (DOSB).[28] In addition, recent publications on measuring social capital exploit new online social media data on social ties from Facebook (e.g. Chetty et al., 2022a,b; Herdagdelen et al., 2022). The Facebook Social Connectedness Index (SCI) broadly measures the propensity of two randomly drawn persons from two locations to have a friendship link in Facebook.[29] I use the likelihood of two users from the same German county being tied in Facebook as a proxy of "internal social network density".

One notable gap in my list of indicators is the subcategory of noise pollution. The UBA provides data on noise pollution by source. However, noise pollution is only measured next to railways, highways, and airports and in densely populated areas. Hence, coverage is highly selective and interpolating and/or aggregating the data to receive a general noise pollution index is not valid.

Most of the data refer to the 31st of December 2017 or the 1st of January 2018. Due to data availability, there are some exceptions: the OSM data covers information on leisure facilities from 2019. Accessibility indicators are from 2015-2020. For example, the access to airports, highways, high-speed railway stations and the next bigger city provided in INKAR (2021) are from 2020, while the access to general practitioners dates to 2015. The reported number of hotel beds refers to 2019. High-speed internet coverage was calculated in 2020. The share of residents in sports clubs refers to January 1st, 2019, or 2020 for some counties. I assume that the indicators do not change much within two years, especially relatively between counties. Section 1.A in the Appendix provides a detailed description of each indicator, including the method, source, and reference time.

It should be mentioned that not all indicators in column 3 of Table 1.1

[28] To be more precise, the data on spots club membership had to be collected manually from sports associations on the state level (or below). In theory, however, the DOSB as the federal umbrella association has all the information included here and will hopefully make them publicly available in the future.

[29] For more information on the Facebook Social Connectedness Index, I refer to Bailey et al. (2018). Access to the data is provided via `https://dataforgood.facebook.com/dfg/tools/social-connectedness-index`.

are treated as measures of amenities in my analysis in section 1.6. Some indicators only serve as controls, indicated by ++. Housing market characteristics (except for the share of new flats with renewable energy) were used as controls when deriving price differentials in section 1.4.1 and can not be interpreted as distinct amenity indicators anymore. As discussed above, labor market characteristics have to be included when identifying the most important indicators in section 1.6. However, they can not be interpreted as amenity indicators as well.

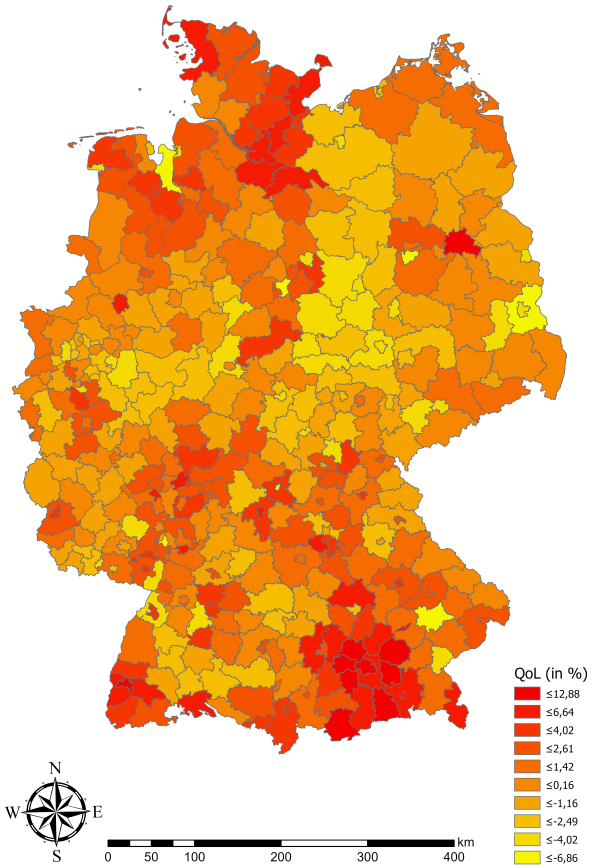
I also distinguish between direct and indirect indicators. For example, the "ZDF Deutschland-Studie 2018" includes life expectancy to rate health care. However, this is more a result of or, at best, a proxy for the quality of health care from a representative household's perspective. Recalling the intuitive logic underlying the Rosen-Roback framework, the number of available doctors is more relevant to a household that decides whether to move to a location or not compared to the average life expectancy of other people around them. A similar argument can be made for tourist overnight stays. A city's attractiveness to tourists might be a decent indicator of the presence of historical sites or natural landmarks as proposed, among others, by Carlino and Saiz (2019). However, it would be preferable to measure those attributes directly, especially as tourists themselves might even be a dis-amenity to residents (Biagi et al., 2020). Hence, I distinguish between direct and indirect indicators, where direct indicators are preferred whenever available. Indirect measures that I assume to measure a similar location characteristic as available direct indicators are marked with a + in column 3 of Table 1.1 and will be excluded from the sparse specification in section 1.6.

1.5 Quality of Life Measure

In this section, I derive a single cardinal measure for regional Quality of Life using equation Eq. (1.11) and the income and cost of living differentials derived in section 1.4.1. Ranging from -10.79% (Salzgitter) to 12.88% (City of Munich), the variation of QoL is sizeable but smaller than differences in income and cost of living. The Figure 1.2 visualizes the QoL measure in a map, while a full ranking of all 401 counties is provided in section 1.E of the

Appendix.

Figure 1.2: Quality of Life of German Counties (2017)



Notes: Quality of life differentials (deviation from the national mean in %) based on equation Eq. (1.11).

As in popular rankings and the ranking by Buettner and Ebertz (2009), many counties with high levels of QoL tend to be located in the south of Germany. Especially several counties from the state of Bavaria including its capital Munich, the Rhine-Main area, and Baden are high on the list. According to my index, QoL is also high near the coast in the north, especially in the greater area of Hamburg. QoL seems to be lower in the center of Germany, especially in Saxony-Anhalt and Thuringia. However, there is no obvious spatial concentration of "losers", as the bottom 15 are located in eight different states, both in East and West Germany.[30] Although some

[30]Note that, since I controlled for east-west level differences, comparisons between east

metropolitan cities of Germany offer relatively high levels of QoL, the overall correlation of QoL differences with the total population ($r_{corr} = 0.24$) is positive but moderate. In line with results from Albouy (2008), the incorporation of federal taxation and non-labor income resolves the puzzle of negative correlation between city size and QoL measures as present in older studies (cf. Burnell and Galster, 1992).

In addition, my QoL ranking shows positive and reasonably high Spearman rank correlations with popular German city or county rankings like the "Prognos Zukunftsatlas 2019" ($r_{spear} = 0.59$), and the "IW Städteranking 2017" ($r_{spear} = 0.6$) and the "ZDF Deutschland-Studie 2018" ($r_{spear} = 0.55$)). However, it should be noted that those rankings include income and costs of housing and therefore do not measure pure QoL as defined in this paper. The structure of the "ZDF Deutschland-Studie 2018" allows for excluding housing and labor market conditions, such that I am able to extract a regional ranking only based on the categories health and security as well as leisure and nature. The resulting ranks are slightly stronger correlated with my QoL measure after excluding labor and housing indicators ($r_{spear} = 0.57$)).[31]

In the next section, the calculated QoL differences are used to identify (statistically) relevant amenity indicators by using supervised statistical learning methods.

1.6 Determinants of Quality of Life

This section aims to investigate which (category) of indicators is relevant to predict the QoL differentials calculated in the previous section. I want to emphasize that the results presented in this section do not allow for any causal interpretation and represent purely statistical relations. As described in section 1.3.2, there are mainly two approaches to derive the (relative) importance of single indicators and categories of indicators. First, I use PCA

and west have to be interpreted with caution.

[31]The almost unchanged degree of correlation when excluding labor and housing market conditions is not very surprising, as the ZDF ranking remains nearly the same. In line with that, the ranking has been criticized for arbitrary weighting and high level of correlation between included indicators: https://www.rwi-essen.de/presse/wissenschaftskommunikation/unstatistik/archiv/2019/detail/gelsenkirchen-401, last accessed 25.01.23.

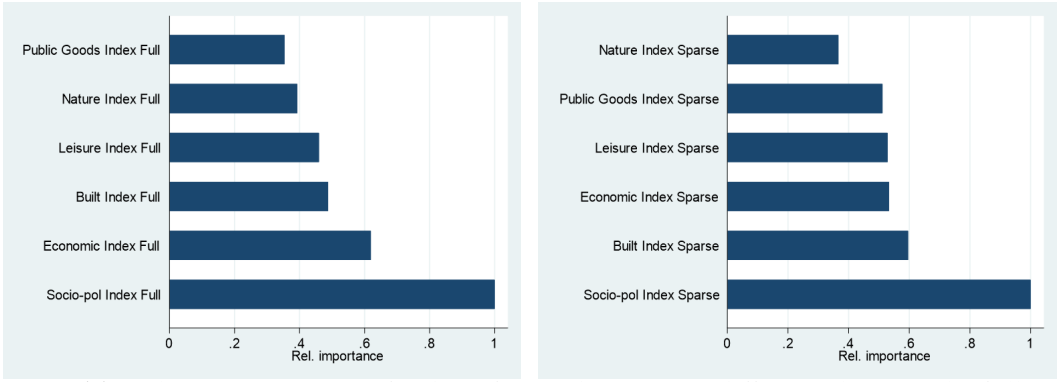
to derive indices of each of the six environments proposed by Lambiri et al. (2007). These indices are fed to a random forest algorithm to identify more and less relevant categories for the prediction of QoL differences. Second, I use Lasso regression to drop statistically irrelevant variables to learn more about single indicators. The remaining indicators are again used in a random forest regression. Both is done using (1) the full set of indicators without any pre-selection based on economic intuition and (2) using a sparse set of indicators dropping the indirect measures marked with a + in Table 1.1.

1.6.1 The relative importance of categories

Following Diamond (2016), I use PCA to derive an index for each of the six categories, namely natural environment, build environment, socio-political environment, economic environment, cultural and leisure environment, and public policy environment. For each category, I take the estimated loads of the first component and multiply them with the standardized indicators that are part of the respective environment as classified in Table 1.1. The six resulting indices are then used as inputs in a random forest regression of the form discussed in section 1.3.2. For the full set of indicators, the number of iterations is set to 500 and the number of variables considered at each split is set to 5. For the sparse set of indicators, tuning variable choices are similar. A graphical representation of the tuning process is provided in section 1.F in the Appendix. As the number of inputs is relatively low, the prediction error (i.e. the out-of-the-bag MSE) is relatively high for both sets. A detailed discussion of the prediction performance can be found in section 1.6.3. The resulting relative importance is represented in Figure 1.3. The left-hand side ranks the indices of the environments where all indicators are included in each index. The right-hand side shows the results for the set of indicators excluding redundant indirect indicators. For both graphs, the relative importance is the total sum of the improvement in the objective function given in the splitting criterion of all splits of a tree and across all trees, where the variable with the highest importance is set to 1. Although two pairs of indices switch ranks, the overall picture looks similar. The socio-political Index is the most important category, being almost twice as important as the next important environment. The second and third

Figure 1.3: Relative importance of categories

(a) Full set (b) Sparse set



Notes: (a): Relative importance of indices for prediction QoL differentials using the full set of indicators from column 3 of Table 1.1. (b): Relative importance of indices for prediction QoL differentials using all indicators of column 3 of Table 1.1 excluding indirect indicators marked with a +.

important indices are the economic and the built index. Both consist of many control indicators and do not represent "true" amenities according to some definitions. Instead, the relative importance of those categories indicates that my QoL measure does not incorporate all relevant differences in the labor and housing markets. Further, the fact that both indices rank higher than the leisure, nature, and the public goods index emphasizes that the relevance of the three latter environments is sometimes overestimated when not properly controlling for structural differences between local labor and housing markets. Nevertheless, it should be mentioned that every category receives a relative importance score of at least 35%, indicating that no environment is entirely irrelevant to predict QoL variation.

1.6.2 The relative importance of indicators

To learn more about the relevance of single indicators, adaptive Lasso is used to drop indicators with low predictive power before applying random forest. The optimal penalty term λ derived by adaptive cross-validation is .0005 for the full and .0023 for the sparse specification. The higher penalty term in the latter regression means that more variables are dropped, although the number of inputs was already reduced from 109 (full) to 88 (sparse). The selected variables are listed in Table 1.2. Column 2 represents the selected

indicators of all indicators, and column 3 lists the surviving variables when indirect indicators are eliminated upfront.

Table 1.2: Variable selection by Lasso

Indicator	Adaptive Lasso full	Adaptive Lasso sparse
(1) **Natural environment**		
precipitation	x	x
temperature	x	x
pm10_mean	x	
pm10_days		x
no2_mean		x
slope		x
agric_space	x	
water	x	x
east^{++}	x	x
(2) **Built environment**		
renewable		x
rooms^{++}	x	x
rural^{+}	x	
habitat_dens^{+}	x	
(3) **Socio-political environment**		
council_wshare	x	x
marriage	x	x
divorce	x	
age_avg^{++}		x
turnout	x	x
sportsclub		x
(4) **Economic environment**		
unemployment_long		x
vacancy_assist^{+}	x	
poverty	x	x
privat_dept	x	x
work_creative		x
work_knowledge^{+}	x	
gender_wagegap	x	x
gender_empgap	x	x
sector_prim^{++}	x	x
sector_ter^{++}	x	x
firms_large^{++}	x	x

Continued on next page

Table 1.2 – *Continued from previous page*

Indicator	Adaptive Lasso full	Adaptive Lasso sparse
firms_tiny^{++}	x	x
firms_birth^{++}	x	
land_tax_B	x	x
public_dept	x	
public_empl	x	x
living_space^{++}	x	x
(5) **Cultural & leisure environment**		
museums	x	
recre_space	x	
(6) **Public policy environment**		
access_pharma	x	x
crime	x	x
road_casual	x	x
grad_non	x	x
grad_low^{++}		x
grad_med^{++}	x	x
schools	x	
kindergardens	x	x
access_airport	x	x
access_bigcity^{++}	x	x
access_grocery	x	
cars	x	x

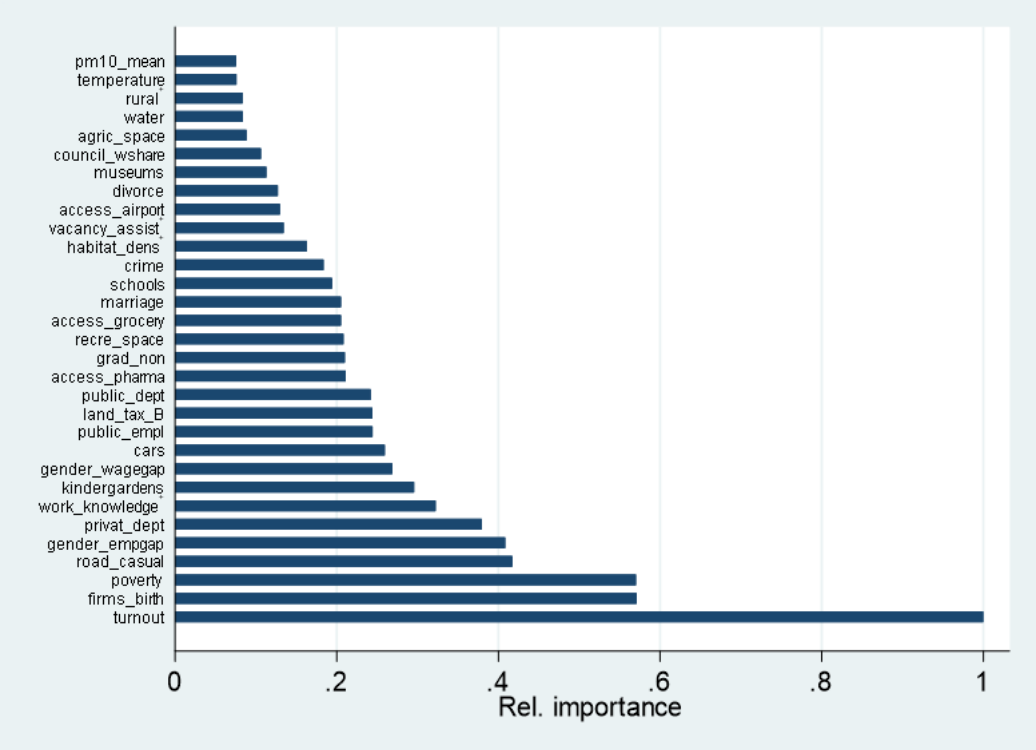
Notes: Indicators selected by adaptive Lasso regression. Column 2 represents selected indicators of the full list of all indicators. Column 3 shows the selected variables of the sparse list excluding indirect indicators marked with a + in column 3 of Table 1.1.

Overall, the selection of variables between the two specifications is very similar, considering that indirect indicators can not be included in the sparse specification by definition. This confirms that variable selection by adaptive Lasso regression is very consistent, even in the presence of a high correlation between inputs. It is visible that some categories are underrepresented after the variable elimination. In most categories, roughly 30%-40% of the included indicators survived. The two exceptions are the economic environment, where 45.5% (full) to 50% (sparse) of the variables are chosen, and the cultural and leisure categories, from which only 28.57% (full) to 0%

(sparse) of the inputs are selected by Lasso.

However, it should be noted that this does not incorporate the importance of each indicator or category. Hence, I again use the random forest algorithm to derive the relative importance of indicators that survived the Lasso selection in the next step. The results for the specification initially involving all indicators are visualized in Figure 1.4. Note that random forest does not drop variables like Lasso. Nevertheless, the plot includes only indicators with a relative importance of at least 0.1 and no control indicators for comprehensibility. In line with the results from the PCA specification, the most

Figure 1.4: Relative importance of indicators (Lasso full)



Notes: Relative importance of indicators for predicting QoL differentials using the full set of indicators from column 3 of Table 1.1. Control indicators are included in the regression but not displayed in this graph. For a graph including the control indicators please see Figure 1.G.20 in the Appendix.

important predictor is turnout in the federal election, an indicator from the socio-political category. Other variables from this group, the number of new marriages per resident and the share of women in the city council belong to

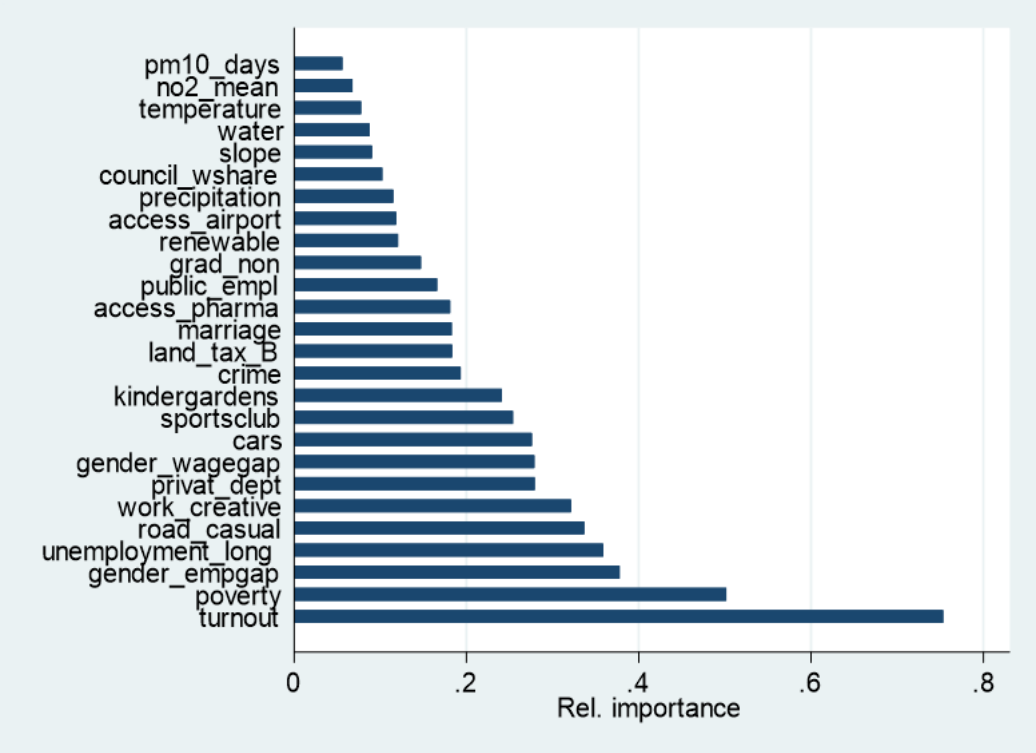
the relevant inputs as well but are ranked significantly lower. As expected, many economic indicators are relatively important as well. For example, two measures that can be related to innovativness, the number of new firms per resident and the employment share in knowledge-intensive firms, are among the top 10. The same is true for two indicators concerning gender equality in the labor market, the gender wage gap and the gender (full-time) employment gap. It should be noted here that the classification of indicators, for example of the two gender variables into the economic instead of the socio-political category, has no impact on this result and is only used to make interpreting the results more comprehensible. On the lower end, the two indicators of the culture and leisure environment, the number of museums per capita and the land share of recreation areas, have a relative importance of only 11.31% and 20.86%, respectively.

Similarly, Figure 1.5 plots the relative importance of predictors for the sparse specification. Again, control variables are not displayed for comprehensibility. Turnout is still top of the list, however not being the most important predictor as the east dummy has an even higher relevance in this specification. Instead, an additional social indicator, the share of residents being members of a sports club, entered the ranking. Overall, the results are quite similar, which is not too surprising given the fact that Lasso selected a relatively similar list of indicators. That means that the exclusion of indirect indicators based on economic intuition instead of statistical methods has no strong effect on the relative importance of the remaining predictors. In fact, as I will show in the next section, the predictive performance has even increased.

1.6.3 Prediction performance

To compare the performance of all versions of the random forest regression above, I use the out-of-bag prediction errors of the optimized algorithms displayed in Table 1.3.
Not surprisingly, the PCA specifications perform worse in terms of the out-of-bag prediction error. Reducing the dimension of the predictor space to only 6 dimensions takes away a lot of the flexibility of the algorithm. However, this exercise did not aim to maximize the predictive power but rather to shed

Figure 1.5: Relative importance of indicators (Lasso sparse)



Notes: Relative importance of indicators for predicting QoL differentials using all indicators of column 3 of Table 1.1 excluding indirect indicators marked with a +. Control indicators are included in the regression but not displayed in this graph. For a graph including the control indicators please see Figure 1.G.21 in the Appendix.

Table 1.3: Prediction performance and parsimony

	Full	Sparse	Lasso full	Lasso sparse	PCA full	PCA sparse
OOB error	.0145	.0148	.0153	.0152	.0201	.0197

Notes: Out-of-bag error (OOB error) for all six specifications discussed above. "Full" refers to specifications including all indicators of column 3 of Table 1.1. "Sparse" labels specifications where indirect indicators marked with a + in column 3 of Table 1.1 are excluded upfront. The first two columns are specifications where the respective list of indicators is directly used by the random forest algorithm. Columns 3 and 4 refer to specifications with pre-selection of variables using Lasso regression from subsection 1.6.2. The last two columns represent both specifications involving PCA discussed in subsection 1.6.1.

light on the relative importance of categories. Nevertheless, one striking fact is also true for the PCA specifications: the estimations using only the sparse list of indicators have lower errors in all of the three approaches. The typical trade-off between parsimony and accuracy seems not to apply to the specifications here. One potential explanation is that the problem of overfitting the training data is still prevalent, despite using bagging. To provide additional evidence, I rerun the random forest algorithm with only the two inputs that have the strongest correlation with the QoL differential of each of the six main categories. The resulting out-of-bag prediction error is 0.0181. Using the ten indicators with the highest relative importance in the sparse lasso specification results in an out-of-bag prediction error of 0.0152, which is in the same region as the error using more than 80 indicators.[32] That suggests that using only a limited number of indicators can be preferable not only to avoid problems of data availability but also to improve the quality of describing QoL differences.

1.7 Conclusion

In this paper, I apply statistical learning methods to identify the most important determinants of Quality of Life (QoL) differences between German counties. First, I derive a cardinal measure of regional QoL in 2017 based on the well-established Rosen-Roback framework incorporating non-labor income and federal taxation. Second, I identify potential amenity indicators previously used in the academic literature and in non-academic city and region rankings. Based on that, I collect and construct over 100 indicators using various sources. This provides an overview for future research on what measures of local amenities are available in the German context and where to obtain the data. Third, I apply a least absolute shrinkage and selection operator (Lasso) regression and a random forest algorithm to identify the indicators that are most relevant for predicting the QoL differences. To investigate the relative importance of groups of indicators, I generate indices for six different broad categories proposed by Lambiri et al. (2007) using

[32] Including control indicators, the ten most important variables of the sparse lasso specification are *east*, *turnout*, *sector_ter*, *poverty*, *firm_tiny*, *gender_empgap*, *unemployment_long*, *road_casual*, *work_creative* and *firms_large*.

principal component analysis (PCA). These six categories are (1) natural environment, (2) built environment, (3) socio-political environment, (4) local economic environment, (5) cultural and leisure environment, and (6) public policy environment.

My results indicate that socio-political indicators as well as labor market characteristics are the most important determinants of QoL differentials. Namely, the turnout at federal elections, poverty, and local informativeness stood out as predictors. The role of turnout is especially important to note, as previous urban economics literature widely ignored social indicators as relevant determinants of local well-being. However, newly constructed more direct indicators of social interaction, the share of residents that are members of a local sports club and the Social Connectedness Index from Facebook, are not capable of replacing election turnout as a relevant predictor.

I also show that reducing the number of included indicators upfront using economic reasoning instead of statistical methods can improve prediction accuracy. Additionally, using only the ten best predictors identified with statistical learning methods produce a prediction accuracy similar to specifications with over 80 or using the full set of over 100 indicators. This result provides evidence that there is no need to collect a large number of amenity indicators to predict or test estimated QoL differentials. However, I suggest including measures representing six different broad categories, since my results indicate that each category has substantial importance in predicting the spatial variation of QoL.

All methods applied in this paper do not allow to draw any causal conclusions about the role of single amenities for local well-being. Instead, I provide guidance on how to decide which and how many amenity indicators to include for future research. This includes papers investigating the causal impact of specific amenities as well as papers building on spatial general equilibrium models that use amenity indicators to validate their model via over-identification.

Future research on this topic could improve in deriving more accurate QoL differentials. First, the problem of spatial correlation for example by commuting can be alleviated by aggregating data on labor market regions. However, aggregating data reduces the number of regions (i.e. observations), which makes prediction even more difficult, especially with a large number

of independent variables. My results can help to tackle this problem by pre-selecting a parsimonious model with a reduced number of indicators. Second, using individual income data, for example by using the Sample of Integrated Labour Market Biographies provided by the Institute for Employment Research (IAB), allows to better control for structural differences in labor markets directly when estimating income differentials. This would yield a more accurate QoL ranking and allows to drop controls in the prediction which are typically not considered as amenities (like the sectoral composition). Third, using individual income data would allow to differentiate between different groups to investigate heterogeneity in QoL and amenity valuation. The differentiation could be done along various dimensions, including education, age, income, or family status (i.e. having children).

Bibliography

Ahlfeldt, G. M., Bald, F., Roth, D., and Seidel, T. (2020). Quality of Life in a Dynamic Spatial Model. *CEPR Discussion Paper*, 15594.

Ahlfeldt, G. M. and Pietrostefani, E. (2019). The economic effects of density: A synthesis. *Journal of Urban Economics*, 111:93–107.

Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., and Wolf, N. (2015). The Economics of Density: Evidence From the Berlin Wall. *Econometrica*, 83(6):2127–2189.

Albouy, D. (2008). Are Big Cities Bad Places to Live? Estimating Quality of Life across Metropolitan Areas. *NBER working paper*, 14472.

Allen, T. and Arkolakis, C. (2014). Trade and the Topography of the Spatial Economy. *The Quarterly Journal of Economics*, 129(3):1085–1140.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, 32(3):259–280.

Basuchoudhary, A., Bang, J. T., and Sen, T. (2017). *Machine-learning Techniques in Economics: New Tools for Predicting Economic Growth*. Springer.

Bayer, P., Ferreira, F., and McMillan, R. (2007). A Unified Framework for Measuring Preferences for Schools and Neighborhoods. *Journal of Political Economy*, 115(4):588–638.

Beeson, P. E. (1991). Amenities and regional differences in returns to worker characteristics. *Journal of Urban Economics*, 30(2):224–241.

Biagi, B., Ladu, M. G., and Meleddu, M. (2018). Urban Quality of Life and Capabilities: An Experimental Study. *Ecological Economics*, 150:137–152.

Biagi, B., Ladu, M. G., Meleddu, M., and Royuela, V. (2020). Tourism and the city: The impact on residents' quality of life. *International Journal of Tourism Research*, 22(2):168–181.

Black, D., Kolesnikova, N., and Taylor, L. (2009). Earnings Functions When Wages and Prices Vary by Location. *Journal of Labor Economics*, 27(1):21–47.

Blomquist, G. C. (2006). Measuring Quality of Life. In Arnott, R. J. and McMillen, D. P., editors, *A Companion to Urban Economics*, pages 483–501. Blackwell Publishing Ltd.

Blomquist, G. C., Berger, M. C., and Hoehn, J. P. (1988). New Estimates of Quality of Life in Urban Areas. *The American Economic Review*, 78(1):89–107.

Boelmann, B. and Schaffner, S. (2019). FDZ Data description: Real-Estate Data for Germany (RWI-GEO-RED) - Advertisements on the Internet Platform ImmobilienScout24). Technical report, RWI Leibniz-Institut für Wirtschaftsforschung, Essen.

Boyer, R., Boyer, R., and Savageau, D. (1985). *Places rated almanac: Your guide to finding the best places to live in America*. Rand McNally, Chicago.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Buettner, T. and Ebertz, A. (2009). Quality of Life in the Regions: Results for German Counties. *The Annals of Regional Science*, 43(1):89–112.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer, Berlin.

Burnell, J. D. and Galster, G. (1992). Quality-of-life Measurements and Urban Size: An Empirical Note. *Urban Studies*, 29(5):727–735.

Campbell, A., Converse, P. E., and Rodgers, W. L. (1976). *The Quality of American Life: Perceptions, Evaluations, and Satisfactions*. Russell Sage Foundation, New York.

Carlino, G. A. and Saiz, A. (2019). Beautiful city: Leisure amenities and urban growth. *Journal of Regional Science*, 59(3):369–408.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., Barberá, P., Bhole, M., and Wernerfelt, N. (2022a). Social capital I: measurement and associations with economic mobility. *Nature*, 608(7921):108–121.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., Johnston, D., Koenen, M., Laguna-Muggenburg, E., Mudekereza, F., Rutter, T., Thor, N., Townsend, W., Zhang, R., Bailey, M., Barberá, P., Bhole, M., and Wernerfelt, N. (2022b). Social capital II: determinants of economic connectedness. *Nature*, 608(7921):122–134.

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2021). On cross-validated Lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317.

Costanza, R., Fisher, B., Ali, S., Beer, C., Bond, L., Boumans, R., Danigelis, N. L., Dickinson, J., Elliott, C., and Farley, J. (2008). An Integrative Approach to Quality of Life Measurement, Research, and Policy. *Surveys and Perspectives Integrating Environment and Society*, 1(1):16–21.

Cragg, M. and Kahn, M. (1997). New Estimates of Climate Demand: Evidence from Location Choice. *Journal of Urban Economics*, 42(2):261–284.

Diamond, R. (2016). The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000. *American Economic Review*, 106(3):479–524.

Douglas, S. (1997). Estimating Relative Standard of Living in the United States Using Cross-Migration Data. *Journal of Regional Science*, 37(3):411–436.

Duranton, G. and Puga, D. (2020). The Economics of Urban Density. *Journal of Economic Perspectives*, 34(3):3–26.

DWD Climate Data Center (CDC) (2018a). Vieljährige mittlere Raster der Lufttemperatur (2m) für Deutschland 1981-2010. Version v1.

DWD Climate Data Center (CDC) (2018b). Vieljähriges Mittel der Raster der Niederschlagshöhe für Deutschland 1981-2010. Version v1.

DWD Climate Data Center (CDC) (2018c). Vieljähriges Mittel der Raster der Sonnenscheindauer für Deutschland 1981-2010. Version v1.

Faller, B., Hellbach, C., and Vater, A. (2009). Möglichkeiten Zur Bildung Eines Regionalindex Wohnkosten Unter Verwendung Von Angebotsdaten. *German Council for Social and Economic Data (RatSWD) Research Notes*, (34).

Florida, R. (2002). Bohemia and economic geography. *Journal of Economic Geography*, 2(1):55–71.

Glaeser, E. L., Kolko, J., and Saiz, A. (2001). Consumer City. *Journal of Economic Geography*, 1(1):27–50.

Gyourko, J., Kahn, M., Tracy, J. B. T. H. o. R., and Economics, U. (1999). Quality of Life and Environmental Comparisons. In *Applied Urban Economics*, volume 3, chapter 37, pages 1413–1454. Elsevier.

Gyourko, J. and Tracy, J. (1991). The Structure of Local Public Finance and the Quality of Life. *Journal of Political Economy*, 99(4):774–806.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2015a). *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York, 2 edition.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015b). Statistical Learning with Sparsity. *Monographs on Statistics and Applied Probability*, 143.

Herdagdelen, A., Adamic, L., and State, B. (2022). Community gifting groups on Facebook. *arXiv working paper*, 2211.09043.

Hiller, N. and Lerbs, O. (2015). The capitalization of non-market attributes into regional housing rents and wages: evidence on German functional labor market areas. *Review of Regional Research*, 35(1):49–72.

Hsieh, C.-m. (2003). Counting Importance: The Case of Life Satisfaction and Relative Domain Importance. *Social Indicators Research*, 61(2):227–240.

INKAR (2021). *Indikatoren und Karten zur Raum- und Stadtentwicklung*. Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) im Bundesamt für Bauwesen und Raumordnung (BBR), Bonn, ausgabe 20 edition.

Joint Economic Committee (2018). The Geography of Social Capital in America. Technical Report 1-18, SCP Report (Social Capital Project).

Kahneman, D., Diener, E., and Schwarz, N., editors (1999). *Well-Being: Foundations of Hedonic Psychology*. Russell Sage Foundation, New York.

Lambiri, D., Biagi, B., and Royuela, V. (2007). Quality of Life in the Economic and Urban Economic Literature. *Social Indicators Research*, 84(1):1.

Lee, S. (2010). Ability sorting and consumer city. *Journal of Urban Economics*, 68(1):20–33.

Lee, S., Lee, S. H., and Lin, J. (2021). The Well-Being of Nations: Estimating Welfare from International Migration. *International Economic Review*, 62(3):1111–1130.

Marans, R. W. and Stimson, R. (2011). An Overview of Quality of Urban Life. In Marans, R. W. and Stimson, R. J., editors, *Investigating Quality of Urban Life: Theory, Methods, and Empirical Research*, pages 1–29. Springer Netherlands, Dordrecht.

McFadden, D. (1974). The Measurement of Urban Travel Demand. *Journal of Public Economics*, 3(4):303–328.

Meinshausen, N. and Bühlmann, P. (2006). High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462.

Nakajima, K. and Tabuchi, T. (2011). Estimating Interregional Utility Differentials. *Journal of Regional Science*, 51(1):31–46.

Okulicz-Kozaryn, A. and Valente, R. R. (2019). Livability and Subjective Well-Being Across European Cities. *Applied Research in Quality of Life*, 14(1):197–220.

Peichl, A., Pestel, N., and Siegloch, S. (2013). Ist Deutschland wirklich so progressiv? Einkommensumverteilung im europäischen Vergleich. *Vierteljahrshefte zur Wirtschaftsforschung*, 82(1):111–127.

Putnam, R. D. (2000). *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster, New York.

Redding, S. J. and Rossi-Hansberg, E. (2017). Quantitative Spatial Economics. *Annual Review of Economics*, 9(1):21–58.

Roback, J. A. (1980). The Value of Local Urban Amenities: Theory and Measurement. *Ph.D. dissertation*.

Roback, J. A. (1982). Wages, Rents, and the Quality of Life. *Journal of Political Economy*, 90(6):1257–1278.

Rosen, S. (1979). Wage-based indexes of urban quality of life. *Current issues in urban economics*, pages 74–104.

Rupasingha, A., Goetz, S. J., and Freshwater, D. (2006). The production of social capital in US counties. *The Journal of Socio-Economics*, 35(1):83–101.

RWI and ImmobilienScout24 (2023a). RWI Real Estate Data - Hauskauf. RWI-GEO-RED. Version: 1. *RWI – Leibniz Institute for Economic Research. Dataset*.

RWI and ImmobilienScout24 (2023b). RWI Real Estate Data - Wohnungskauf. RWI-GEO-RED. Version: 1. *RWI – Leibniz Institute for Economic Research. Dataset*.

RWI and ImmobilienScout24 (2023c). RWI Real Estate Data - Wohnungsmiete. RWI-GEO-RED. Version 1. *RWI – Leibniz Institute for Economic Research. Dataset*.

Savageau, D. (2007). *Places Rated Almanac: The Classic Guide for Finding Your Best Places to Live in America*. Places Rated Books Llc, Washington D.C., 7 edition.

Schaeffer, Y. and Dissart, J.-C. (2018). Natural and Environmental Amenities: A Review of Definitions, Measures and Issues. *Ecological Economics*, 146:475–496.

Shapiro, J. M. (2006). Smart Cities: Quality of Life, Productivity, and the Growth Effects of Human Capital. *The Review of Economics and Statistics*, 88(2):324–335.

Smith, D. M. (1973). *The Geography of Social Well-being in the United States An Introduction to Territorial Social Indicators*. McGraw Hill, New York.

Sollis, K., Yap, M., Campbell, P., and Biddle, N. (2022). Conceptualisations of wellbeing and quality of life: A systematic review of participatory studies. *World Development*, 160:106073.

Statistisches Bundesamt (Destatis) (2017). Statistischer Jahrbuch 2017. Technical report.

Statistisches Bundesamt (Destatis) (2020). Einkommens- und Verbrauchsstichprobe 2018. Technical report.

Thorndike, E. L. (1939). *Your City*. Harcourt Brace College Publishers, New York.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (1997). The Lasso Model for Variable Selection in the Cox Model. *Statistics in Medicine*, 16(4):385–395.

Wall, H. J. (2001). Voting with your feet in the United Kingdom: Using cross-migration rates to estimate relative living standards. *Papers in Regional Science*, 80(1):1–23.

Weinand, S. and von Auer, L. (2020). Anatomy of regional price differentials: evidence from micro-price data. *Spatial Economic Analysis*, 15(4):413–440.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Appendix

1.A Data

In this section, I provide detailed information on each indicator including a precise description, how it was obtained/calculated, and some information on the source or alternative sources. The order of indicators will be the same as in Table 1.1.

Climate

All variables on climate were provided by the *Deutscher Wetterdienst* (DWD) and are visualized in Figure 1.A.1.

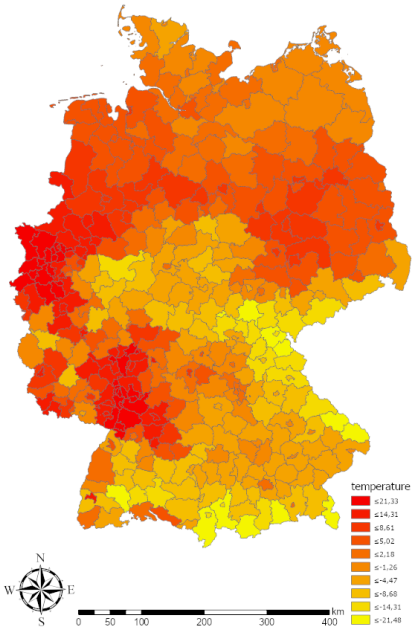
Average temperature (variable name *temperature*) is the average temperature in Celsius degrees of the monthly averages of 1981 - 2010, geocoded on a 1km x 1km grid level (DWD Climate Data Center (CDC), 2018a). Over 500 measuring stations collect data on the air temperature two meters above the ground. Measures for grids without stations are obtained by interpolation by the DWD. I aggregated the average temperature on the county level by taking the mean of all grids that are (entirely) within the counties' borders.

Precipitation (variable name *precipitation*) is the average amount of rainfall in millimeters of the yearly averages of 1981 - 2010, geocoded on a 1km x 1km grid level (DWD Climate Data Center (CDC), 2018b). Measures for grids without measuring stations are obtained by interpolation by the DWD. I aggregated the average precipitation on the county level by taking the mean of all grids that are (entirely) within the counties' borders.

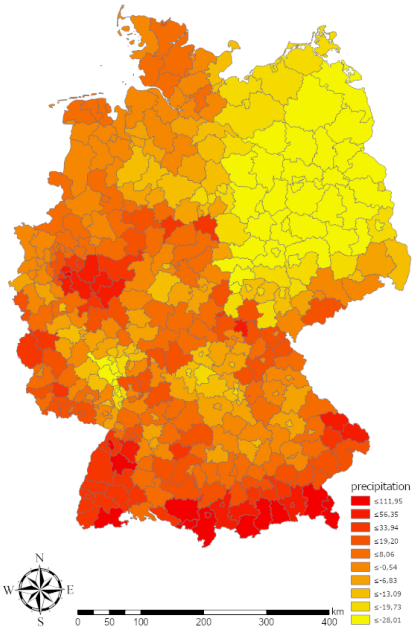
Sun hours (variable name *sunhours*) indicates the average duration of sunshine in hours of the yearly averages of 1981 - 2010, geocoded on a 1km

Figure 1.A.1: Climate

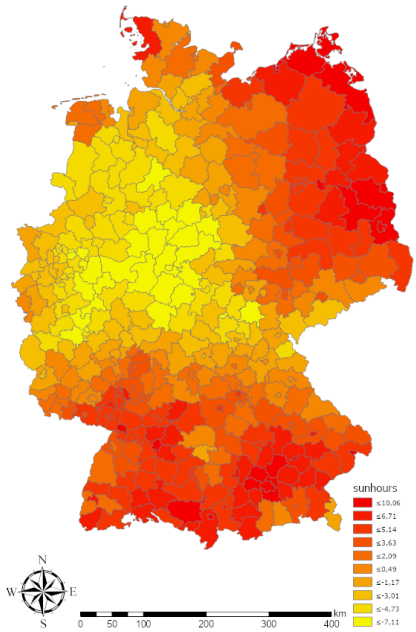
(a) Average temperature



(b) Precipitation



(c) Sun hours



Notes: All values are deviations from the national mean in %.

x 1km grid level (DWD Climate Data Center (CDC), 2018c). Measures for grids without measuring stations are obtained by interpolation by the DWD. I aggregated the average sunshine duration on the county level by taking the mean of all grids that are (entirely) within the counties' borders.

Air quality

All variables on air quality were provided by the *Umweltbundesamt* (UBA). Figure 1.A.2 plots four of the measures.

Particulate (PM_{10}) emission is represented by two variables, *pm10_mean* and *pm10_days*. *pm10_mean* measures the average concentration of particulates with diameter $< 10\mu m$ in $\mu g/m^3$ in 2017, geocoded on a 2km x 2km grid level. The UBA corrects the measures for biases by non-representative measure station locations and interpolates data for grids without stations. I aggregated the concentration on the county level by taking the mean of all grids that are (entirely) within the counties' borders. *pm10_days* is the number of days where the PM_{10} concentration exceeds the legal limit of 50 $\mu g/m^3$.

Particulate ($PM_{2.5}$) emission (variable name *pm25_mean*) measures the average concentration of particulates with diameter $< 2.5\mu m$ in $\mu g/m^3$ in 2017, geocoded on a 2km x 2km grid level. The UBA corrects the measures for biases by non-representative measure station locations and interpolates data for grids without stations. I aggregated the concentration on the county level by taking the mean of all grids that are (entirely) within the counties' borders.

Nitrogen dioxide (NO_2) emission (variable name *no2_mean*) measures the average concentration of nitrogen dioxide in $\mu g/m^3$ in 2017, geocoded on a 2km x 2km grid level. The UBA corrects the measures for biases by non-representative measure station locations and interpolates data for grids without stations. I aggregated the concentration on the county level by taking the mean of all grids that are (entirely) within the counties' borders.

Ozone (O_3) emission (variable name *o3_days*) measures the number of days the nitrogen dioxide concentration exceeded the legal limit of 120 $\mu g/m^3$ in 2017, geocoded on a 2km x 2km grid level. The UBA corrects the measures for biases by non-representative measure station locations and

interpolates data for grids without stations. I aggregated the number of days on the county level by taking the mean of all grids that are (entirely) within the counties' borders. In contrast to the emissions listed above, ozone is usually not directly emitted but depends mostly on weather conditions (e.g. sun exposure, especially with low levels of protection by the ozone layer).

Natural environment

Variables describing the natural environment were obtained from the *Digitales Geländemodell 200m* (DGM200) by *Bundesamt für Kartographie and Geodäsie* (BKG), *OpenStreetMap* (OSM), and INKAR (2021).[33] Maps including some of the variables are presented in Figure 1.A.3.

Distance from coast is measured by the distance of each municipality's centroid to the closest coastline identified in OSM (in km). Then, I took the mean of the distances of all municipalities within one county. Note distances were calculated by using not only German but also international coastlines like the ones in the Netherlands and Italy. OSM shapefiles were downloaded from `https://www.geofabrik.de/de/index.html` and date to 21.05.2019.

Average slope (variable name *slope*) is calculated using DGM200.[34] The DGM200 provides information on elevation on the 200m x 200m grid level. The slope of each grid is obtained by taking the average steepness to all eight surrounding grid cells (in incline degrees). If data on elevation is missing for more than two of the neighboring grid cells, no slope is calculated, which is especially the case at the national German border. Finally, I took the mean of each grid's slope that is (entirely) within the counties' borders to obtain the average slope of each county. Therefore, I do not distinguish between regions with moderate steepness in all grid cells and counties with high incline degrees in only a few locations of the area.

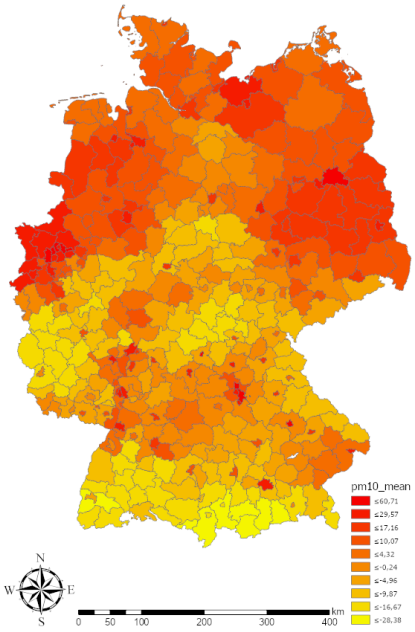
The **% of bodies of water** (variable name *water*) of a county's area is directly obtained from INKAR (2021). The information provided by INKAR (2021) is based on the *Flächenerhebung nach Art der tatsächlichen Nutzung des Bundes und der Länder* and considers all areas covered by water, including rivers, lakes, and harbor basins in 2017. The sea is not taken into account.

[33] Information on land use is also available in the *Regionaldatenbank Deutschland* (RDB) by *Statististische Ämter des Bundes und der Länder.*

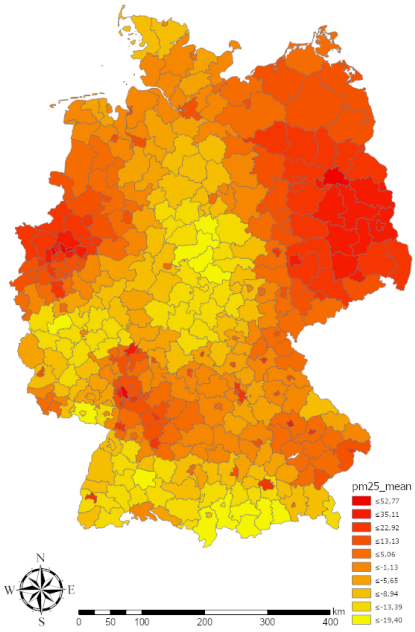
[34] DGM200 was downloaded from `www.geodatenzentrum.de`. ©GeoBasi-DE / BKG 2019

Figure 1.A.2: Air quality

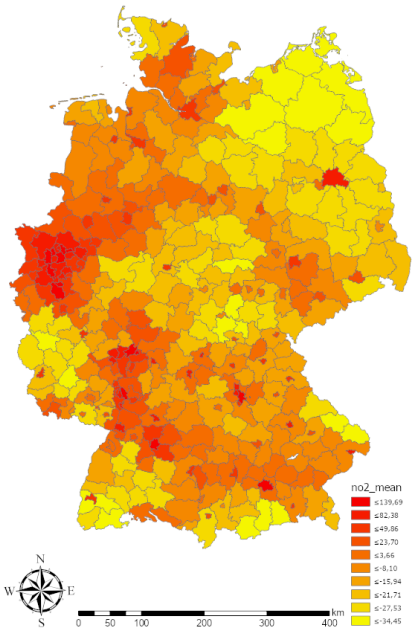
(a) Particulate (PM_{10}) emission



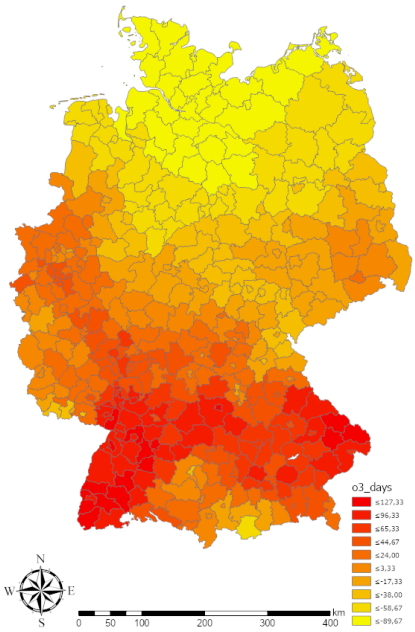
(b) Particulate ($PM_{2.5}$) emission



(c) Nitrogen dioxide (NO_2) emission



(d) Ozone (O_3) emission



Notes: All values are deviations from the national mean in %.

This measure is independent of how the body of water is used and could therefore represent an economic factor if one considers access to water trade

routes or a recreational factor when considering lakes to swim in. Also, especially smaller bodies of water like lakes or ponds in parks are often not of natural origin. Still, when looking at the total share of space covered by water, I assume the majority to be natural bodies of water or at least the endogenous variation over time to be very limited. Hence, **% of bodies of water** is part of the natural environment sub-category.

The **% natural spaces** (variable name *nat_space*) of a county's area is also provided by INKAR (2021) and based on the *Flächenerhebung nach Art der tatsächlichen Nutzung des Bundes und der Länder*. In contrast to *water*, *nat_space* considers the function of (natural) spaces. For example, it also includes bodies of water like rivers and lakes, but no harbor basins since it is the goal to measure the area that is still in its original natural state in 2017. Agricultural spaces as well as land covered by forest are excluded, which are represented by their own variables (see below).

The **% of forest** (variable name *forest*) of a county's area is also provided by INKAR (2021) and based on the *Flächenerhebung nach Art der tatsächlichen Nutzung des Bundes und der Länder*. It measures the share of space covered by forests and groves in 2017.

The **% of open spaces** (variable name *openspace*) of a county's area is also provided by INKAR (2021) and based on the *Flächenerhebung nach Art der tatsächlichen Nutzung des Bundes und der Länder*. It represents the share of land that is not covered by construction in 2017. Besides agricultural spaces, forests, and water bodies, this includes the parts of settlement areas without buildings on them like parks or cemeteries.

The **& of agricultural spaces** (variable name *agric_space*) of a county's area is also provided by INKAR (2021) and based on the *Flächenerhebung nach Art der tatsächlichen Nutzung des Bundes und der Länder*. It measures the share of space covered by the area that is used for agricultural production in 2017 (which excludes e.g. agricultural buildings). As with other land use variables, agricultural space is not of natural origin and can be categorized as an economic indicator as well. However, I take the share of land used for agricultural production (and other land use indicators) as a measure of how the space appears to an average household living in that area. The economic effect of agriculture is supposed to be measured by the share of employment in the primary sector. This especially makes a difference when

comparing counties with a lot of animal husbandry with counties with filed based agriculture which can differ substantially in terms of land and labor intensity. In fact, the correlation coefficient between *agric_space* and the share of employment in the primary sectors is only 0.56. Hence, *agric_space* is part of the natural environment sub-category.

Location

The **East / West dummy** (variable name *east*) indicates whether a county was part of the German Democratic Republic (GDR) before reunification in 1990. This includes all counties in the so-called "new states": Brandenburg, Mecklenburg-West Pomerania, Saxony, Saxony-Anhalt, and Thuringia. The city of Berlin is a special case since it contains districts from both the former GDR and West Germany. I decided to count Berlin as a whole as "western".

Housing market

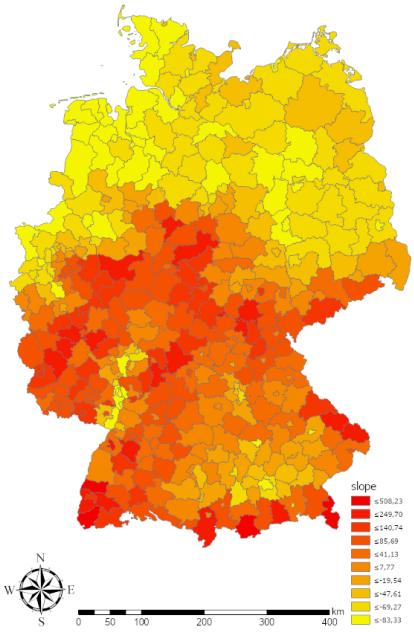
Information on the housing market is mainly provided by the FDZ Ruhr based on data from the online market platform ImmoScout24 (RWI and ImmobilienScout24, 2023b,a,c). A detailed data description can be found in Boelmann and Schaffner (2019). Free access to the data can be requested by researchers on the FDZ Ruhr homepage. Note that variables that were only used to calculate the cost of living differential in section 1.4.1 are not included here. Figure 1.A.4 visualizes that controlling for house characteristics is essential since there are systematic differences between regions in terms of the average floor space, the number of rooms, and the age of the housing units.

The **share of new flats with renewable energy** (variable name *renewable*) in 2017 measures how modern the change in the housing stock is. It is based on the *Statistik der Baufertigstellungen des Bundes und der Länder* and provided by INKAR (2021). More precisely, it is the share of all new housing units, not just flats, that can be heated with renewable energy like solar thermal, geothermal, and bioenergy.

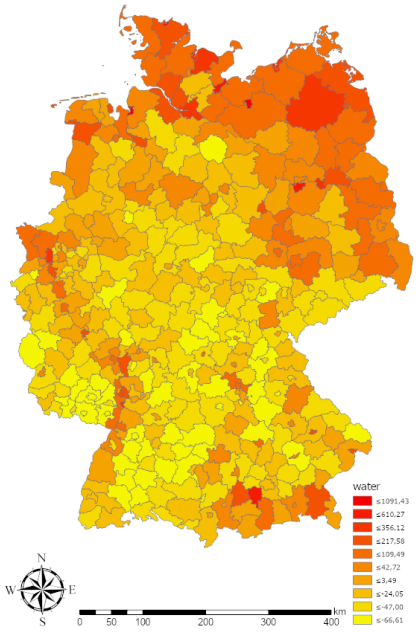
The **number of rooms per flat** (variable name *rooms*) in 2017 is calculated using RWI and ImmobilienScout24 (2023b) and RWI and ImmobilienScout24 (2023c). It represents differences in the housing stock in terms of the size of

Figure 1.A.3: Natural Environment

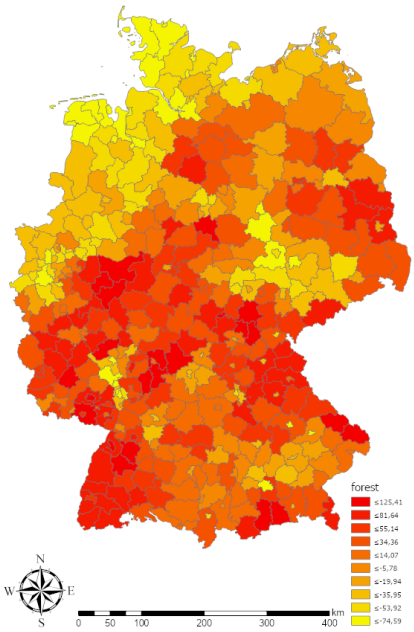
(a) Average slope



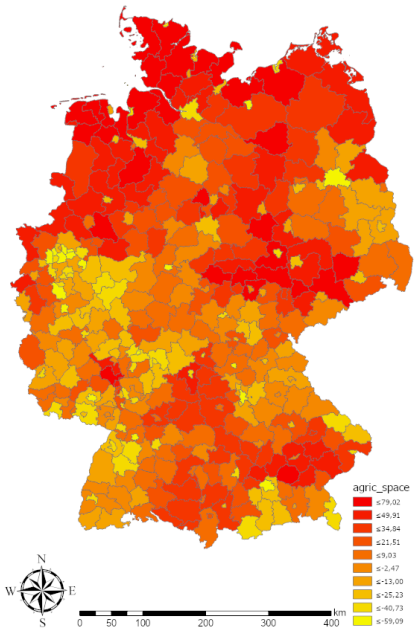
(b) Bodies of water



(c) Forest



(d) Agricultural spaces



Notes: All values are deviations from the national mean in %.

flats. The availability of larger flats can generate additional utility independent of the price per m^2.

Similarly to the number of rooms, the average **living space per flat** (variable name *floorspace*) in 2017 is a measure of the size of housing units in one location. It is also calculated using RWI and ImmobilienScout24 (2023b) and RWI and ImmobilienScout24 (2023c). As one can argue that the average floorspace is just an inverse measure of population density, I want to emphasize that the correlation coefficient with population density $r = -0.3$ is negative, but not close to 1.

The average **age of housing** units (variable name *housing_age*) in 2017 is calculated using purchase adverts from RWI and ImmobilienScout24 (2023b) and RWI and ImmobilienScout24 (2023a).[35] The age of housing can both reflect the state of modernization of the regional housing stock but also the availability of historic buildings, which might look more appealing.

Urbanization

The **total population** (variable name *pop_total*) in 2017 is based on the (updated) micro census from 2011, which is also the officially used population statistic and is provided by numerous sources including INKAR (2021) and *Regionaldatenbank Deutschland* (RDB). The population size is an important (indirect) indicator if there are amenities like, for example, theaters that require a certain number of potential customers. Many papers have already shown that consumption variety increases with the population size of a location.

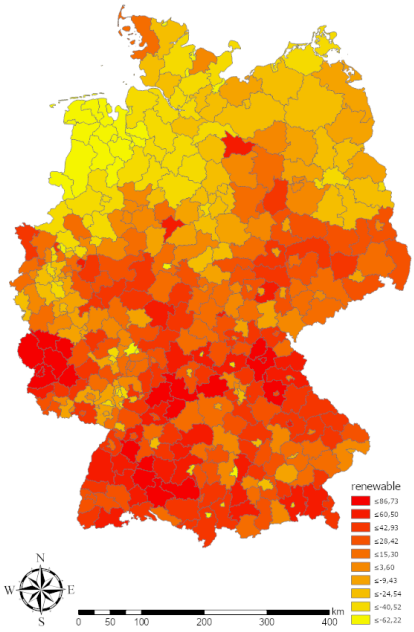
Population density (variable name *pop_dens*) is calculated by dividing the total population in 2017 by the total area of a county in km^2. Information on both is obtained from INKAR (2021) but is also available in the RDB. In contrast to total population, population density makes comparing regions of different land sizes more reasonable when it comes to the degree of urbanization.

Habitat density (variable name *habitat_dens*) divides the total population in 2017 only by the area used for settlements and transportation infrastructure and is provided by INKAR (2021). It is an attempt to obtain a more precise measure of experienced population density.

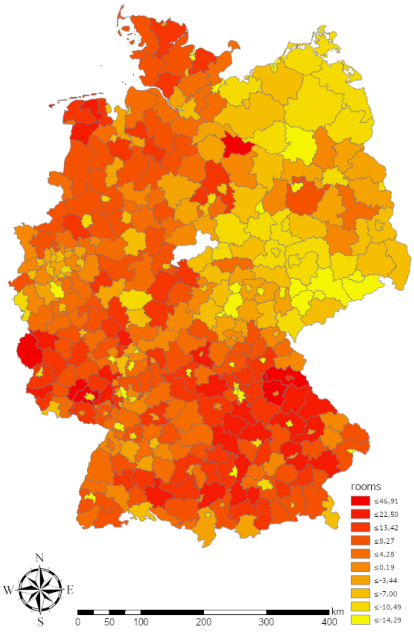
[35] Averts for rent from RWI and ImmobilienScout24 (2023c) are excluded here since information on the age of the whole housing unit is less reliable in this dataset.

Figure 1.A.4: Housing market

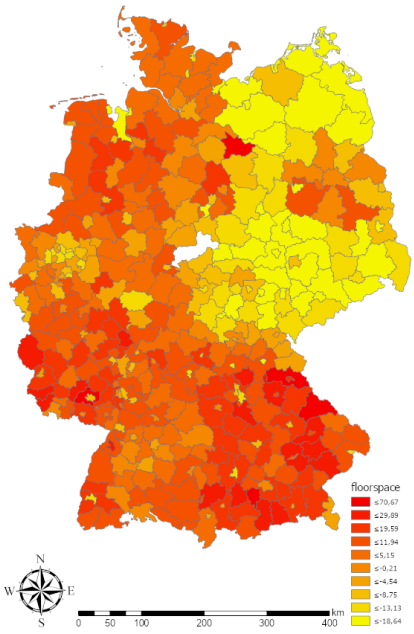
(a) Flats with renewable energy



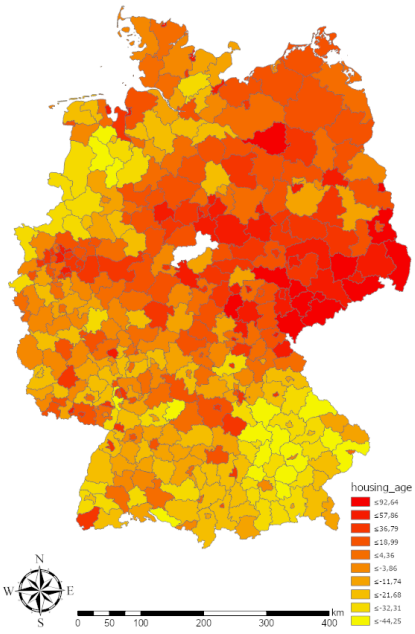
(b) Number of rooms



(c) Floor space



(d) Age of housing units



Notes: All values are deviations from the national mean in %.

The **degree of urbanization** is measured by a continuous variable (variable name *rural* obtain from INKAR (2021). It represents the share of the county's

residents that live in municipalities with a population density below 150 residents per km^2 in 2019. Data on population Hence, if *rural* is high it actually means that the degree of urbanization is low in that county. In contrast to raw population density, this measure considers more precisely how residents are distributed within the county.

Demographics

Indicators of demographic characteristics intend to measure how the neighbors or more generally the local society is perceived. All variables in the sub-category are obtained from INKAR (2021). Figure 1.A.5 plots two of the variables as an example.

The **female share of city council** (variable name *concil_wshare*)in 2017 from *Statistische Ämter des Bundes und der Länder - Ergebnisse der Kommunalwahlen* is my equivalent of a diversity measure. Florida (2002) proposes to measure diversity by an index of the gay population. His goal was to get a proxy not only for diversity but also for the openness of the local population. Therefore, I use the share of women in local councils and parliaments as a proxy for liberal and progressive values.

I calculate the average **population growth rate** (variable name *popgrowth_5y*) of the five years prior to 2017, 2012-2016, using the information on total population from INKAR (2021). This indicator does not only capture (perceived) dynamics of a region but is sometimes used as an indirect measure of overall attractiveness, as more appealing locations are expected to attract more people to move in and stay.

As population growth is also depending on the natural population flow by deaths and births, I also calculate the net **migration balance** (variable name *migrbalance_5y*) for the five years prior to 2017, 2012-2016. Using data on net migration from INKAR (2021) based on the *Wanderungsstatistik des Bundes und der Länder*, this indicator only measures the number number of immigrants minus the number of emigrants per 1000 inhabitants, excluding natural changes in the population size. Both intra-, as well as international migration, are considered.

Nevertheless, the number of birth can be interpreted as a proxy for how suitable a location is perceived to raise children. Hence, I also include the

fertility rate (variable name *fertility*) in 2017, provided by INKAR (2021) using *Statistik der Geburten des Bundes und der Länder*.

The **number of marriages** (variable name *marriage*) is defined as the number of new marriages per 1000 residents older than 18 years in 2017. It is based on the *Statistik der Eheschließungen des Bundes und der Länder* and provided by INKAR (2021). The geographic assignment to a county uses the location of the registry office.

Similar to the number of marriages, the **number of divorces** (variable name *divorce*) is defined as the number of divorces per 1000 residents older than 18 years in 2017. It is based on the *Statistik rechtskräftiger Urteile in Ehesachen des Bundes und der Länder* and provided by INKAR (2021). Divorces are geographically assignment to a county by the location of the responsible office of the family court. Both the number of marriages and the number of divorces are clearly influenced by other demographic characteristics like age and gender balance. However, it means that the likelihood of finding a partner may differ between counties. As a new marriage is usually associated with well-being (and a divorce vice-versa), the two variables are included as potential amenity indicators.

The **share of the population aged 18-29** (variable name *young_pop*) is calculated using data from INKAR (2021) based on the *Fortschreibung des Bevölkerungsstandes des Bundes und der Länder* (Census 2011) for the year 2017. Besides many direct measures included in his "coolness-index", Florida (2002) proposes the population share of young adults being a proxy for how attractive a location is for what he calls "the creative class".

A more broad measure of the age structure is the **average age** (variable name *age_avg*) of the population in 2017, again based on the *Fortschreibung des Bevölkerungsstandes des Bundes und der Länder*. In contrast to the population share of young adults, it also (partially) captures higher shares of old residents.

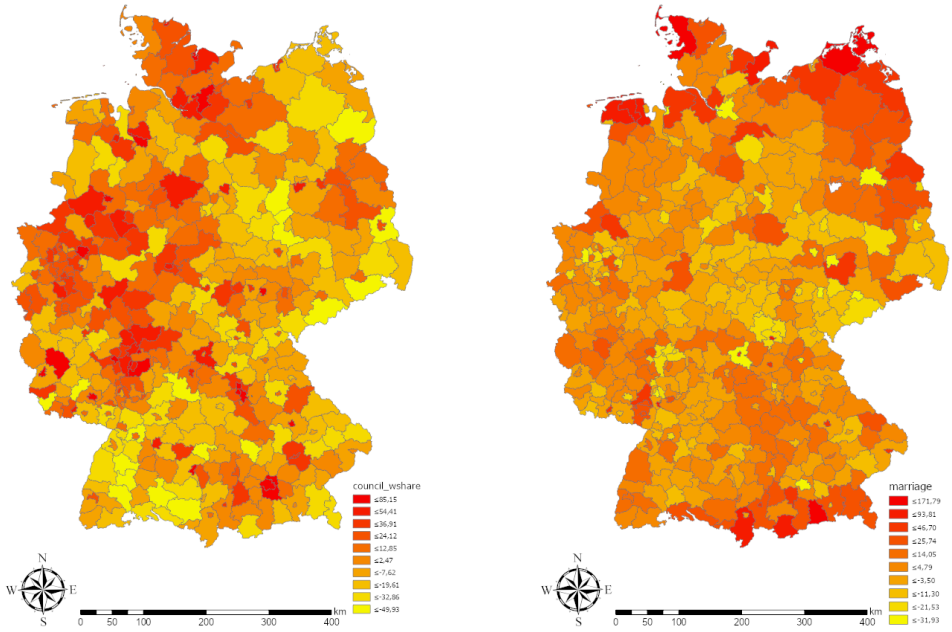
It may be important how the population of young adults is constituted. Therefore, I also include the **number of students** (variable name *students*) per 1000 residents in 2017. The variable provided by INKAR (2021) based on data from the *Hochschulstatistik des Bundes* does not distinguish between universities and university colleges (Fachhochschulen).

In the German education system, the majority of young adults that do

not attend a tertiary education institution apply for an apprenticeship. The **number of apprentices** (variable name *apprentices*) per 1000 workers in 2017 measures how important that form of education is for the local labor force. The variable is provided by INKAR (2021) utilizing information from the *Statistik der Bundesagentur für Arbeit*.

Figure 1.A.5: Demographics

(a) Female share of city council (b) Marriages



Notes: All values are deviations from the national mean in %.

Social participation

As mentioned already in section 1.4.3, indicators of social participation are supposed to measure how well-integrated residents feel in their local community. All indicators are mapped in Figure 1.A.6.

Election turnout (variable name *turnout*) is measured as the number of votes in the federal election in 2017 divided by the total number of residents eligible to vote. The data is provided by INKAR (2021) using information from the *Allgemeine Bundestagswahlstatistik des Bundes und der Länder*. All valid and non-valid votes of the secondary or main votes (Zweitstimme) are considered as participation in the election. Arguably, participation in the

local elections would be an even better indicator. However, no data on local election turnout with sufficient coverage was available.

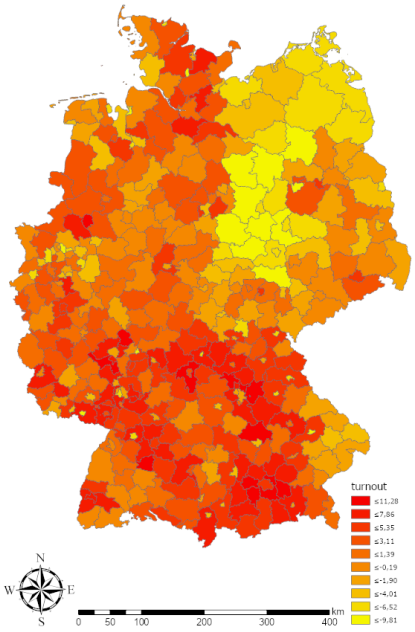
Information on the **share of residents in sports clubs** (variable name *sportsclubs*) is collected from 20 sports club associations, mainly on the state level.[36] The variable is defined as the sum of members of all sports clubs in the respective county divided by the number of residents. Drawbacks of the data are that members of more than one sports club were counted twice and that members do not necessarily live in the same county. This is especially problematic in the presence of large professional football clubs. Most of the states were able to provide data dating to January 1st of 2018. Exceptions are Saarland (2019), Rheinhessen (part of Rhineland-Palatinate; 2020), Thuringia (2020), Baden North and Baden South (part of Baden-Württemberg; 2019), and Württemberg (part of Baden-Württemberg; 2020). It is important to note that all data was collected before the start of the Covid-19 pandemic. In Bavaria, data for many cities is provided only together with the surrounding county. I approximated the share of members in sports clubs by dividing the reported number of members by the sum of the population of both counties.

The **Social Connectedness Index** (SCI) (variable name *sci_fb*) was obtained via the Partner Portal of Facebook in 2020. At the time of publishing this paper, a constantly updated version of the SCI is publicly available at `https://data.humdata.org/dataset/social-connectedness-index`. The data used here date to December 31st of 2017. The SCI is a scaled version of the probability that two randomly drawn residents of location i and location j, who are both members of the online social media platform Facebook, are linked by a Facebook friendship. The variable *sci_fb* only uses information for $i = j$, i.e. the likelihood that Facebook members within the same county are linked on Facebook. High values of *sci_fb* represent a relatively high density of the local social network under the assumption that online friendships are correlated with real-life interaction.

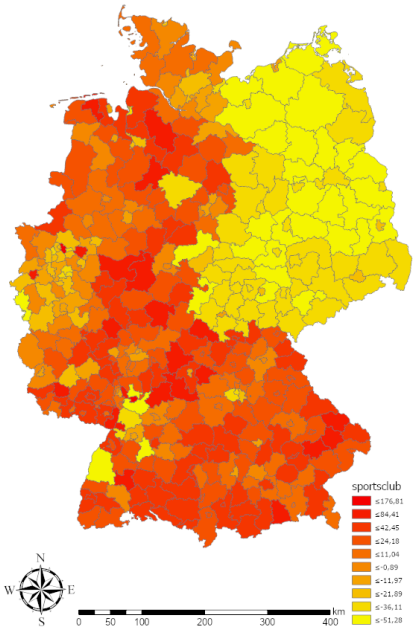
[36] Rhineland-Palatinate and Baden-Württemberg collect data not on the state level but in three distinct regional associations each. Not all associations provide the data on county level directly on their website. Some had to be contacted individually.

Figure 1.A.6: Social participation

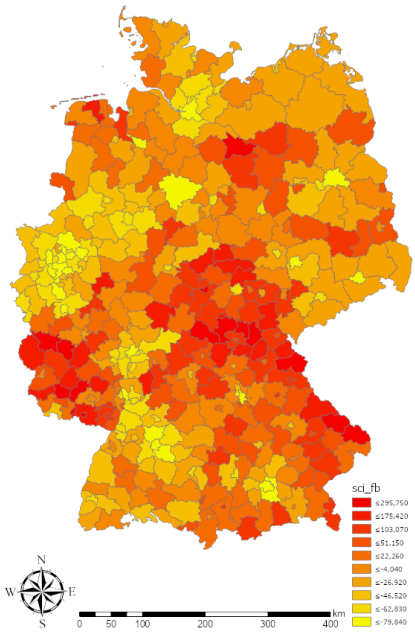
(a) Election turnout



(b) Sports clubs



(c) Social Connectedness Index



Notes: All values are deviations from the national mean in %.

Labor market

Indicators on the labor market are included to incorporate characteristics of the local labor market that are not covered by the income differentials of gross median income derived in section 1.4.1. All variables are obtained from INKAR (2021), although some variables are available at *Regionaldatenbank Deutschland* (RDB), too. Some representative indicators are visualized in Figure 1.A.7.

The **unemployment rate** (variable name *unemployment*) uses information from the *Arbeitsmarktstatistik der Bundesagentur für Arbeit* (AMS) and reflects the share of unemployed of the labor market population (Erwerbspersonen) in 2017.

The **long-term unemployment rate** (variable name *unemployment_long*) is defined as the share of all unemployed workers that is unemployed for at least one year in 2017. The data also stems from the AMS (see above).

The qualification requirements of **vacancies** are included for four different levels. The qualification requirement groups are defined by the employment agency (Arbeitsargentur): assistant (German: Helfer; variable name *vacancy_assist*), skilled (German: Fachkraft; variable name *vacancy_skilled*), expert (German: Experte; variable name *vacancy_expert*), and specialist (German: Spezialist; variable name *vacancy_spec*).[37] For each level, the variable measures the average share of all vacancies reported to the employment agency in 2017 that fall into the respective requirement group. A high value indicates that labor market migration is relatively easy for workers with the respective level of qualification. Information is also based on AMS (see above).

The **poverty rate** (variable name *poverty*) is defined to be the share of recipients of unemployment benefits of type SGB II (Arbeitslosengeld II) among all residents under 65 years in 2017. This includes both (long-time) unemployed as well as non-employable. The original data source is the *Statistik der Grundsicherung für Arbeitsuchende nach dem SGB II der Bundesagentur für Arbeit*.

[37] A more detailed definition of each group can be found here: https://statistik.arbeitsagentur.de/DE/Statischer-Content/Grundlagen/Methodik-Qualitaet/Methodische-Hinweise/uebergreifend-MethHinweise/Anforderungsniveau-Berufe.html, last accessed November 2022.

The **children poverty rate** (variable name *children_poverty*) is defined equivalently to the poverty rate, except that only recipients and residents under 15 years are considered.

The **share of residents in dept** (variable name *privat_dept*) is the share of debtors of all adult residents in 2017 in percent. The data originates from the *Schuldneratlas Deutschland des Verbands der Vereine creditreform e.V.*.

The **share of out-commuters** (variable name *commute_out*) and the **share of in-commuters** (variable name *commute_in*) are included for similar reasons as the migration balance. The share of out-commuters is defined as the share of all employed residents that do not work in the same county in 2017. It is an indirect proxy for how attractive neighboring labor markets are (relatively). The share of in-commuters is the share of all employed in the respective county who live in a different county in 2017. This is included as a proxy for the relative attractiveness of surrounding regions in terms of living conditions. Both variables are based on the *Beschäftigtenstatistik der Bundesagentur für Arbeit*.

The **share of highly qualified worker** (variable name *work_highskilled*) measures the share of workers employed in a county that have a college degree in 2017. This variable, based on information from the *Beschäftigtenstatistik der Bundesagentur für Arbeit*, should not be interpreted as an amenity indicator as in Shapiro (2006) since it was used to calculate the income differentials in section 1.4.1 already.

The **share employed in the creative sector** (variable name *work_creative*) and the **share employed in the knowledge-intensive sector** (variable name *work_knowledge*) measure the employment share of specific sectors in 2017. The creative sector includes the publishing industry, film industry, recording industry/music publishing, broadcasting industry, cultural industries, libraries/museums, trade in cultural goods, architecture, design, advertising, and software/games. It is the best available indicator of the bohemian class proposed by Florida (2002). The knowledge-intensive sector consists of the chemical industry, pharmaceutical industry, electronic industry, mechanical engineering, and automobile industry. Both variables are based on information from the *Beschäftigtenstatistik der Bundesagentur für Arbeit*.

The **gender wage gap** (variable name *gender_wagegap*) measures the gross full-time labor income of women relative to the full-time labor income

of men in 2017. Therefore, a higher value corresponds to higher female or lower male labor income. The variable provided by the *Statistik der Bundesagentur für Arbeit* does not control for differences in occupations or levels of hierarchy.

The **gender employment gap** (variable name *gender_empgap*) is defined as the (full-time) employment rates of women divided by the full-time employment rate of men living in the respective county in 2017. Hence, a higher value of *gender_empgap* represents more (less) women (men) in full-time labor. Original data stems from the *Beschäftigtenstatistik der Bundesagentur für Arbeit*.

Average working hours (variable name *workinghours*) are calculated as the total number of working hours in 2017 divided by the number of workers. This includes working hours in both, self- and dependent employment. Holidays and other working hours that were taken off are excluded. The variable is derived from the *Volkswirtschaftliche Gesamtrechnung der Länder*.

Firms

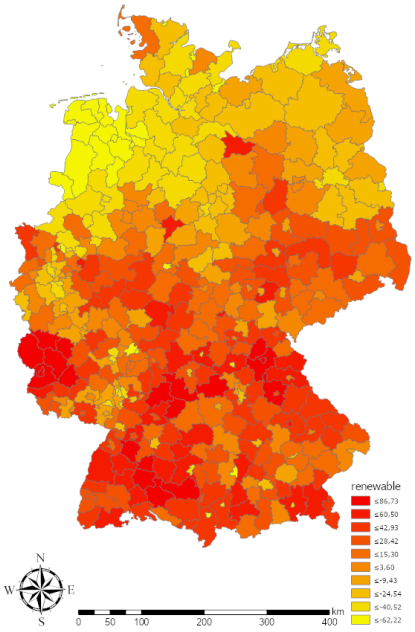
Similar to the labor market sub-category, indicators here are mainly included to take systematic differences of the local economy into account that are not covered by the income differentials derived in section 1.4.1. The variables are again obtained from INKAR (2021) and *Regionaldatenbank Deutschland* (RDB).

Sectoral employment shares of 2017 are all based on the *Beschäftigtenstatistik der Bundesagentur für Arbeit*. The primary sector (variable name *sector_prim*) represents agriculture, forestry, and fishing. The secondary sector (variable name *sector_sec*) consists of the producing industry including the construction industry. Workers of the tertiary sector (variable name *sector_ter*) belong to the service industry.

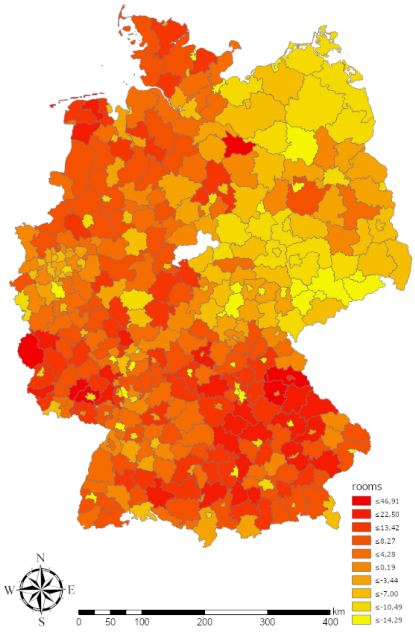
The **share of firms of different size** controls for potential income differences that are caused by wage premia that depend on firm size. Large firms (variable name *firms_large*) are defined to have more than 250 employees. Firms with 50 to 249 employees are labeled medium sized firms (variable name *firms_med*). Small firms (variable name *firms_small*) are those with 10-49 employed workers. The remaining tiny firms (variable name

Figure 1.A.7: Poverty

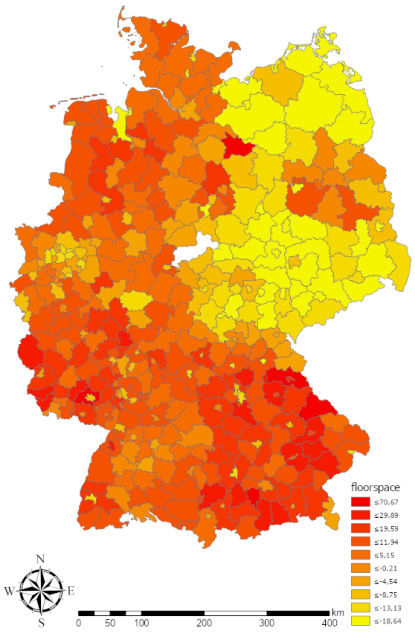
(a) Flats with renewable energy



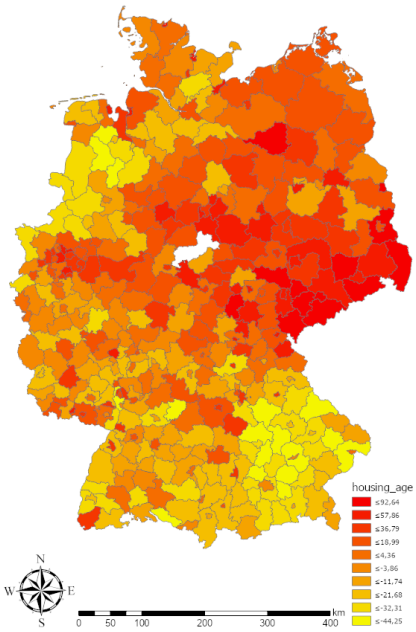
(b) Private dept



(c) Gender wage gap



(d) Gender employment gap



Notes: All values are deviations from the national mean in %.

firms_tiny) have less than 10 employees. All variables are based on the *Unternehmensregister-System des Bundes und der Länder* obtained via INKAR

(2021) and represent the universe of firms in 2017. The shares are plotted on the county level in Figure 1.A.8.

To have a better sense of the dynamics of the local economy I also include the **number of new firms** per capita (variable name *firms_birth)* and the **number of firm closures** per capita (variable name *firms_death*) in 2017. Both variables are calculated using data obtained at RDB from the *Gewerbeanzeigenstatistik des Statistischen Bundesamts*. Moving firms and mergers are not considered as firm birth or closure.

While the number of new firms can be seen as an indicator of innovativeness, **investment** per worker (variable name*invest*) would be a more direct measure. Data on firms' investments can be obtained via RDB (section Regionalatlas). However, many counties including all from the state of North Rhine-Westphalia are missing for all available years in this data set. Since excluding *invest* from the analysis in section 1.6 does not change the results, the variable is dropped from the main specification to keep the number of observations as high as possible.

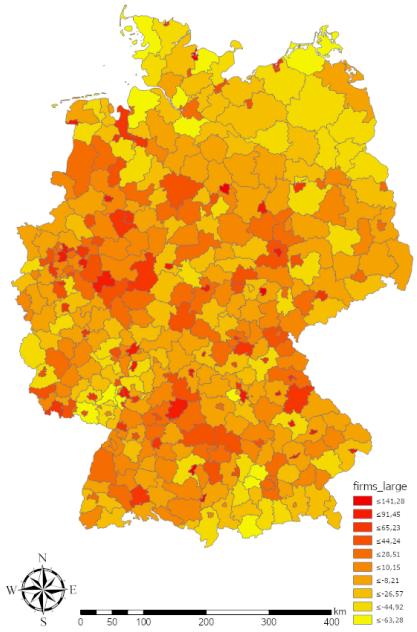
Fiscal variables

There are three main **local tax rate** multipliers that can be set by municipalities autonomously. All three variables are defined as multipliers (in percent) and can be obtained from the *Statistikportal* of the *Statististische Ämter des Bundes und der Länder*.[38] The land tax rate type A (variable name *land_tax_A*) is the average rate at which the tax base amount for land used for agriculture and forestry is multiplied in the municipalities of a county in 2017. All the remaining land is taxed using the land tax rate type B (variable name *land_tax_B*), including privately owned property. Both land tax multipliers can be set without any federal restrictions. The business tax multiplier (variable name *business_tax*) has to be at least 200% by federal law and taxes local firms' profits above a certain allowance. The local tax rate multipliers represent two factors relevant to this paper. First, a higher multiplier means a higher tax burden, both for businesses (all multipliers) and for households (land tax multiplier type B). Second, higher multipliers

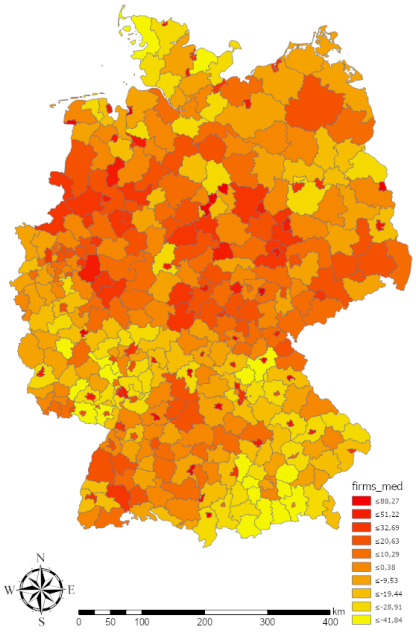
[38] See https://www.statistikportal.de/de/veroeffentlichungen/hebesaetze-der-realsteuern-deutschland, last accessed on 21.01.23, for more information.

Figure 1.A.8: Firm size

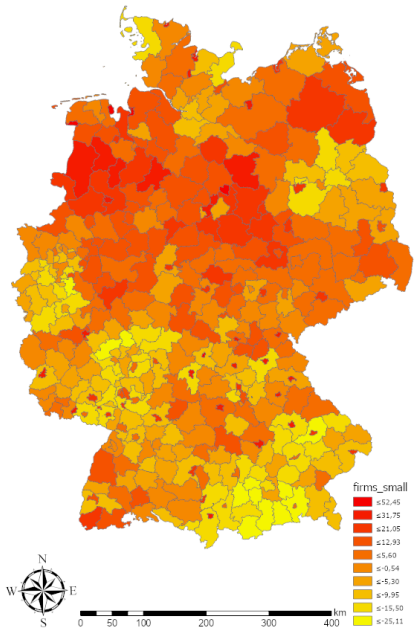
(a) Share of large firms



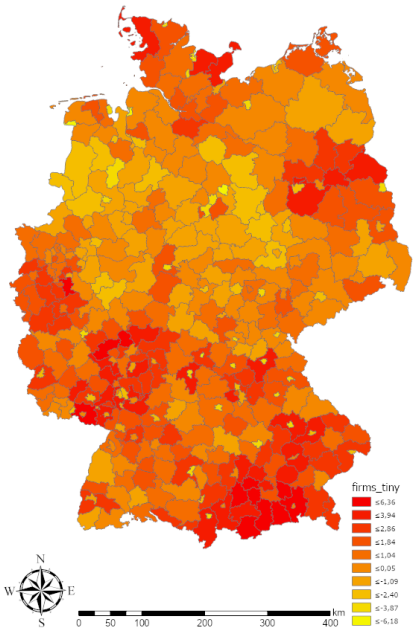
(b) Share of medium firms



(c) Share of small firms



(d) Share of tiny firms



Notes: All values are deviations from the national mean in %.

result in higher tax income for the municipality (ceteris paribus), which can be spent to provide public goods.

The provision of local goods also heavily depends on the **local public debt** per resident (variable name *public_dept*) in 2017. High public debt does not only comes with an incentive to cut down spending for public goods for faster debt repayment but also results in high interest liabilities. The information originates from the *Statistik über Schulden des Bundes und der Länder* and is provided by INKAR (2021).

One of the local public goods that is part of the fiscal variable sub-category is the size of the **staff of the local administration** per 10,000 residents (variable name *public_empl*) in 2017. The *Personalstandsstatistik der Länder, Gemeinden und Gemeindeverbände* obtained via INKAR (2021) also includes the staff of hospitals and other businesses that are run by the public locality here. A higher number of employees per resident is assumed to increase the quality of public service, e.g. by reducing waiting times in the local administration.

The main fiscal variables are visualized in Figure 1.A.9.

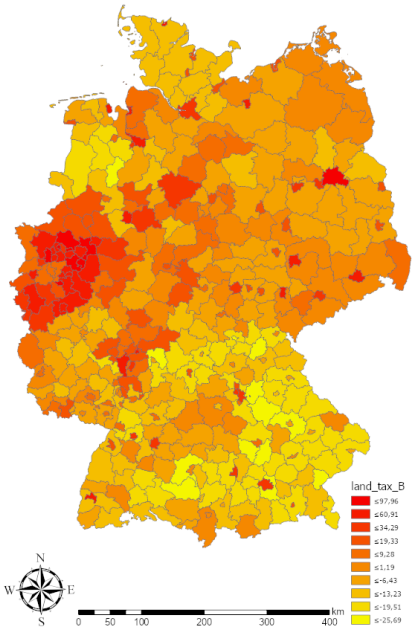
Cost of living

As discussed above, not all price differences are incorporated in the derived rent differentials. To make sure that unobserved variations in price levels are not misinterpreted as (dis-) amenities, I include **consumer price indices (CPIs)** for goods (variable name *CPI_goods*) and services (variable name *CPI_services*) from Weinand and von Auer (2020) as controls. Weinand and von Auer (2020) calculate local CPIs using CPI micro-data provided by the *Statististische Ämter des Bundes und der Länder* dating to May 2016.

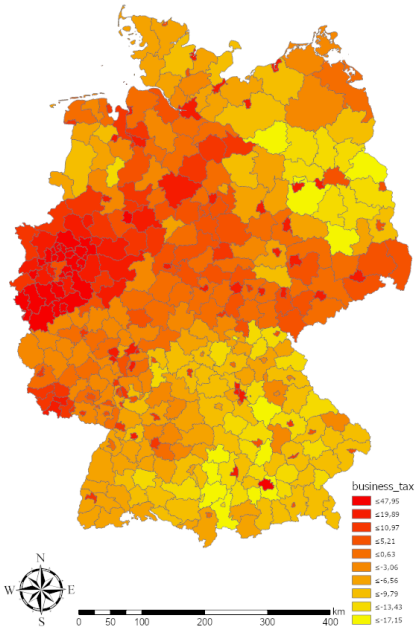
Another concern is that behavioral responses might affect the impact of income and cost of living differentials on utility. Suppose households in a relatively expensive location are willing to trade off less living space for a higher level of consumption. In that case, the negative (positive) effect of high (low) prices on utility is weakened, especially on the tails of the price differential distribution. Using national averages, e.g. for the expenditure share on housing, in section 1.4.2 would then lead to an over- (under-)estimation of QoL. To partially control for this, I include the **average living space** (in m^2) per resident (variable name *living_space*) in 2017 as a control. Based on the *Fortschreibung des Wohngebäude- und Wohnungsbestandes des Bundes*

Figure 1.A.9: Fiscal variables

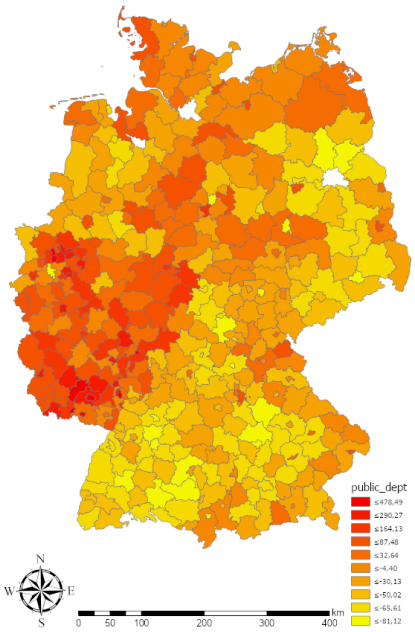
(a) Land tax B



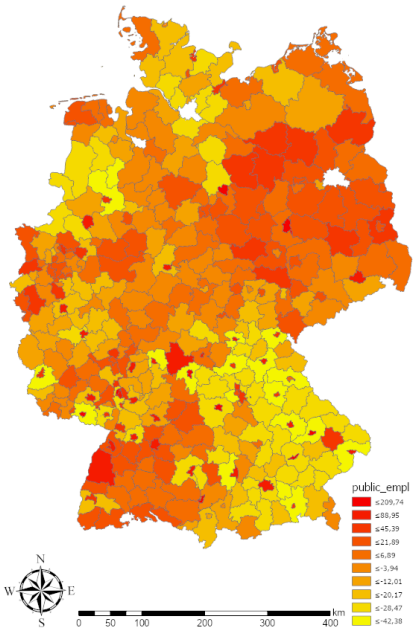
(b) Business tax



(c) Public dept



(d) Size of local administration



Notes: All values are deviations from the national mean in %.

und der Länder and obtained from INKAR (2021), the variable has one main drawback, as living space of empty and non-residential buildings is not

excluded.

Recreation & entertainment

Indicators on recreation and entertainment facilities are supposed to measure the quality and variety of opportunities to spend enjoyable free time. They are mainly obtained from OpenStreetMap (OSM). Shapefiles dating to May 2019 were obtained from *Geofabrik*.[39] OSM is a non-profit user-generated map service.

The **number of museums** per resident (variable name *museums*), the **number of restaurants** per resident (variable name *restaurants*), the **number of theaters** per resident (variable name *theaters*), and the **number of shops** per resident (variable name *shops*) are all obtained by aggregation the number of so-called points of interest in each county. That means that everything in a respective category is counted. There is no differentiation in quality, small theaters with few seats have the same weight as large state theaters. Restaurants can also include takeaway and fast-food restaurants. The number of shops, however, excludes supermarkets and grocery stores, as they build their own category. Figure 1.A.10 presents the four indicators.

Additionally, the availability of sports grounds, parks, and other green spaces are measured by the **land share of recreation areas** (variable name *recre_space*) in 2017 obtained via INKAR (2021). As with other land use indicators, the variables represent the share of land used for the above-mentioned purposes in percent. The original data is provided by the *Flächenerhebung nach Art der tatsächlichen Nutzung des Bundes und der Länder*.

Culture

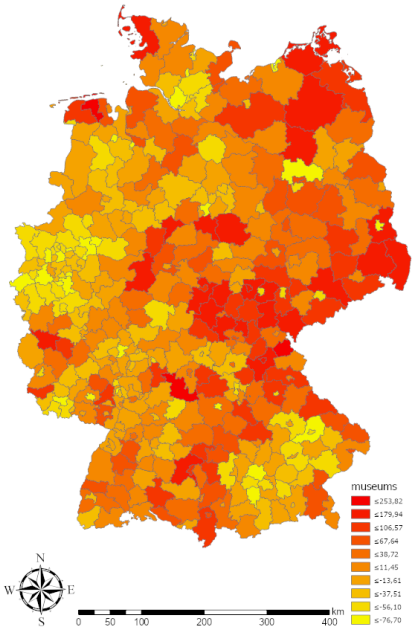
Cultural indicators have a strong overlap with indicators of recreation & entertainment. The density of museums and theaters, for example, could be part of this sub-category, too. I decided to include only broad indirect measures of (cultural) attractiveness here, namely how touristic a location is.[40] What exactly makes a region appealing to tourists can be related to

[39] `https://www.geofabrik.de/de/index.html`, last accessed 31.01.23.

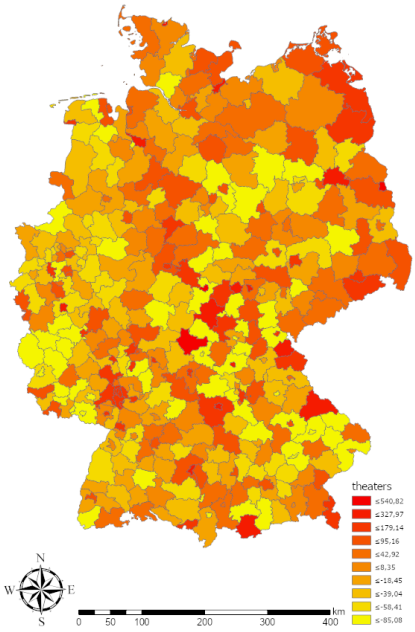
[40] It should be noted that the classification in sub-categories does not play a role in my analysis. It only serves the purpose to organize the data.

Figure 1.A.10: Recreation & entertainment facilities

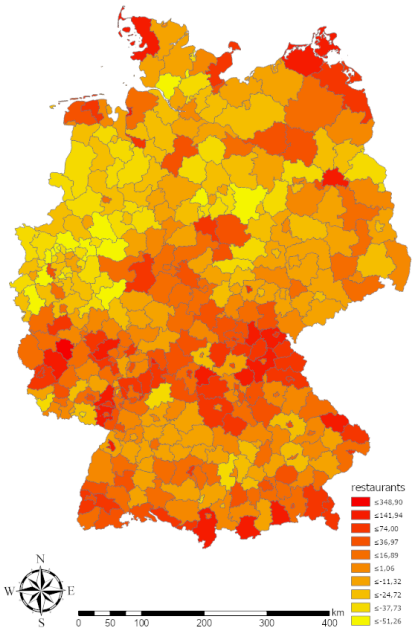
(a) Museums



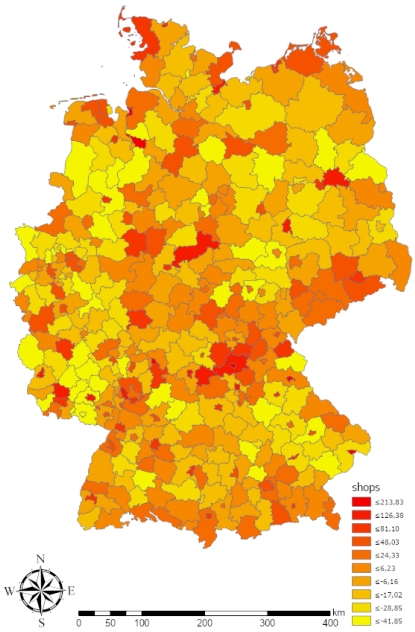
(b) Theaters



(c) Restaurants



(d) Shops



Notes: All values are deviations from the national mean in %.

the former sub-category. But natural features, scenery old towns, or even business-related factors can play a role as well. Both variables are available

both on INKAR (2021) and RDB and are based on administrative data from the *Monatserhebung im Tourismus des Bundes und der Länder*.

The **number of tourist overnight stays** per resident (variable name *tourist_stays*) measures the total number of overnight stays in accommodation establishments in 2017 divided by the number of residents. This includes among others hotels, inns, vacation homes and apartments, youth hostels, campsites, and also preventive and rehabilitation clinics that can accommodate at least 10 guests at the same time.

Number of hotel beds per resident (variable name *hotelbeds*) represents the average number of beds in opened accommodation establishments that can accommodate at least 10 guests at the same time in 2019 divided by the number of residents.

Health

The quality and accessibility of healthcare are especially important to elderly people. All indicators in this sub-category are obtained from INKAR (2021).

The **number of physicians** per 10,000 residents (variable name *doctors*) in 2017 uses data from *Kassenärztliche Bundesvereinigung*. The variable includes not only physicians but also their staff to weigh larger offices stronger. Psychological psychotherapists are not covered here.

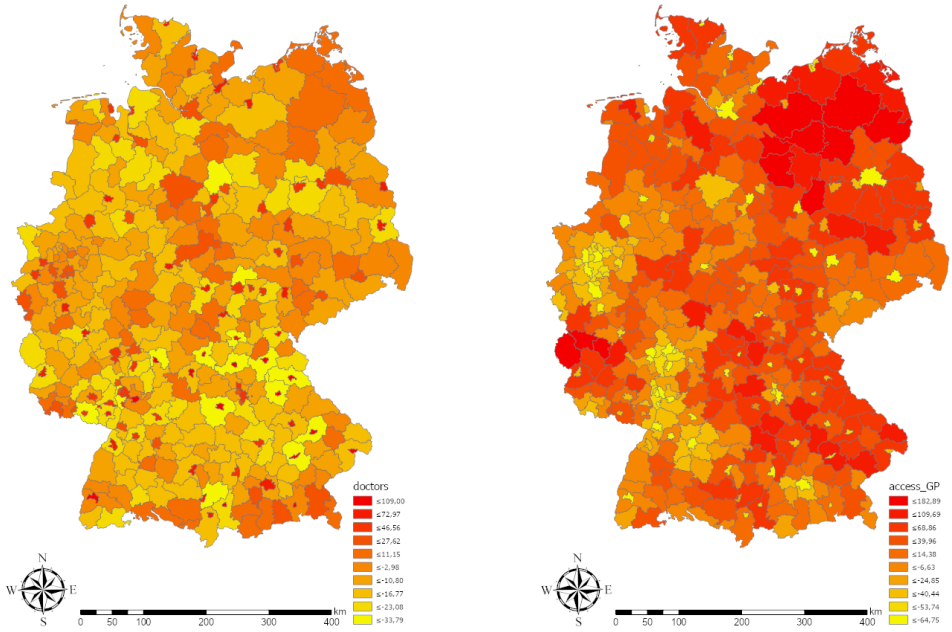
Especially older people lack mobility. Therefore, the **access to general practitioners (GPs)** (variable name *access_GP*) is probably more crucial than the overall density of physicians. The variable is a population-weighted average distance to the closest GP in 2015 and is calculated using various data sources by the *Bundesinstitut für Bau-, Stadt-, und Raumforschung* (BBSR).[41] A high value of that indicator represents a high distance and therefore a relatively low level of average accessibility. Figure 1.A.11 compares the physician density (*doctors*) and **access to general practitioners (GPs)** (*access_GP*) graphically.

The **number of hospital beds** per 1000 residents (variable name *hospital_beds*) is a measure for the inpatient provision of medical services in 2017. The variable uses data from the *Krankenhausstatistik des Bundes und der*

[41] The interested reader is referred to `https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/Komponenten/LaufendeRaumbeobachtung/laufenderaumbeobachtung.html`; last accessed 03.02.23.

Figure 1.A.11: Physicians

(a) Physician density (b) Access to general practitioners



Notes: All values are deviations from the national mean in %.

Länder.

The *Pflegestatistik des Bundes und der Länder* allows to include indicators on the availability of nursing services. The **staff in nursing homes** per 100 inpatient patients (variable name *nursing_emp*) represents the quality of nursing homes in 2017. Nursing homes are inpatient care facilities in which persons in need of care are cared for under the constant responsibility of a trained nurse and can be accommodated and fed all day and/or only during the day or night. The availability of such nursing homes is measured by the **number of nursing home places** per 10,000 residents (variable name *nursing_places*). I also consider outpatient care facilities that provide care and housekeeping services to persons in need of care in their homes by including the size of the **staff in nursing services** (variable name *nursing_service*).

Similar to the access to GPs, the **access to pharmacies** (variable name *access_pharma*) is defined as the population-weighted average distance to the closest pharmacy in 2017. The indicator is also generated by the (BBSR).

The **infant mortality** (variable name *infant_mortality*) and **life expectancy** (variable name *life_expect*) are rather outcomes than direct indicators of the

quality of healthcare. Infant mortality is defined as the number of deaths among 1000 children younger than one year in 2017 and calculated based on the *Statistik der Sterbefälle des Bundes und der Länder*. Life expectancy represents the average number of years a newborn (younger than one year old) would have to live if the mortality ratios measured during the reporting period of 2017 did not change throughout that child's life. The indicator is not affected by the age structure of the population but may partially represent the general health status of parents, especially mothers. It is based on the *Statistik der Geburten und Sterbefälle des Bundes und der Länder*.

Public safety

Indicators of public safety have a long tradition of dis-amenities, especially in the US-based literature. Besides typical measures of crime, I also include an indicator of the risk of being involved in a traffic accident.

I collect three different crime rates. All data on crimes is obtained from the *Polizeiliche Kriminalstatistik* provided by the *Bundeskriminalamt*.[42] The **total crime rate** (variable name *crime*) is calculated by dividing the number of all cases (excluding violations of the laws on residence and asylum) in 2017 by the number of residents. The **violent crime rate** (variable name *crime_violent*) only considers cases of murder, homicide, rape, sexual coercion, sexual assault, robbery, serious bodily harm, extortionate kidnapping, hostage-taking, and attacks on air and sea transport and might be a more coherent indicator of overall public safety. To be more focused on perceived safeness, I also include the **street crime rate** (variable name *crime_street*). This variable only incorporates crimes that (usually) happen on the streets like pickpocketing and property damage but also more serious violent crimes like sexual assault or bodily harm as well as more generally crimes from groups.

Road traffic casualties (variable name *road_casual*) are defined as the number of killed or injured people in accidents per 100,000 residents in 2017. The indicator is based on data from the *Statistik der Straßenverkehrsunfälle des Bundes und der Länder* and provided by INKAR (2021).

[42] https://www.bka.de/DE/AktuelleInformationen/StatistikenLagebilder/PolizeilicheKriminalstatistik/pks_node.html, last accessed 03.02.23.

Education

Similar to health care, education is a public good where quality and accessibility are important. Municipalities are mainly in control of many factors in the provision of education, including the number and location of schools. Figure 1.A.12 visualized representative indicators for both, quality (dropout rate) and accessibility (school density).

The **student/teacher ratio** (variable name *stud_teach_ratio*) is a typical proxy for the quality of education, where a high value stands for low quality. I calculate the ratio using administrative data of the school year 2017/2018 from the *Kommunale Bildungsdatenbank* (KBD) run by the *Statististische Ämter des Bundes und der Länder*.[43] It should be noted that the state of Saarland is generally not included in the KBD and was also not able to provide data directly. In addition, Saxony-Anhalt does not publish data on the number of teachers. Therefore, *stud_teach_ratio* is missing for counties of these two states. It was not possible to weigh teachers according to their working hours. Hence, part-time teachers are weighted similarly to full-time teachers.

The **high school dropout rate** (variable name *grad_non*) is defined as the share of all school leavers that do not have any degree in 2017. The indicator is obtained from INKAR (2021), but can similarly be calculated using data from KDB. Again, counties from the state of Saarland are missing.

INKAR (2021) also provides readily usable variables on the **share of graduates by the level of degree**. The levels basically represent the three different tracks of secondary education, which are low (German: Hauptschulabschluss; variable name *grad_low*), medium (German: mittlerer Schulabschluss; variable name *grad_med*) and high (German: Hochschulreife; variable name *grad_high*). The share is calculated by dividing the number of school leavers with the respective degrees by the total number of school leavers in 2017, including those without degrees. A greater share of graduates with a high degree could be interpreted as higher quality of education. However, the numbers heavily depend on the state and serve only as controls for the supply of labor.

The **number of children in childcare** of all children (variable name *childcare*) in 2017 measures the share of under 3-years-old children in daycare

[43] https://www.bildungsmonitoring.de/bildung/online/; last accessed 03.02.23.

facilities. It is calculated by INKAR (2021) building on data from *Kindertagesbetreuung regional (Gemeinschaftsveröffentlichung Statistische Ämter des Bundes und der Länder)*.

The **number of kindergartens** per resident (variable name *kindergardens*) is obtained from OpenStreetMap (OSM) (see section 1.A for details on OSM). The variable represents the overall accessibility of childcare for children usually aged 3 to 5 in 2019. It is not possible to take the size of a kindergarten into account. Since all children from the age of 3 are entitled to a kindergarten place, the kindergarten density is not a useful measure of availability.

The **number of schools** per resident (variable name *schools*) in the school year 2017/2018 are mainly obtained from KBD. Missing data for counties of the state of Saarland were substituted by numbers obtained from OSM (2019). As for kindergartens, children are entitled to a place in school. Hence, this indicator measures accessibility more than availability.

A more precise indicator of accessibility is provided by INKAR (2021). The **access to primary schools** (variable name *access_primschool*) is a population-weighted average distance to the closest primary school in 2016 to 2018. For more details on how the variable is derived please check the description in section 1.A. A high value of that indicator represents a high distance and therefore a relatively low level of average accessibility.

Infrastructure & mobility

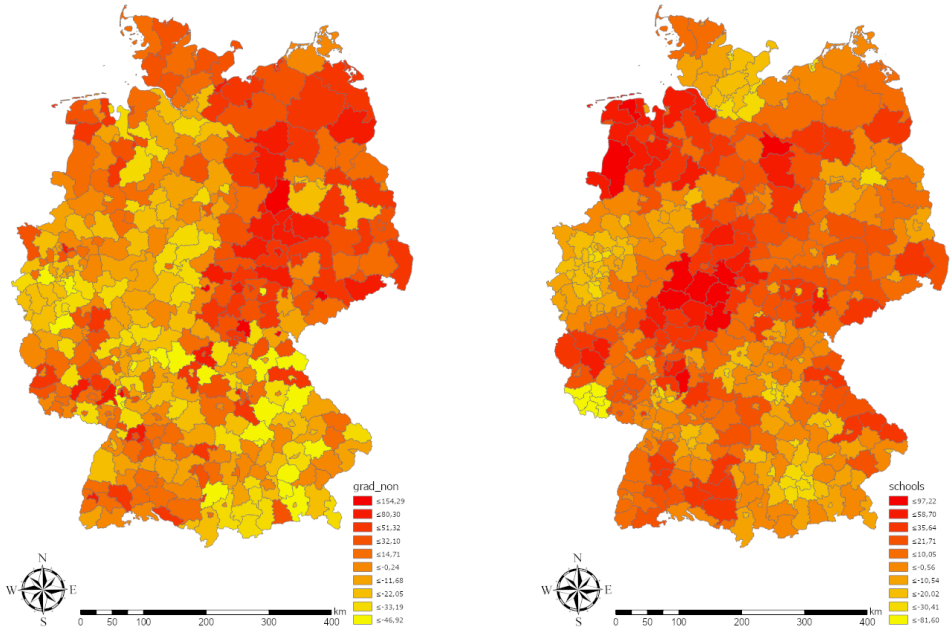
This sub-category mainly contains variables on accessibility calculated using various data sources by the *Bundesinstitut für Bau-, Stadt-, und Raumforschung* (BBSR) provided via INKAR (2021).[44] All access variables are weighted averages of the distance or driving time to the closest point of interest. That means that a high value of the respective accessibility indicator represents a high distance and therefore a relatively low level of average accessibility.

The **coverage with high speed internet** (variable name *internet*) is the

[44]The interested reader is referred to https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/Komponenten/LaufendeRaumbeobachtung/laufenderaumbeobachtung.html, last accessed 03.02.23, and https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/Komponenten/Erreichbarkeitsmodell/erreichbarkeitsmodell.html; last accessed 03.02.23.

Figure 1.A.12: Education

(a) Drop out rate (b) School density



Notes: All values are deviations from the national mean in %.

share of households with access to an internet connection with at least 100 Mbit/s in 2020. Building on the *Breitbandatlas des Bundesministeriums für Verkehr und digitale Infrastruktur* the indicator is obtained via INKAR (2021). A fast internet connection is important for firms with digital infrastructure as well as households, especially with members that work from home.

The **access to high speed railway** (variable name *access_railway*) is the area-weighted average of the driving time (by car) to the closest high-speed railway stop in 2020, measured in minutes.

Similar to the accessibility of the railway network, the **access to highways** (variable name *access_highway*) is the area-weighted average of the driving time (by car) to the closest motorway junction in 2020, measured in minutes.

Access to airports (variable name *access_airport*) is defined as the area-weighted average of the car driving time to the closest international commercial airport in 2020, measured in minutes.

Access to public transport (variable name *publictrans*) is derived as the population-weighted average distance to the closest public transport stop in 2018. Only stops with at least 20 departures on a weekday were considered.

Access to a big city (variable name *access_bigcity*) is the area-weighted average of the car driving time to the closest big city in 2020, measured in minutes. The indicator on the **access to a medium sized city** (variable name *access_medcity*) is defined equivalently. The exact definition of a big (Oberzentrum) and a medium-sized city (Mittelzentrum) does not only depend on the total population but also on the relative importance to a region.[45]

The **access to grocery stores** (variable name *access_grocery*) is the population-weighted average distance to the closest grocery store in 2017.

The **number of cars** per 1000 residents (variable name *cars*) measures the car density in 2017. Original data stems from the *Statistik des Kraftfahrzeugbestandes des Bundes und der Länder* and is provided by INKAR (2021). A high density of cars can be interpreted as an indicator of congestion or of a high level of mobility. However, especially for more rural areas, a high density of cars could also be an inverse measure of the availability of alternative modes of transportation like public transport.

[45] More information on the definition of the relative importance of cities including a map can be found here: https://www.bmwsb.bund.de/Webs/BMWSB/DE/themen/raumentwicklung/raumordnung/zentrale-orte/zentrale-orte-node.html; last accessed 04.02.23.

1.B Detailed derivation of the model

Starting with the equilibrium condition of the model, equation 1.3,

$$E(\mathbf{p_i}, w_i, \tau, \bar{u}, Q_i) = 0 \; \forall \; i \tag{1.12}$$

totally differentiating (at national averages denoted by $\bar{x}$) yields:

$$\frac{\partial E}{\partial \mathbf{p_i}} \mathbf{dp_i} + \frac{\partial E}{\partial w_i} dw_i + \frac{\partial E}{\partial Q} dQ_i = 0 \tag{1.13}$$

With Shephard's Lemma ($\frac{\partial E}{\partial \mathbf{p_i}} = \mathbf{y}$ and $\frac{\partial E}{\partial w_i} = (-1 + \tau')$) this is

$$\mathbf{y dp_i} + (-1 + \tau') dw_i = -\frac{\partial E}{\partial Q} dQ_i. \tag{1.14}$$

Extending the equation by national averages yields:

$$\mathbf{y}\bar{\mathbf{p}} \frac{\mathbf{dp_i}}{\bar{p}} + (-1 + \tau')\bar{w} \frac{dw_i}{\bar{w}} = -\frac{\partial E}{\partial Q} dQ_i. \tag{1.15}$$

Using hat-notation for percentage changes, $\hat{w}_i \equiv dw_i/\bar{w}, \hat{\mathbf{p}}_i \equiv \mathbf{dp}_i/\bar{\mathbf{p}}$ gives

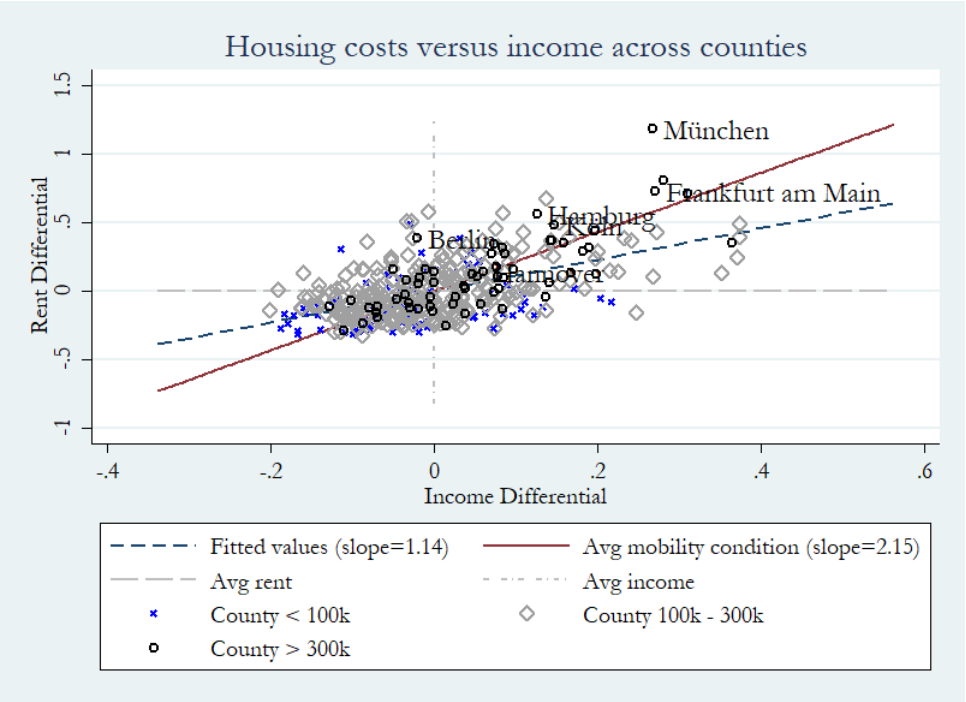
$$\mathbf{y}\bar{\mathbf{p}}\hat{\mathbf{p}}_\mathbf{i} - (1 - \tau')\bar{w}\hat{w}_i = -\frac{\partial E}{\partial Q} dQ_i \tag{1.16}$$

By dividing both sides by the national average total income $\bar{m}$ and normalizing $\hat{Q}_i \equiv -(\frac{\partial E}{\partial Q}) dQ_i/\bar{m}$ we arrive at equation 1.5 from section 1.3.1:

$$\hat{Q}_i = \frac{\mathbf{y}\bar{\mathbf{p}}}{\bar{m}} \hat{\mathbf{p}}_\mathbf{i} - (1 - \tau') \frac{\bar{w}}{\bar{m}} \hat{w}_i \tag{1.17}$$

1.C Mobility condition

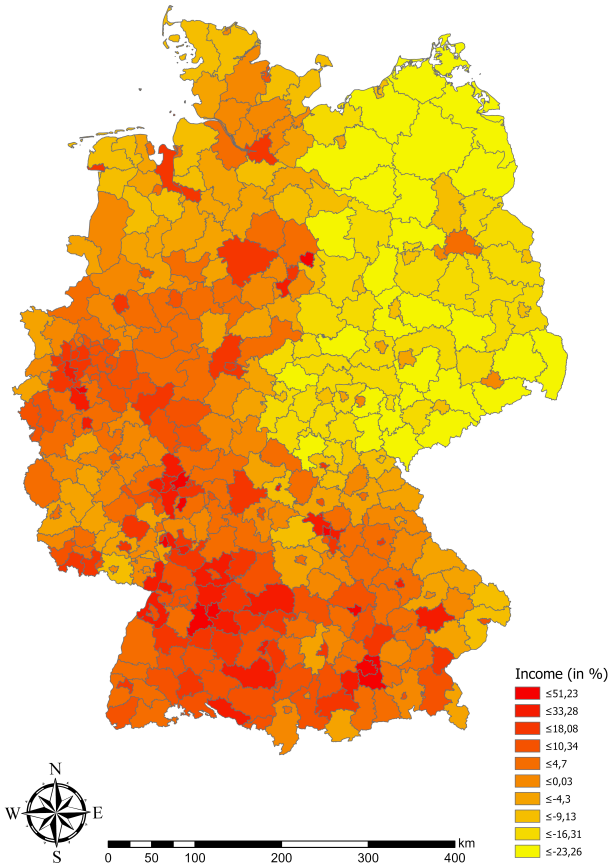
Figure 1.C.13: Mobility condition



Notes: Income and rent differentials and the average mobility condition in the form of equation 1.11.

Figure 1.C.13 plots the income and rent differentials for all counties, separated by population size. The red solid line represents the average mobility condition given the parameters of equation 1.11 for counties with an average QoL ($\hat{Q}_i = 0$)). In line with the Rosen-Roback framework, rents rise with increasing levels of income for a given level of QoL. In locations above this line, the cost of living is higher than in regions with an average level of attractiveness. Therefore, QoL has to be higher in the county given the assumption of free mobility.

Figure 1.D.14: Gross Income



Notes: Gross median labor income (deviation from the national mean in %) without controlling for east-west differences.

1.D Alternative Income and Price Indices

Gross Income

Figure 1.D.14 visualized the structural differences between East and West Germany in terms of gross labor income in 2017. Wages are systematically lower in counties of the former GDR, highlighting the importance to control for these differences when calculating income differences for my QoL index. Otherwise, the historic impact of the GDR on today's wages translates into systematically higher levels of QoL in all East German counties.

Prices

Figure 1.D.15 compares alternative price differentials. In panel (a), purchase prices of houses from the online platform ImmobilienScout24 are used. Similar to my price differential derived from rents, the house price index controls for housing characteristics (see Boelmann and Schaffner (2019)). Overall, the distribution of house price differentials has a higher standard deviation and ranges from 210% above the national average to 59% below. In comparison, my preferred rent-based differential varies between -32% and 118%. This is not surprising as house prices are likely to include future returns. To make them comparable to rents, Albouy (2008) suggests multiplying them with a discount rate. Besides that, the two measures are very similar with a Spearman correlation of 0.92 and a Pearson correlation of 0.93.

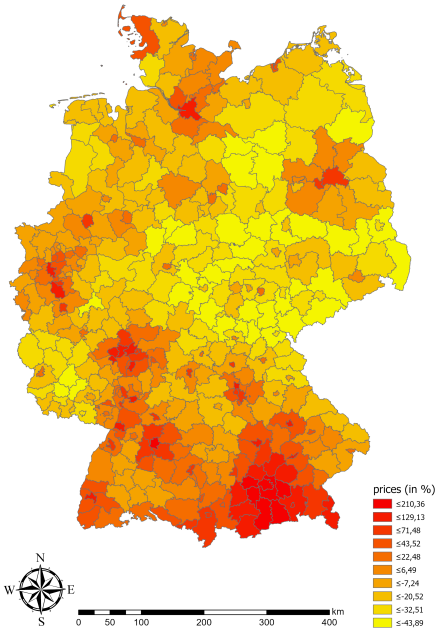
The overall picture of the publicly available rent data from INKAR (2021) in panel (b) looks similar. The correlation with my preferred price differentials based on ImmobilienScout24 data is even higher (Spearman correlation of 0.96 and Pearson correlation of 0.97). Again, the variation is larger, reflecting the absence of controlling for flat characteristics. Figure 1.A.4 shows that there are systematic differences in the housing stock, emphasizing their relevance when deriving a local rent price index.

The house price index from Weinand and von Auer (2020), is visualized in panel (c) of Figure 1.D.15, uses rents as well as the cost of owner-occupied housing. The overall correlation with my preferred rent index is 0.86. The deviations can be explained by the different timing, as Weinand and von Auer (2020) use data from May of 2016, and of course by the inclusion of owner-occupied housing. The most notable difference, however, is that the index of Weinand and von Auer (2020) represents the %-deviation from the population-weighted average, rather than the raw county average. Hence, the national average is higher as more populated counties tend to have higher costs of housing. In line with that, the house price index in panel (c) is shifted downwards, especially at the top of the distribution.

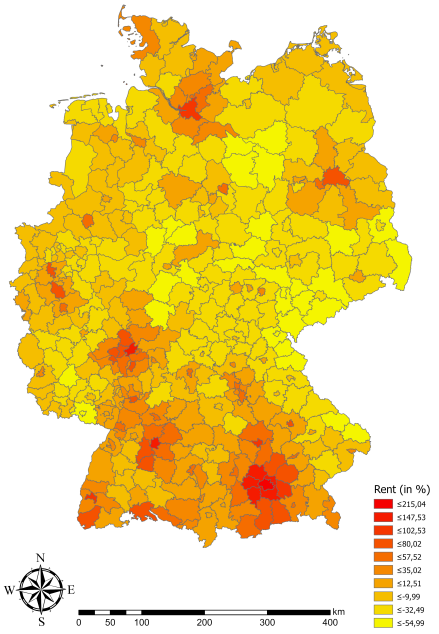
The correlation with the overall consumer price index (CPI) of Weinand and von Auer (2020) including price differences of non-housing goods and services in panel (d) is very similar with 0.8. This is not surprising as cost-of-living differentials are mainly driven by the variation in the cost of housing.

Figure 1.D.15: Alternative price differentials

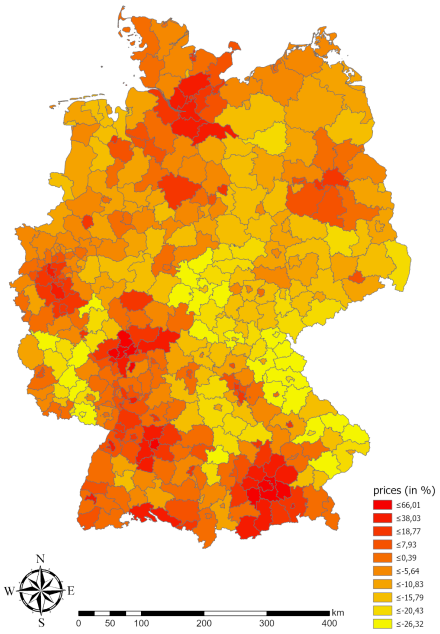
(a) House purchase prices



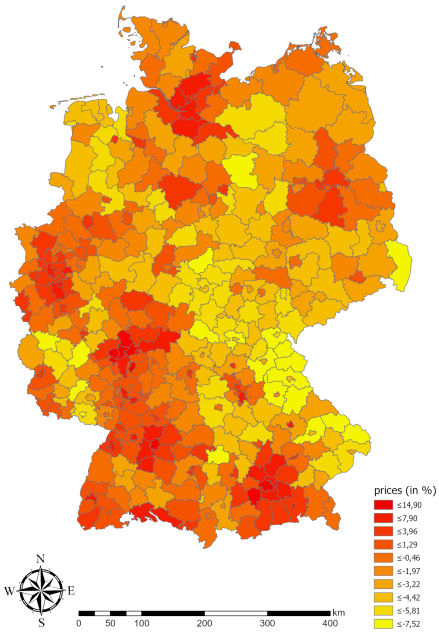
(b) Rents (INKAR)



(c) Housing index



(d) CPI



Notes: Deviation from the national mean in %. (a): Price differentials based on hedonic price regression on house purchase prices of 2017 from the online market platform ImmobilienScout24. (b): Rent differentials based on rent data of 2017 from INKAR (2021). (c) Housing cost differentials using data from Weinand and von Auer (2020). (d): consumer price index from Weinand and von Auer (2020).

Moreover, the CPI has a much smaller range from only -9.6% to 14.9%. This reflects the fact that housing accounts only for a fraction of a household's total spending. As noted above, an average German household spends not more than 15% of the total gross income on housing, which is roughly 30% of the total net income. The rest is used for non-housing goods and services, which have a much smaller spatial price variation (Weinand and von Auer, 2020). Hence, it is not surprising that the CPI has a smaller standard deviation, highlighting the importance of incorporating the expenditure share on housing when calculating Quality of life differentials with Eq. 1.11.

1.E Quality of Life ranking

Table 1.E.1: QoL ranking

Position	County/City code	County/City name	QoL
1	09162	München, Stadt	12.88
2	09182	Miesbach	9.73
3	09179	Fürstenfeldbruck	9.69
4	11000	Berlin, Stadt	8.71
5	09175	Ebersberg	8.66
6	09174	Dachau	8.57
7	09188	Starnberg	8.48
8	09180	Garmisch-Partenkirchen	8.42
9	08311	Freiburg im Breisgau, Stadt	7.99
10	09177	Erding	7.55
11	09184	München	6.64
12	09173	Bad Tölz-Wolfratshausen	6.58
13	03353	Harburg	6.52
14	03355	Lüneburg	6.07
15	02000	Hamburg, Stadt	5.89
16	09163	Rosenheim, Stadt	5.64
17	09187	Rosenheim	5.52
18	01055	Ostholstein	5.27
19	09663	Würzburg, Stadt	5.24
20	08315	Breisgau-Hochschwarzwald	5.17
21	08221	Heidelberg, Stadt	5.03
22	09172	Berchtesgadener Land	4.83
23	09178	Freising	4.8
24	09573	Fürth	4.75
25	09261	Landshut, Stadt	4.72
26	01062	Stormarn	4.53
27	05515	Münster, Stadt	4.53
28	08335	Konstanz	4.49
29	01054	Nordfriesland	4.49
30	07315	Mainz, Stadt	4.39
31	09176	Eichstätt	4.38
32	06412	Frankfurt am Main, Stadt	4.37
33	09772	Augsburg	4.26
34	09771	Aichach-Friedberg	4.25

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
35	05315	Köln, Stadt	4.02
36	09776	Lindau (Bodensee)	4.01
37	08416	Tübingen	3.92
38	09780	Oberallgäu	3.88
39	01057	Plön	3.77
40	09563	Fürth, Stadt	3.7
41	01060	Segeberg	3.7
42	07211	Trier, Stadt	3.68
43	03403	Oldenburg (Oldenburg), Stadt	3.55
44	01056	Pinneberg	3.54
45	09181	Landsberg am Lech	3.53
46	03356	Osterholz	3.52
47	07313	Landau in der Pfalz, Stadt	3.5
48	08336	Lörrach	3.47
49	08111	Stuttgart, Stadt	3.43
50	03458	Oldenburg	3.43
51	09763	Kempten (Allgäu), Stadt	3.38
52	09462	Bayreuth, Stadt	3.36
53	01053	Herzogtum Lauenburg	3.33
54	09565	Schwabach, Stadt	3.24
55	09564	Nürnberg, Stadt	3.24
56	03451	Ammerland	3.15
57	07316	Neustadt an der Weinstraße, Stadt	3.14
58	06432	Darmstadt-Dieburg	3.1
59	08211	Baden-Baden, Stadt	3.08
60	09679	Würzburg	3.06
61	03452	Aurich	3.01
62	03154	Helmstedt	2.89
63	09362	Regensburg, Stadt	2.88
64	09186	Pfaffenhofen a.d.Ilm	2.86
65	06440	Wetteraukreis	2.85
66	03159	Göttingen	2.84
67	01003	Lübeck, Stadt	2.77
68	08316	Emmendingen	2.71
69	06413	Offenbach am Main, Stadt	2.69
70	09678	Schweinfurt	2.68
71	09476	Kronach	2.67
72	08118	Ludwigsburg	2.66
73	06431	Bergstraße	2.61

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
74	09761	Augsburg, Stadt	2.59
75	01059	Schleswig-Flensburg	2.57
76	09473	Coburg	2.53
77	08121	Heilbronn, Stadt	2.52
78	08212	Karlsruhe, Stadt	2.47
79	08337	Waldshut	2.46
80	06439	Rheingau-Taunus-Kreis	2.45
81	06631	Fulda	2.44
82	03453	Cloppenburg	2.44
83	07332	Bad Dürkheim	2.37
84	07338	Rhein-Pfalz-Kreis	2.37
85	09478	Lichtenfels	2.37
86	08436	Ravensburg	2.31
87	03251	Diepholz	2.31
88	09474	Forchheim	2.25
89	03457	Leer	2.24
90	08421	Ulm, Stadt	2.15
91	09461	Bamberg, Stadt	2.15
92	05378	Rheinisch-Bergischer Kreis	2.13
93	06414	Wiesbaden, Stadt	2.13
94	09575	Neustadt a.d.Aisch-Bad Windsheim	2.12
95	09363	Weiden i.d.OPf., Stadt	2.08
96	07331	Alzey-Worms	2.08
97	09263	Straubing, Stadt	2.06
98	05382	Rhein-Sieg-Kreis	2.04
99	01001	Flensburg, Stadt	2.04
100	03401	Delmenhorst, Stadt	2.03
101	06438	Offenbach	2.02
102	09464	Hof, Stadt	2.01
103	01002	Kiel, Stadt	2.0
104	09275	Passau	2.0
105	06531	Gießen	1.99
106	09262	Passau, Stadt	1.9
107	05111	Düsseldorf, Stadt	1.88
108	03460	Vechta	1.86
109	08119	Rems-Murr-Kreis	1.84
110	07235	Trier-Saarburg	1.83
111	03352	Cuxhaven	1.81
112	03158	Wolfenbüttel	1.79

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
113	03361	Verden	1.79
114	09762	Kaufbeuren, Stadt	1.77
115	09661	Aschaffenburg, Stadt	1.77
116	03101	Braunschweig, Stadt	1.75
117	09576	Roth	1.71
118	09190	Weilheim-Schongau	1.7
119	03462	Wittmund	1.69
120	09185	Neuburg-Schrobenhausen	1.67
121	03360	Uelzen	1.65
122	03404	Osnabrück, Stadt	1.64
123	09375	Regensburg	1.63
124	01058	Rendsburg-Eckernförde	1.59
125	09574	Nürnberger Land	1.59
126	12063	Havelland	1.58
127	09671	Aschaffenburg	1.58
128	09278	Straubing-Bogen	1.57
129	07337	Südliche Weinstraße	1.55
130	07131	Ahrweiler	1.53
131	06435	Main-Kinzig-Kreis	1.49
132	05362	Rhein-Erft-Kreis	1.42
133	09189	Traunstein	1.39
134	06534	Marburg-Biedenkopf	1.37
135	01004	Neumünster, Stadt	1.34
136	09273	Kelheim	1.32
137	09471	Bamberg	1.32
138	08226	Rhein-Neckar-Kreis	1.31
139	13075	Vorpommern-Greifswald	1.31
140	07318	Speyer, Stadt	1.3
141	07339	Mainz-Bingen	1.3
142	03151	Gifhorn	1.23
143	14628	Sächsische Schweiz-Osterzgebirge	1.2
144	05314	Bonn, Stadt	1.2
145	03359	Stade	1.16
146	03459	Osnabrück	1.15
147	16053	Jena, Stadt	1.14
148	09274	Landshut	1.14
149	13073	Vorpommern-Rügen	1.12
150	06434	Hochtaunuskreis	1.11
151	08116	Esslingen	1.1

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
152	03241	Region Hannover	1.08
153	09277	Rottal-Inn	1.05
154	08215	Karlsruhe	1.05
155	05774	Paderborn	1.05
156	05334	Städteregion Aachen	1.04
157	09271	Deggendorf	1.01
158	04011	Bremen, Stadt	0.99
159	09373	Neumarkt i.d.OPf.	0.98
160	09475	Hof	0.97
161	14522	Mittelsachsen	0.96
162	08415	Reutlingen	0.94
163	13072	Landkreis Rostock	0.92
164	06411	Darmstadt, Stadt	0.92
165	05154	Kleve	0.87
166	09376	Schwandorf	0.76
167	05166	Viersen	0.76
168	03357	Rotenburg (Wümme)	0.75
169	09571	Ansbach	0.75
170	07340	Südwestpfalz	0.71
171	08317	Ortenaukreis	0.7
172	07111	Koblenz, Stadt	0.67
173	10042	Merzig-Wadern	0.66
174	08231	Pforzheim, Stadt	0.65
175	06436	Main-Taunus-Kreis	0.6
176	03456	Grafschaft Bentheim	0.6
177	06533	Limburg-Weilburg	0.55
178	09472	Bayreuth	0.54
179	03153	Goslar	0.53
180	09777	Ostallgäu	0.52
181	09561	Ansbach, Stadt	0.47
182	09672	Bad Kissingen	0.47
183	09775	Neu-Ulm	0.45
184	09764	Memmingen, Stadt	0.45
185	03254	Hildesheim	0.31
186	09183	Mühldorf a.Inn	0.26
187	12069	Potsdam-Mittelmark	0.24
188	03455	Friesland	0.16
189	14625	Bautzen	0.14
190	14626	Görlitz	0.14

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
191	03354	Lüchow-Dannenberg	0.09
192	05558	Coesfeld	0.09
193	05370	Heinsberg	0.07
194	05566	Steinfurt	0.06
195	08435	Bodenseekreis	0.03
196	14521	Erzgebirgskreis	-0.01
197	07231	Bernkastel-Wittlich	-0.01
198	09372	Cham	-0.02
199	03157	Peine	-0.03
200	08235	Calw	-0.04
201	07319	Worms, Stadt	-0.04
202	07137	Mayen-Koblenz	-0.06
203	08425	Alb-Donau-Kreis	-0.1
204	03257	Schaumburg	-0.12
205	05116	Mönchengladbach, Stadt	-0.12
206	07143	Westerwaldkreis	-0.13
207	06437	Odenwaldkreis	-0.14
208	06632	Hersfeld-Rotenburg	-0.15
209	06433	Groß-Gerau	-0.21
210	08222	Mannheim, Stadt	-0.23
211	08236	Enzkreis	-0.24
212	07335	Kaiserslautern	-0.26
213	03155	Northeim	-0.28
214	05978	Unna	-0.28
215	05711	Bielefeld, Stadt	-0.3
216	09477	Kulmbach	-0.33
217	05170	Wesel	-0.37
218	06611	Kassel, Stadt	-0.42
219	09774	Günzburg	-0.42
220	14729	Leipzig	-0.48
221	05913	Dortmund, Stadt	-0.51
222	06636	Werra-Meißner-Kreis	-0.57
223	03256	Nienburg (Weser)	-0.59
224	09272	Freyung-Grafenau	-0.6
225	09675	Kitzingen	-0.6
226	16064	Unstrut-Hainich-Kreis	-0.62
227	05915	Hamm, Stadt	-0.63
228	03454	Emsland	-0.64
229	12061	Dahme-Spreewald	-0.65

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
230	14627	Meißen	-0.65
231	09676	Miltenberg	-0.66
232	08117	Göppingen	-0.67
233	16074	Saale-Holzland-Kreis	-0.68
234	01051	Dithmarschen	-0.72
235	10046	St. Wendel	-0.77
236	12054	Potsdam, Stadt	-0.77
237	12068	Ostprignitz-Ruppin	-0.79
238	05122	Solingen, Stadt	-0.8
239	09577	Weißenburg-Gunzenhausen	-0.81
240	03358	Heidekreis	-0.83
241	01061	Steinburg	-0.89
242	09276	Regen	-0.9
243	09778	Unterallgäu	-0.9
244	08127	Schwäbisch Hall	-0.94
245	07133	Bad Kreuznach	-0.96
246	12060	Barnim	-0.96
247	12062	Elbe-Elster	-0.98
248	05158	Mettmann	-0.99
249	16075	Saale-Orla-Kreis	-1.0
250	05911	Bochum, Stadt	-1.06
251	05554	Borken	-1.07
252	08128	Main-Tauber-Kreis	-1.09
253	12072	Teltow-Fläming	-1.1
254	07135	Cochem-Zell	-1.16
255	07141	Rhein-Lahn-Kreis	-1.18
256	05162	Rhein-Kreis Neuss	-1.2
257	05366	Euskirchen	-1.2
258	09773	Dillingen a.d.Donau	-1.22
259	16069	Hildburghausen	-1.27
260	09674	Haßberge	-1.31
261	07336	Kusel	-1.31
262	05770	Minden-Lübbecke	-1.36
263	05754	Gütersloh	-1.38
264	05113	Essen, Stadt	-1.39
265	12065	Oberhavel	-1.39
266	07138	Neuwied	-1.41
267	12064	Märkisch-Oderland	-1.44
268	16077	Altenburger Land	-1.47

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
269	16076	Greiz	-1.47
270	16070	Ilm-Kreis	-1.48
271	09374	Neustadt a.d.Waldnaab	-1.48
272	03351	Celle	-1.5
273	05762	Höxter	-1.51
274	08326	Schwarzwald-Baar-Kreis	-1.52
275	09562	Erlangen, Stadt	-1.55
276	05758	Herford	-1.56
277	14523	Vogtlandkreis	-1.6
278	05512	Bottrop, Stadt	-1.61
279	07134	Birkenfeld	-1.62
280	09673	Rhön-Grabfeld	-1.62
281	15091	Wittenberg	-1.72
282	12073	Uckermark	-1.79
283	06535	Vogelsbergkreis	-1.79
284	15082	Anhalt-Bitterfeld	-1.83
285	14612	Dresden, Stadt	-1.83
286	05974	Soest	-1.87
287	07232	Eifelkreis Bitburg-Prüm	-1.87
288	07311	Frankenthal (Pfalz), Stadt	-1.88
289	08225	Neckar-Odenwald-Kreis	-1.88
290	09377	Tirschenreuth	-1.91
291	05570	Warendorf	-1.96
292	08417	Zollernalbkreis	-1.97
293	06532	Lahn-Dill-Kreis	-2.01
294	09779	Donau-Ries	-2.02
295	07233	Vulkaneifel	-2.04
296	10044	Saarlouis	-2.05
297	16051	Erfurt, Stadt	-2.07
298	16071	Weimarer Land	-2.07
299	13071	Mecklenburgische Seenplatte	-2.09
300	05562	Recklinghausen	-2.1
301	10041	Regionalverband Saarbrücken	-2.14
302	06634	Schwalm-Eder-Kreis	-2.17
303	09479	Wunsiedel i.Fichtelgebirge	-2.19
304	12067	Oder-Spree	-2.22
305	07140	Rhein-Hunsrück-Kreis	-2.25
306	09361	Amberg, Stadt	-2.27
307	16061	Eichsfeld	-2.28

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
308	08125	Heilbronn	-2.4
309	07312	Kaiserslautern, Stadt	-2.4
310	09463	Coburg, Stadt	-2.42
311	05766	Lippe	-2.42
312	14730	Nordsachsen	-2.49
313	05954	Ennepe-Ruhr-Kreis	-2.49
314	13003	Rostock, Stadt	-2.53
315	16073	Saalfeld-Rudolstadt	-2.57
316	16068	Sömmerda	-2.6
317	15087	Mansfeld-Südharz	-2.62
318	05358	Düren	-2.63
319	05117	Mülheim an der Ruhr, Stadt	-2.66
320	12070	Prignitz	-2.67
321	03252	Hameln-Pyrmont	-2.72
322	15090	Stendal	-2.75
323	09371	Amberg-Sulzbach	-2.76
324	05114	Krefeld, Stadt	-2.77
325	10043	Neunkirchen	-2.85
326	08437	Sigmaringen	-2.88
327	16066	Schmalkalden-Meiningen	-2.89
328	08135	Heidenheim	-2.9
329	05119	Oberhausen, Stadt	-2.92
330	07132	Altenkirchen (Westerwald)	-2.94
331	15081	Altmarkkreis Salzwedel	-2.97
332	09161	Ingolstadt, Stadt	-3.01
333	08136	Ostalbkreis	-3.01
334	13076	Ludwigslust-Parchim	-3.04
335	05374	Oberbergischer Kreis	-3.05
336	03405	Wilhelmshaven, Stadt	-3.05
337	16052	Gera, Stadt	-3.13
338	13074	Nordwestmecklenburg	-3.14
339	05970	Siegen-Wittgenstein	-3.24
340	06635	Waldeck-Frankenberg	-3.25
341	08237	Freudenstadt	-3.27
342	15084	Burgenlandkreis	-3.31
343	08327	Tuttlingen	-3.34
344	14713	Leipzig, Stadt	-3.36
345	08126	Hohenlohekreis	-3.38
346	16067	Gotha	-3.39

Continued on next page

Table 1.E.1 – *Continued from previous page*

Position	County/City code	County/City name	QoL
347	05966	Olpe	-3.41
348	05124	Wuppertal, Stadt	-3.41
349	05916	Herne, Stadt	-3.45
350	08325	Rottweil	-3.46
351	08426	Biberach	-3.54
352	15086	Jerichower Land	-3.55
353	09572	Erlangen-Höchstadt	-3.56
354	07317	Pirmasens, Stadt	-3.6
355	10045	Saarpfalz-Kreis	-3.62
356	05513	Gelsenkirchen, Stadt	-3.62
357	16055	Weimar, Stadt	-3.74
358	09677	Main-Spessart	-3.81
359	16063	Wartburgkreis	-3.86
360	05120	Remscheid, Stadt	-3.9
361	05958	Hochsauerlandkreis	-3.96
362	14524	Zwickau	-4.02
363	16062	Nordhausen	-4.03
364	12066	Oberspreewald-Lausitz	-4.03
365	05914	Hagen, Stadt	-4.05
366	16072	Sonneberg	-4.09
367	04012	Bremerhaven, Stadt	-4.1
368	16054	Suhl, Stadt	-4.12
369	03255	Holzminden	-4.14
370	15085	Harz	-4.19
371	08216	Rastatt	-4.29
372	15089	Salzlandkreis	-4.38
373	14511	Chemnitz, Stadt	-4.57
374	09171	Altötting	-4.58
375	06633	Kassel	-4.62
376	07320	Zweibrücken, Stadt	-4.71
377	15088	Saalekreis	-4.72
378	05962	Märkischer Kreis	-4.76
379	15002	Halle (Saale), Stadt	-4.78
380	08115	Böblingen	-4.9
381	16065	Kyffhäuserkreis	-4.91
382	07333	Donnersbergkreis	-4.91
383	12052	Cottbus, Stadt	-4.98
384	15001	Dessau-Roßlau, Stadt	-4.99
385	05112	Duisburg, Stadt	-5.0

Continued on next page

Table 1.E.1 – *Continued from previous page*

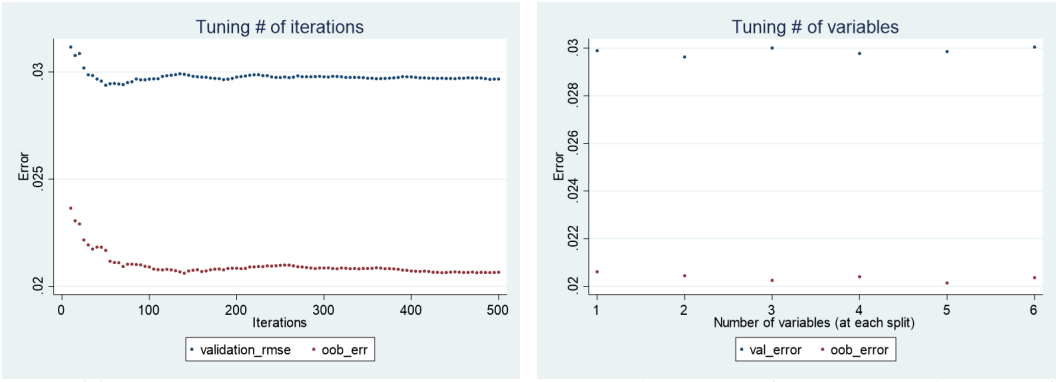
Position	County/City code	County/City name	QoL
386	03402	Emden, Stadt	-5.28
387	15083	Börde	-5.41
388	15003	Magdeburg, Stadt	-5.55
389	05316	Leverkusen, Stadt	-5.65
390	13004	Schwerin, Stadt	-5.69
391	12053	Frankfurt (Oder), Stadt	-5.93
392	03103	Wolfsburg, Stadt	-6.09
393	16056	Eisenach, Stadt	-6.21
394	07334	Germersheim	-6.3
395	12071	Spree-Neiße	-6.86
396	03461	Wesermarsch	-7.15
397	12051	Brandenburg an der Havel, Stadt	-7.2
398	09662	Schweinfurt, Stadt	-8.24
399	07314	Ludwigshafen am Rhein, Stadt	-8.48
400	09279	Dingolfing-Landau	-8.65
401	03102	Salzgitter, Stadt	-10.78

Notes: Full quality of life ranking of all 401 German counties in 2017.

1.F Tuning

Figure 1.F.16: Tuning random forest parameters (PCA full)

(a) Number of iterations (b) Number of variables



Notes: (a): Prediction error depending on the number of iterations (i.e. number of sub-trees). (b): Prediction error depending on the number of variables considered at each split. Both panels belong to the PCA specification using the full sample of indicators.

Breiman (2001) showed that the out-of-bag (OOB) error always converges with an increasing number of iterations. Therefore, I just need to make sure that the number of iterations, that is, the number of grown sub-trees, is large enough to obtain a high predictive accuracy. Panel (a) of Figure 1.F.16 shows that the OOB error already converges after 200 iterations in the full PCA specification. Hence, setting the number of iterations to 500 is sufficient. Panel (b) of Figure 1.F.16 plots the prediction error against the number of variables randomly drawn as candidates for the splitting criteria at each split. With only six inputs, the variation in the OOB error is very moderate. Still, the error is minimized at 5. Results do not differ significantly when considering the test sample-based cross-validation error.

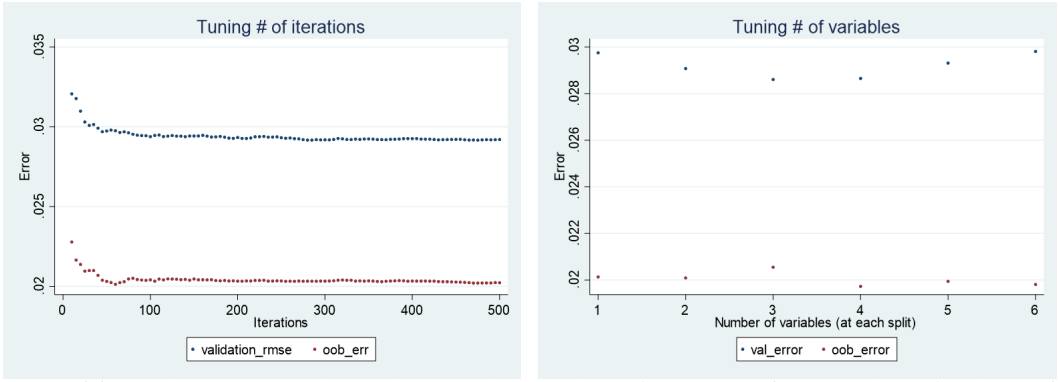
Similarly, Figure 1.F.17 plots the prediction errors for the PCA specification with the sparse list of indicators. Again, the error rates converge after 200 iterations. The optimal number of variables considered at each split is 4.

For the adaptive Lasso specification using the full set of indicators, errors are visualized in Figure 1.F.18. The OOB error converges already after 100 interactions. The number of variables that minimizes the OOB error is 20.

The error behaves similarly when using the sparse list of indicators, as visualized in Figure 1.F.19. Again, the error rates converge at 200 iterations

Figure 1.F.17: Tuning random forest parameters (PCA sparse)

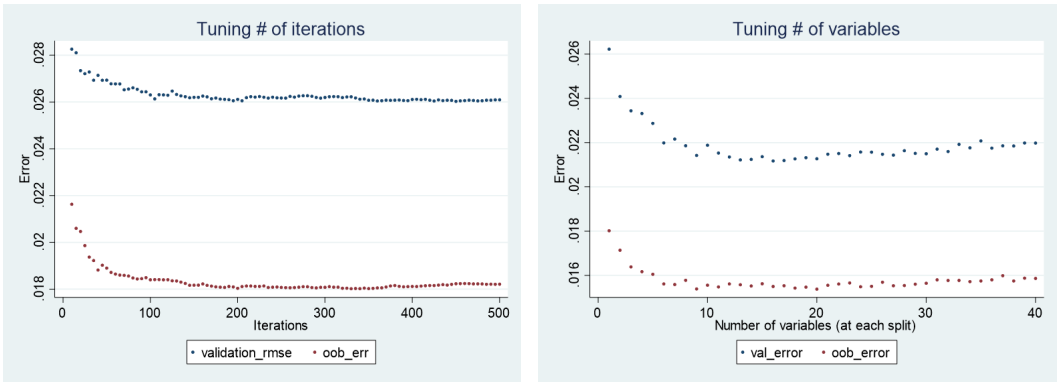
(a) Number of iterations (b) Number of variables



Notes: (a): Prediction error depending on the number of iterations (i.e. number of sub-trees). (b): Prediction error depending on the number of variables considered at each split. Both panels belong to the PCA specification using the sparse sample of indicators.

Figure 1.F.18: Tuning random forest parameters (Lasso full)

(a) Number of iterations (b) Number of variables



Notes: (a): Prediction error depending on the number of iterations (i.e. number of sub-trees). (b): Prediction error depending on the number of variables considered at each split. Both panels belong to the adaptive Lasso specification using the full sample of indicators.

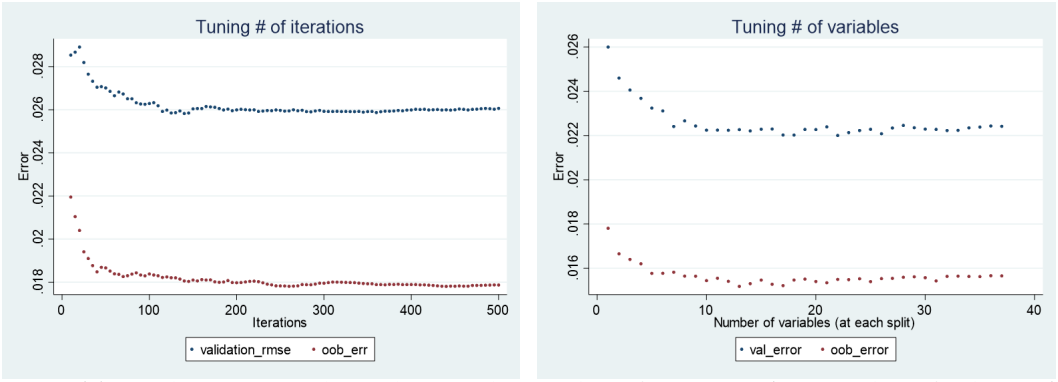
and the optimal number of variables at each split is 13.

1.G Relative importance including control variables

Figure 1.G.20 and Figure 1.G.21 list all variables selected by Lasso of both, the full and the sparse specification, respectively, by the relative importance

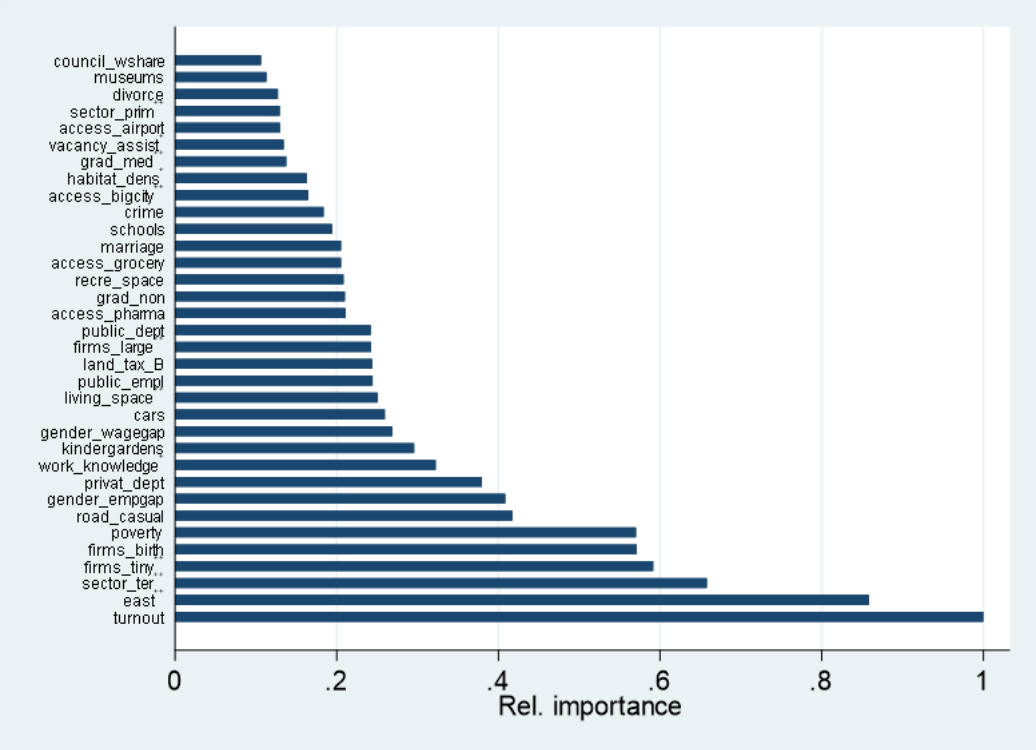
Figure 1.F.19: Tuning random forest parameters (Lasso sparse)

(a) Number of iterations (b) Number of variables



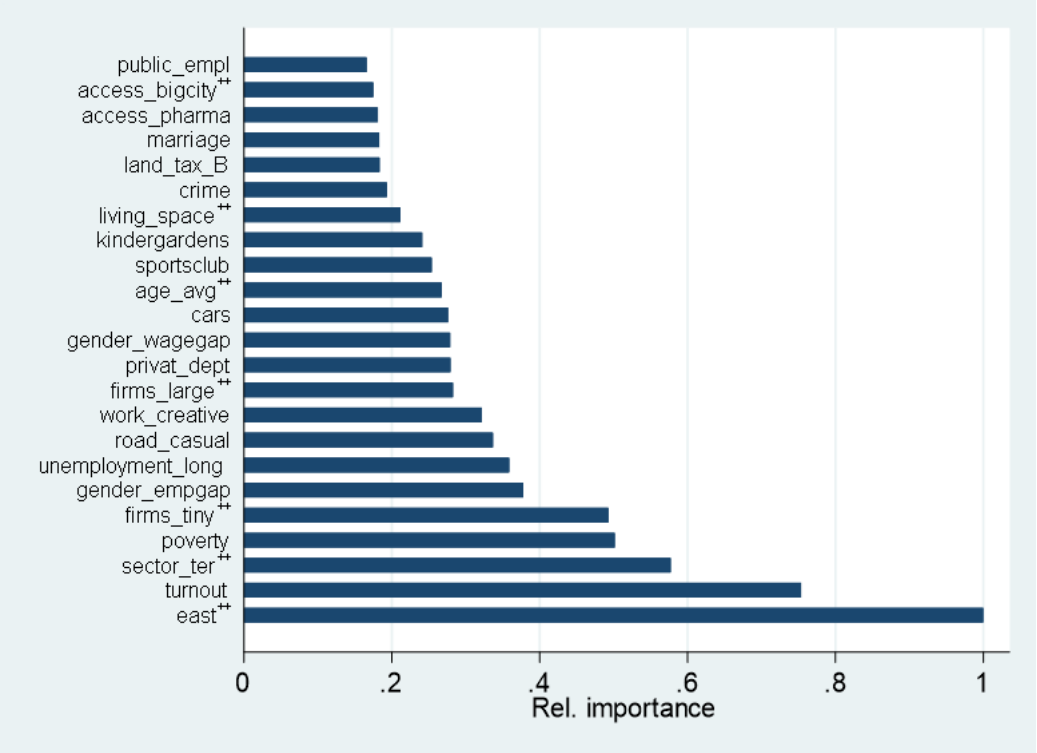
Notes: (a): Prediction error depending on the number of iterations (i.e. number of sub-trees). (b): Prediction error depending on the number of variables considered at each split. Both panels belong to the adaptive Lasso specification using the sparse sample of indicators.

Figure 1.G.20: Relative importance of indicators (Lasso full)



Notes: Relative importance of indicators for predicting QoL differentials using all indicators of column 3 of Table 1.1. Only indicators with a relative importance of above 0.1 are plotted.

Figure 1.G.21: Relative importance of indicators (Lasso sparse)



Notes: Relative importance of indicators for predicting QoL differentials using all indicators of column 3 of Table 1.1 excluding indirect indicators marked with a +. Only indicators with a relative importance of above 0.15 are plotted.

due to random forest prediction of QoL differentials. In contrast to Figure 1.4 and Figure 1.5, controls are not dropped from the graphical representation here. Control variables are marked with ++, including the *east* dummy variable, the most relevant indicator in the sparse specification. However, the control can not be interpreted as amenity indicators directly.

CHAPTER 2

Effects of Access to Universities on Education and Migration Decisions

This chapter is published as Markus, Philipp (2023): "Effects of Access to Universities on Education and Migration Decisions", Ruhr Economic Papers No. 996.

Chapter Abstract

Most studies investigating the causal effect of education on mobility focus on the labor market mobility of high-skilled workers while migration before finishing the educational career is largely ignored. I exploit the exogenous variation in access to universities induced by a large-scale tertiary education expansion reform beginning in the 1970s in Sweden to estimate the impact on college education and migration patterns of high school graduates. Using individual administrative data, I find that a new higher education institution increases college participation rates of local high school graduates by 6.6% while mobility decreases by 10.1% in years after finishing secondary education. In contrast, graduates in the catchment area of the new institution show no change in education outcomes and, if anything, an increased propensity to leave the municipality of high school graduation.

2.1 Introduction

Internal migration is a decisive factor in softening or fostering regional disparities along many dimensions. The flow of workers is important for local labor markets, and the demographic composition of a region's population can shape the provision of amenities and local public goods. Politicians, locally but also at the national level, consider place-based policies as a legitimate way to make certain regions more appealing, either to attract new inhabitants or to make existing residents more likely to stay. One of these place-based policies is to open a new educational institution, which is usually intended to increase local human capital and make a region more attractive, specifically for young adults. However, evidence of the effectiveness of such local policy interventions is very mixed and often depends heavily on internal migration behavior (see Neumark and Simpson (2015) for an overview).

In this paper, I estimate the effects of opening new higher education institutions (HEI) by analyzing a Swedish tertiary education expansion reform beginning in the 1970s. I answer the following questions: Does opening a new HEI makes the local youth more likely to receive college education? And how do migration patterns in that region but also its catchment area change?

Answering these questions is challenging since location and education decisions are known to interact. People move to live closer to an institution providing access to education while the level of education impacts mobility patterns over the life cycle. Especially workers with a college degree tend to be more mobile both at the extensive as well as the intensive margin (see among others DaVanzo, 1978; Corcoran and Faggian, 2017; Plane, 1993). There are some studies investigating the reverse relationship by documenting a negative correlation between the distance to the closest higher education institution and college enrollment (see among others Groen, 2004; Frenette, 2006; Cooke and Boyle, 2011; Alm and Winters, 2009). However, the location of HEI and the place of residence are both likely to be non-random. Some universities were founded hundreds of years ago and have shaped the local demographic and economic development until today. I exploit an arguably exogenous variation in access to higher education by focusing on the openings of new universities and university colleges in Sweden between

1968 and 2012. By using a two-way fixed effects approach I control for time-constant differences between municipalities as well as national time trends. Since the new institutions were founded in different years I use a dynamic Difference-in-Difference (DiD) estimation method, usually referred to as event study (see Roth et al. (2022) and Chaisemartin and D 'Haultfoeuille (2022) for a review of the latest development). It compares individuals from "treated regions" (i.e. municipalities where a university or university college was newly opened within a certain radius) with individuals from "control regions" (i.e. municipalities that never had a higher education institution close by), relative to the difference between those two groups that existed already before the new institution opened.

I have three main findings. First, opening new higher education institutions increases participation in tertiary education by about 6.6% in the same municipality, but there are no spatial spillovers to neighboring regions. Second, young adults are, on average, 10.1% less likely to move out of their home municipality when a new HEI opens in that region. Both effects are significant on all conventional levels. Third, young adults from the catchment area have a slightly but non-significantly increased propensity to move out of their home municipality, where the new nearby HEI is the main destination of migration. Hence, the regional demographic effects of a higher education expansion are very heterogeneous. While municipalities with a new university experience a growing share of young adults due to increased levels of in-migration and lower levels of out-migration, surrounding municipalities see the other side of the same coin: more young adults move away than would without the new HEI close by. I do not find evidence for effects on mobility at later stages of life. There is no significant impact on location choices or total labor market mobility. This has important implications for policymakers that intend to use the founding of a new education institution as a placed-based policy to foster local education outcomes or more general regional development.

There already are some studies that evaluate similar education expansion reforms.[1] Most recently, Berlingieri et al. (2022) find that opening colleges and universities in Germany led to an increasing supply of high-skilled labor without any drops in wages, similar to Carneiro et al. (2023) in Norway. Liu

[1] See Kyvik (2009) for an overview of higher education expansion reforms.

(2015) finds evidence for positive long-run effects on income via general agglomeration economies caused by an increase in population. Besides labor market effects, local innovative activities have been shown to be positively related to higher education expansion reforms in Sweden (Andersson et al., 2009) and Switzerland (Lehnert et al., 2020). Suhonen and Karhunen (2019) find evidence that a Finnish higher education expansion reform increases spillover effects from parental to children's education, while Kamhöfer and Westphal (2019) document negative effects on fertility in Germany. Of course, the direct effect on education outcomes has been studied as well. Frenette (2009) finds comparable positive effects on university attendance among the local youth. Others, like Gibbons and Vignoles (2012), were able to reproduce this result only for low-income households. I contribute to this strand of the literature in three ways. First, I confirm that the distance to the closest HEI matters for participation in university education, even in a context where the monetary costs of moving out are low. Second, I provide evidence that newly opened universities affect not only the educational and economic but also the demographic characteristics of a region via changes in migration patterns. Third, my results emphasize the importance of distance to the new institution and potential negative spillovers to neighboring regions that so far have been overlooked.

This is also relevant for the large literature estimating marginal effects of education, most prominently on wages (e.g. Card, 2001) or non-pecuniary benefits (e.g. Oreopoulos and Salvanes, 2011). There is a tradition of using the proximity to educational institutions as an instrument (see for example Carneiro et al., 2011), based on the assumption that distance to the closest school or college is (negatively) correlated with educational outcomes. My findings of notable effect heterogeneity by distance suggest that the instrument has to be used with a lot of caution for some outcomes.

I also contribute to the large migration literature. Although surveys show that education is one of the main reasons for migration among young adults (Lundholm et al., 2004), most of the papers focus on the role of labor markets as determinants of (internal) migration (see for example Molloy et al., 2011). My results confirm that access to (higher) education affects the migration decisions of young adults as well, especially before entering the labor market.

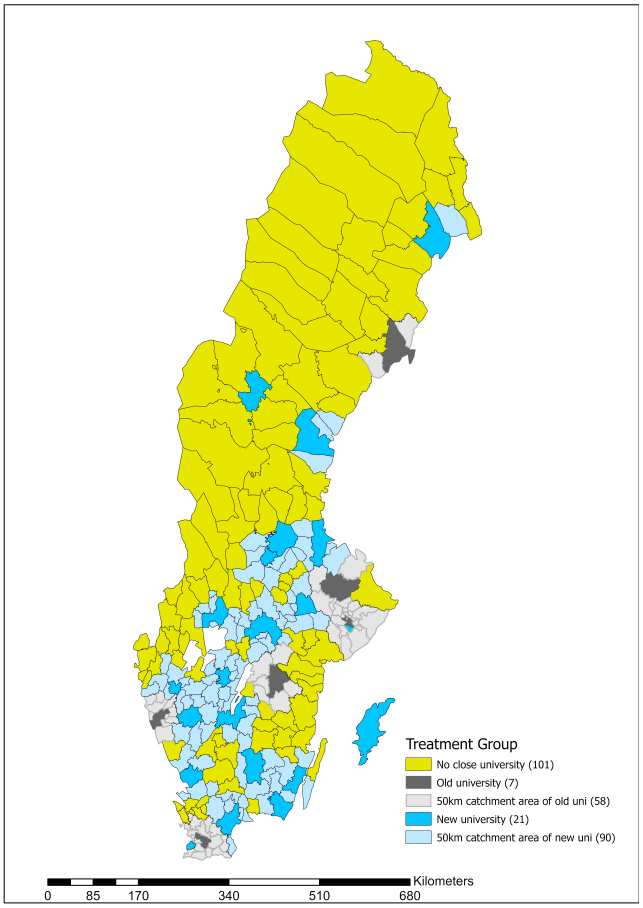
The next section summarizes the institutional background of the Swedish

tertiary education landscape and the expansion reform beginning in the 70s. Section 2.3 describes the data. After discussing the empirical method, I show my main empirical results in section 2.5. The final section concludes.

2.2 Institutional Background

In 1970, 19 higher education institutions (HEIs) operated in seven different municipalities in Sweden. The locations are depicted in dark gray in Figure 2.1. Only six of the 19 institutions were universities providing general

Figure 2.1: Access to tertiary education in Sweden



Notes: Sweden's municipalities with boundaries of 1977 grouped by treatment status defined in section 3.3.3. The number of municipalities of each group is in parentheses.

tertiary education at that time.[2] The other HEIs were more specialized and affiliated with a university. Therefore, they were located in the same municipalities as the general universities. The only exception is the Karolinska Institute in Solna, which is a part of the greater area of Stockholm. In Sweden, tertiary education is provided by universities (*Universitet*) and university colleges (*Högskolan*). In contrast to universities, university colleges do not provide doctoral education. As this difference is of no relevance to the scope of my paper, the terms university, college, and HEI are used interchangeably. I also do not distinguish between specialized universities or those providing education in all academic fields throughout this paper, unless stated otherwise.

Due to the size of the country, having only seven locations offering access to tertiary education means that a substantial part of the population lives relatively far away from a university. In 1970, the (population-weighted) average distance to the next college was over 80 km.[3] Since education including post-secondary education is traditionally free of tuition fees in Sweden (Deen, 2007), the (lack of) geographical access was seen as a major college education friction. Politicians feared that such distances imposed a prohibitive high economic, social or psychological cost to attend college education for some (Premfors, 1984). Therefore and due to the generally increasing number of students, the government decided to establish a significant number of new HEIs in the 70s.[4] The Luleå University of Technology was already founded in 1971. But the substantial change in the higher education landscape in Sweden happened in 1977 when 14 new HEIs were established in 14 different locations where no university was operating before.[5] From 1977 on, there was a total of 22 municipalities with at least one HEI offering access to tertiary education. As intended by the government, this massive expansion more than halved the average distance to the closest college to below 40 km

[2]Uppsala, Lund, Göteborg, Stockholm, and Umeå universities. Linköping university already offered a wide range of programs but got the official status of a university in 1975.

[3]The average distance to the closest HEI over time is also visualized in Figure 2.A.1 in the Appendix.

[4]See e.g. Varga (1998) and Anselin et al. (1997) for a review on the growing number of students.

[5]The 14 municipalities receiving a HEI in 1977 were Jönköping; Vaxjö; Kalmar; Kristianstad; Borås; Skövde; Karlstad; Örebro; Västerås; Falum; Borlänge; Gävle; Sundvall and Östersund.

(see Figure 2.A.1 in the Appendix). Additionally, roughly 40% lived under 50 km away from the next university in 1970 before that share jumped up to almost 70% in 1977.[6]

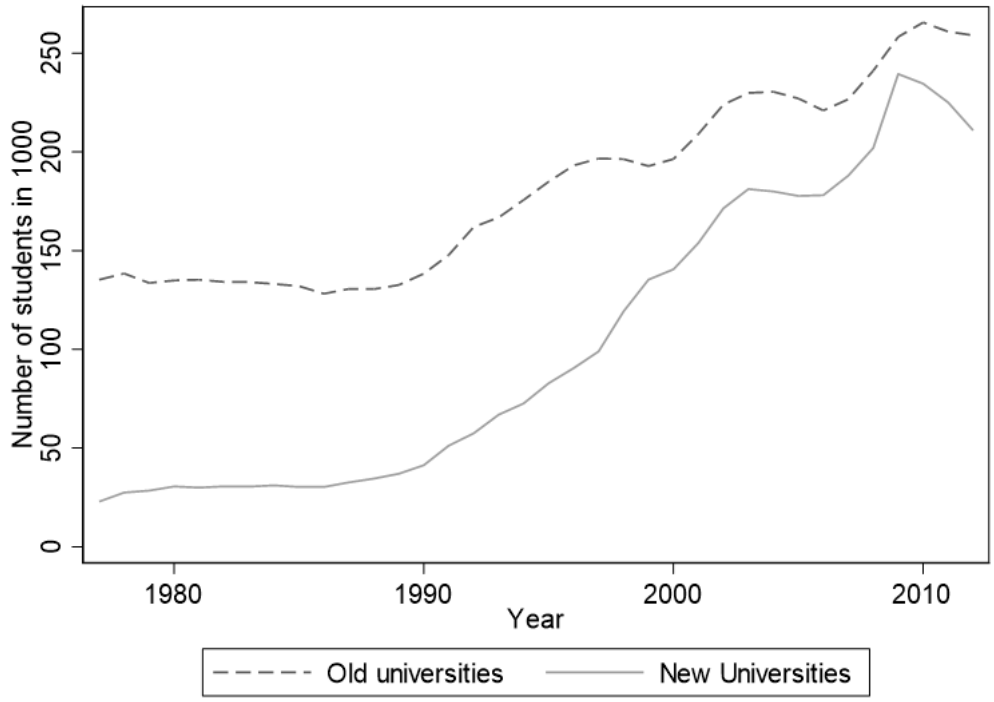
After 1977, six more universities were established, again in municipalities without any HEI until then.[7] Figure 2.1 shows the "new" structure of tertiary education in Sweden where municipalities with new colleges are represented in dark blue. Notation-wise, I will refer to universities that existed already before 1970 as "old" universities and the according municipality as "old uni" municipalities. Equivalently, colleges established after 1970 are termed "new" universities and their municipalities "new uni" municipalities. The "new" colleges started with relatively low numbers of enrolled students as shown in Figure 2.2 and did not start to catch up until the 1990s.

As mentioned above, one of the political goals of the Swedish college expansion reform was to make post-secondary education more accessible, geographically as well as socially (Andersson et al., 2004). The reasoning was to reduce education frictions by lowering the costs of migration or commuting, something that might affect those from lower social classes stronger. The unofficial slogan of the enabling legislation for the initial expansion in 1977 (Swedish Government, 1977) was "En hogskoleenhet i varje ort", which roughly translates to "a unit of higher education in every locality". It can not be ruled out entirely that other factors like regional economic or demographic characteristics have played a role in the location choices of the new institutions. However, interviews with responsible policymakers of that time (Andersson et al., 2009), as well as the reports of the responsible commissions (Premfors, 1984), confirm the hypothesis that geographic dispersion of access to higher education was the primary objective when choosing the locations. Table 2.1 compares "new uni" municipalities (column 3) in 1970 before the first new HEI was opened in 1971, with other potential candidates for a new college, i.e. municipalities without any HEI (columns 4 and 5), as well as with "old uni" municipalities (column 1) and their catchment area (column 2). In line with the above-described objective of dispersion,

[6]The full dynamic is visualized in Figure 2.A.2 in the Appendix.

[7]The universities opened after 1977 are Halmstad University in 1983, Blekinge Institute of Technology in Karlskrona in 1988, University of Trollhättan/Uddevalla in Trollhättan in 1990, Södertörn University in Huddinge, again a part of the larger Stockholm area, and Malmö University as well as Gotland University in 1998.

Figure 2.2: Total number of students enrolled in universities



Notes: "Old" universities were founded before 1968 and "new" universities after 1968.

the new colleges were established in municipalities that had, on average, a high distance to the closest "old" university.[8] When focusing on comparing "new uni" municipalities (column 3) to other municipalities outside of the catchment area of "old" colleges in columns 4 and 5, the distance is not significantly different. In contrast, the "new uni" municipalities' population density is higher and remoteness is lower than in other municipalities without a HEI in 1970. A higher value in remoteness, defined as the sum of population-weighted distances to all municipalities, means that people in that region live relatively far away from the rest of the Swedish population. That emphasizes the importance of municipality-level fixed effect to control for level differences between municipalities, as will be discussed later.

Indicators regarding the population show only little differences. People living in "new uni" municipalities were similar in terms of age, both on

[8]Distance is measured between the centroids or mid-points of the municipalities.

Table 2.1: Population characteristics by treatment group in 1970

	Old (1)	Old Catchment (2)	New (3)	New Catchment (4)	Never uni (5)
Distance to uni (km)	0	27	112	109	146
	(0)	(13)	(71)	(43)	(93)
Pop density	2,142	313	327	39	42
	(1,828)	(530)	(601)	(27)	(74)
Remoteness	310.76	313.91	354.31	307.81	425.56
	(84.7)	(65.97)	(121.55)	(80.71)	(204.79)
Age	31.42	28.6	30.5	31.4	31.93
	(1.47)	(2.8)	(1.11)	(1.12)	(1.57)
Pop share 18y	0.016	0.015	0.017	0.017	0.017
	(0.0008)	(0.0021)	(0.0011)	(0.0013)	(0.0017)
Mobility (past 2y)	0.042	0.073	0.041	0.038	0.035
	(0.199)	(0.259)	(0.2)	(0.192)	(0.183)
College	0.29	0.25	0.24	0.19	0.2
	(0.04)	(0.07)	(0.02)	(0.03)	(0.03)
Individuals	1,318,525	1,129,965	1,409,444	1,273,169	1,605,757
Municipalities	7	58	21	90	101

Notes: Old: Municipalities with universities established before 1968; Old Catchment: Municipalities without university but where a university was established before 1968 within 50 km; New: Municipalities with universities established between 1968-2012; New Catchment: Municipalities without university but where a university was established between 1968-2012 within 50 km; Never uni: No university until 2012. Remoteness is the population-weighted sum of all distances to all other municipalities. High remoteness means that a region is relatively far away from the rest of Sweden's population. Mobility measures the share of people that moved at least once between municipalities in 1968 and 1969. College is the share of the 1970's population that already has a college degree or will attain a college degree at some point. Therefore, it also includes future education outcomes. All variables were calculated on the municipality level and aggregated to the group level as population-weighted averages. Standard deviations in parentheses.

average and in the population share of 18-year-olds.[9] The probability to move at least once during the two years before 1970 was only slightly higher in "new uni" municipalities. Inhabitants of "new uni" municipalities had a higher likelihood of having a college degree at some point in their life compared to those in other municipalities that did not have a university. However, it should be noted that the measure includes future education outcomes, which might include outcomes of the reform already. The used

[9]Comparing Figure 2.A.2 and Figure 2.A.3 in the Appendix provides additional evidence that my study population of high school graduates is similarly distributed as the total population when it comes to distance to the closest HEI.

data will be described in more detail in the next section.

2.3 Data

2.3.1 Individual-Level Data

To investigate the effect of changes in access to higher education, I need to observe individuals before they decide to go to college. In Sweden, students have to choose one of several national programs if they apply for secondary education after the 9th grade (Holmlund, 2021).[10] Until 1993, the vocationally orientated programs (lower secondary education) had a duration of two years while the theoretical track (higher secondary education) took three years, making students eligible to enroll at a university. However, the duration of the vocational programs was extended to three years as well in 1993 such that they also qualify students for accessing higher education (Deen, 2007). Summing up, it has always been necessary to finish 12 years of schooling before entering a university, which is usually the case in the year students turn 19. Hence, my base study population consists of the full Swedish population born between 1950 and 1990 who finished at least higher secondary education.[11] I combine several Swedish administrative records. The data used in this paper comes from the Swedish Interdisciplinary Panel (SIP), administered by the Centre for Economic Demography, Lund University, Sweden.

The register of the total population, available from 1968 onward, includes yearly information on the municipality of residence and links to spouses and parents. The place of residence is a central variable in my analysis. It is defined as the municipality in which the individual was registered by the end of the year. The municipality of residence is also used to calculate the distance to the closest HEI and, therefore, is the determinant of each individual's treatment status (see section 2.3.3 below for more details). One notable

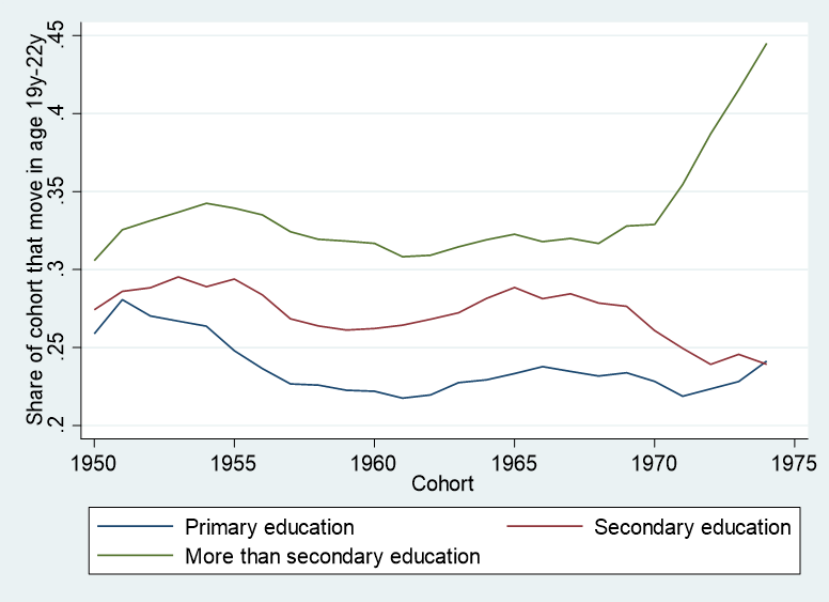
[10] See Fischer et al. (2020) for a more detailed assessment of the Swedish education system before 1950.

[11] The cohort born 1950 turned 18 in 1968, which is the first year where I can observe (among others) the municipality of residence. I cannot use older cohorts, since the first time I observe their place of living is after they (potentially) moved out of their place of schooling.

concern is that people might not live at the place where they are registered. For instance, young adults might move out of their parent's homes without registering the new address. That would result in an underestimation of mobility. However, it should be noted that the law requires citizens to register their place of living and non-compliance is penalized.

Using the information on the municipality of residence, I construct one of my two main outcome variables. To answer the question of whether a change in access to higher education changes migration decisions, my main focus lies on the few years after graduating from high school. During that time, graduates decide whether to apply to a university to attain a college degree and if so, which university they pick. To investigate migration behavior in that crucial period of life, I generate a dummy that equals one if an individual changed her municipality of residence at least once in the four years after finishing secondary education i.e. moving from the place where she lived during high school graduation. The development of mobility during early adulthood is displayed in Figure 2.3 by the level of education. The

Figure 2.3: Early adulthood mobility by level of education



Notes: Share of each cohort that moves at least once between age 19 years - 22 years over time by the highest level of education.

graph confirms the stylized fact that education and mobility are positively

correlated, even before the highest educational degree is completed. In addition, the difference between the educational groups is increasing over time. While the low educated of the cohort born in 1950 are about 5%-points less likely to move as young adults than those who attain a college degree in their life, the difference almost doubles to 9%-points in the cohort born in 1962.

I construct a proxy for migration costs with the provided link to parents and grandparents. Previous research has shown that individuals who live in the same location as their parents and grandparents have a lower probability of moving away, all other factors constant, since the social costs of moving are higher (see for example Mulder and Malmberg (2014) for Sweden). This allows me to split the sample into individuals with low migration costs (without local family ties) and high migration costs (with local family ties). The birth register contains information on the year of birth, gender, and place of birth of the full population.

Data on earnings are taken from the official tax register based on official tax returns. Hence, it is only available on a yearly basis. The exact definition follows Edin and Fredriksson (2000). Consumer Price Index (CPI) adjust incomes to SEK in 2011.

The educational register provides information on the highest achieved level of education. It allows me to distinguish between only primary education without a high school degree (less than 11 years of schooling), secondary education with a high school degree (11-12 years of schooling), and (some) tertiary education, i.e. college education (more than 12 years of schooling). As highlighted above, not every high school degree makes students eligible to enroll in a college. Hence, the group of secondary education is separated into "eligible for college" (12 years of schooling; higher secondary education) and "not eligible for college" (11 years of schooling; lower secondary education) when estimating the treatment effect of the tertiary education reform, where only those eligible for college are considered relevant. However, the decision to finish higher secondary education may depend on access to tertiary education as well. Hence, the two groups with less than 12 years of schooling are used for robustness tests. The education register was recorded in 1990 for the first time and includes all degrees obtained until 2019. Therefore, individuals who died before 1990 are not covered. This is primarily a problem for

the parents of the study population. To complement information on parents' education as well as economical background the Swedish Census of 1970 was added whenever information from the educational register is missing. In contrast to the education register, information in the Census is based on self-enumeration and refers to October 1970 instead of the status in 1990, which might explain little differences. It should be emphasized that the register data provides some information on the time when the highest degree was obtained, but only for a relatively small sub-sample of the total population. Hence, all information on education obtained from the educational register is time-invariant and represents only the highest educational degree. The Census of 1970 does not necessarily contain the highest educational degree, but the highest degree obtained by October 1970. Nevertheless, information from the census is mainly used to complement information for individuals who died before 1990 and therefore can be expected to have completed their educational careers at the time of the census.

2.3.2 University and Municipality Data

To analyze the effect of university openings, the time and location of these openings are crucial. I mainly follow a report of the Swedish National Agency for Higher Education (*Högskoleverket*) on the Swedish higher education landscape from 2006 (Högskoleverket (National Agency for Higher Education), 2006). However, the official founding year is not always the year in which a higher education institution (HEI) becomes a notable provider of tertiary education. For that reason, information from the Higher Education Register (*Högskoleregistret*) on the number of enrolled students and hand collected data was added to determine the de-facto start of a new college. Throughout this paper, a HEI is considered as such if it is labeled as a university or university college in the report of the National Agency for Higher Education, offers (at least) undergraduate education in more than one field, or has more than 500 students enrolled. That excludes specialized HEIs like nursing, military, or theater schools.

The location of the main campus is linked to the respective municipality, where I use centroids to determine the distance to other municipalities and their inhabitants. All municipalities are defined in the administrative

borders of 1977 to keep the borders constant over time. That leaves 277 municipalities of today's 290. In case a university has numerous branches in different locations or more than one campus, only the location of the main campus is considered. The only exception is secondary campuses which were independent universities and continued to host a sizable number of students after a merger. Given that rule, there are no university closures between 1968 and 2012.

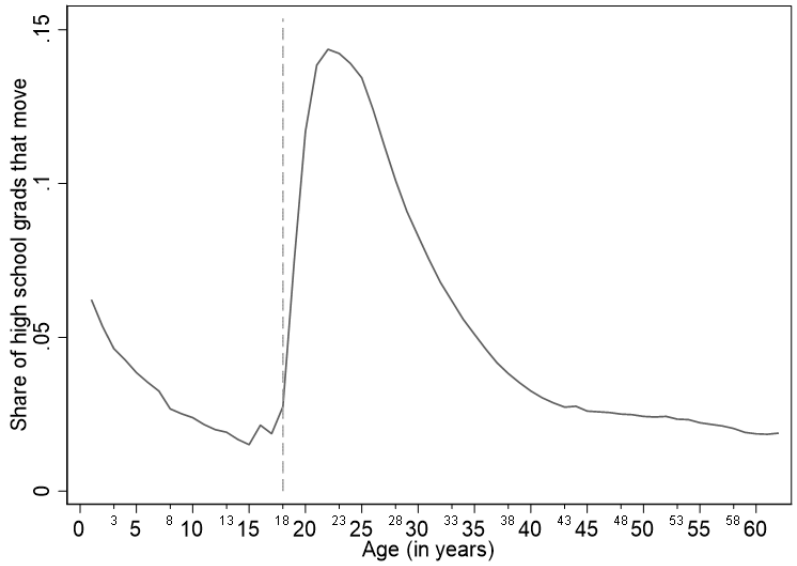
2.3.3 Assigning Treatment Status

An individual is part of the treatment group if she lived in a treated municipality **before** finishing high school. As mentioned above, students usually finish higher secondary education in the year they turn 19. For that reason, individuals are assigned to the treatment or control group according to the treatment status of their place of residence in the year they turn 18 in the main specification. Figure 2.4 visualizes the likelihood to move by age within my sample. Although mobility starts to increase with the age of 15 years already, the big jump is exactly after turning 18 years, supporting my approach to assigning individuals to municipalities in the year of turning 18 years.

The treatment status of a municipality depends on the distance to the closest HEI. In general, there are five different groups of municipalities: (1) "Old uni" municipalities already have a HEI at the beginning of the observation period in 1968 and therefore always have a distance of 0 km. (2) The catchment area of these "old uni" municipalities includes municipalities that have a distance below a certain threshold throughout the whole observation period. These two groups are usually referred to as always-treated municipalities. (3) "New uni" municipalities are municipalities without a university in 1968 but where a new HEI has opened afterward (i.e. the distance to the closest HEI dropped from above 0 km to 0 km).[12] (4) Equivalently to the "old uni" municipalities, the "new uni" municipalities have their catchment area as well. In general, groups 3 and 4 are the treatment regions, depending on

[12]Depending on the size of the catchment area, there are two municipalities (Malmö and Huddinge) that belong, by definition, to group 2 *and* 3. They are classified as part of group 3 in the main specification. Treating them as always-treated (group 2) and dropping them from the sample does not change the results, as shown in section 2.C.

Figure 2.4: Propensity to move by age

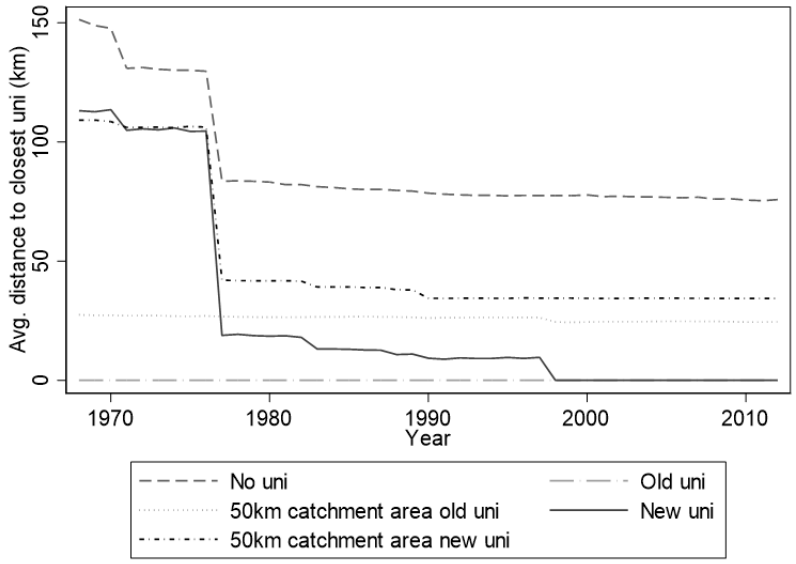


Notes: Propensity to move between municipalities by age. All individuals born between 1950 and 1990 with (at least) secondary education are included.

the specification. (5) Finally, all municipalities that always have a distance above the catchment area threshold are considered never-treated municipalities. Figure 2.1 provides a geographical overview of the five groups for a catchment area threshold of 50 km.

There might be changes in the distance to the closest university for some municipalities of the never-treated group as can be seen in Figure 2.5. However, these changes are considered non-relevant when it comes to education and migration decisions, either because the changes are very small in magnitude or because the distance is very sizable even after the drop. Individuals graduating from high school in these regions are making up the control group. Section 2.C in the Appendix shows that the main results do not depend on the definition of the catchment area threshold. The exact empirical strategy including the underlying assumptions is discussed in section 2.4.

Figure 2.5: Distance to the closest higher education institution

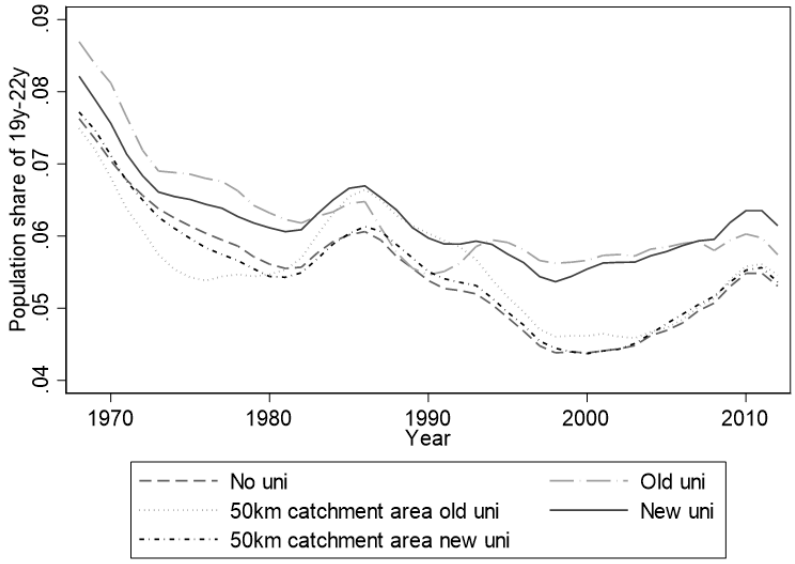


Notes: Population weighted average of the distance to the closest higher education institution, by treatment group. The population weights only use the 18 years old population. The exact definition of the treatment groups is described in section 2.3.3.

2.3.4 Comparing the Treatment Groups

I already used the same definition of treatment groups before in Table 2.1 where I compare the treatment groups before the first "new" university opened in 1971. When comparing the municipalities over time, the distribution of young adults gives some potential descriptive evidence of changes in migration patterns in early adulthood. Figure 2.6 plots the share of 19 - 22 years old residents by treatment groups defined above. In 1970, before any of the new HEI were opened, roughly 8% of the population in "new uni" municipalities were between 19 and 22 years old. That is somewhere between the 7.5% of young adults in other municipalities without a university and the 8.4% in municipalities with an "old" institution. When the trends start to diverge in the 1970s, the share of young adults in "new uni" municipalities begins to converge to the level of "old uni" municipalities and closed the gap already in the 1980s.

Figure 2.6: Population share of young adults



Notes: Population share of young adults (age 19 years - 22 years) over time, by treatment group. The exact definition of the treatment groups is described in section 2.3.3.

2.4 Empirical Strategy

I exploit the variation in access to higher education caused by openings of new universities to provide causal evidence for the relevance of geographical distance to HEIs on the education and migration decisions of young adults. As mentioned above, the higher education expansion led to a massive drop in the average geographical distance to tertiary education.[13] Figure 3.5 documents a notable variation in the change of the distance to the closest HEI between municipalities, which is exploited in this paper. Since outcome variables as well as the treatment status are constant over time, my sample is a repeated cross-section consisting of cohorts of 19-year-old high-school graduates born between 1950 and 1990.[14] As the treatment (the opening of a new college) happened in different years, I estimate the dynamic treatment

[13] See also Figure 2.A.1 in the Appendix.

[14] My period of observation is 1968-2012. For the assignment of the treatment status, I need to observe individuals in the year they turn 18. Therefore, the first cohort I included in the sample is born in 1950. The upper bound of my sample is limited by outcome variables that require me to observe individuals up to the year they turn 22.

effect using a staggered DiD or event-study design. In line with the recent development regarding two-way-fixed-effect literature (see Roth et al. (2022) and Chaisemartin and D 'Haultfoeuille (2022) for an overview), I exclude treated observations from the control group. However, not-yet-treated observations are part of the control group to increase the statistical power. I follow the approach of Gardner (2022) with the event-study regression equation

$$Y_{ist} = \lambda_s + \gamma_t + \sum_{r \neq -1} \beta_r D_{isr} + X_{ist} + \epsilon_{ist}, \tag{2.1}$$

where individual i graduates in municipality s in year t.[15] Y_{ist} is the outcome of interest, for example, a college degree dummy or a dummy indicating whether an individual moved in the year of graduating from high school or three years after. $r \in \{-9, ..., -2, 0, ..., 16\}$ indexes the relative time-wise distance from the treatment and D_{isr} are indicators of treatment adoption.[16] D_{isr} equals one if person i graduates in a treatment municipality r years **after** the new university was opened for $r \geq 0$. Equivalently for the case $r < 0$,

[15] It should be emphasized here that there is a difference between the year where I assign individuals to the place of residence and the year of (potential) treatment. As noted in section 3.3.3, I use the place of living in the year individuals turn 18 to ensure to not pick up a move right after finishing high school. Nevertheless, the actual year of potential treatment is the year t when graduating individuals turn 19, which is the year they finish secondary education. Therefore, the place of residence in the year of turning 18 is a proxy for the place of high school graduation. This is based on the assumption that there is no movement in the year of graduation before secondary education is actually finished. Taking the year of high school graduation as the period of potential treatment is especially important for cohorts turning 19 around the time the HEI is opened and, therefore, for the reference cohort. Take a cohort of high school graduates born in 1968, who turn 18 in 1976 in a municipality that is treated one year later, in 1977. If I used the year where observations turn 18 as the year of (potential) treatment, these individuals would be considered as not-yet-treated since they live in a treatment municipality in the year before the treatment happens, i.e. the university opens. However, that is only true for the year where I assign individuals to the place of residence, not the year where the young adults actually graduate from high school, which is 1977 when the new institution was already operating.

[16] The lower bound of the observation window is limited by the time of the majority of variation, which happens in 1977. The first observed cohort, born in 1950, graduates in 1969, which is eight years before 1977. Coefficients for the relative years before that can be estimated as well, but they rely on a small number of treated observations only. This is an unavoidable problem of event studies that have an unbalanced panel by construction. For that reason, the endpoints of the interval are binned following Schmidheiny and Siegloch (2022): Let l index the non-binned relative time, then $r = -9$ if $l \leq -9$ and $r = 16$ if $l \geq 16$. In contrast to the lower bound, the upper bound of the observation window is limited due to economic reasons. To limit the problem caused by potential general equilibrium effects, the observation window ends after 15 years after the university was opened.

D_{isr} is zero for individuals completing secondary education while living in a treatment municipality $|r|$ years **before** the new university is established. For young adults that graduate from high school in control municipalities, $D_{isr} = 0 \; \forall \; r$. The coefficients β_r capture the dynamic treatment effect of interest for $r \geq 0$, while they can be used as evidence for the plausibility of the parallel trend assumption before the treatment (i.e. when $r < 0$). λ_s and γ_t are municipality and time or cohort fixed effects. The former controls for all time-invariant differences between municipalities, while the latter controls for national trends. Therefore, the coefficients are estimated by exploiting the variation between units *and* time only. Intuitively, coefficient β_r captures the change in the level-difference between treatment and control group r years after (or before if $r < 0$) the intervention, relative to the difference in a reference period. Here, the reference period is $r = -1$, i.e. one year before a new college is opened. Or more precisely, the cohort graduating from high school one year before a new university opens in their place is the reference cohort.

Following the logic of Gardner (2022), I obtain the coefficients by applying a two-stage approach. First, I estimate the model

$$Y_{ist} = \lambda_s + \gamma_t + X_{ist} + \epsilon_{ist} \tag{2.2}$$

on the sample of individuals that lived in control municipalities at the age of 18 (i.e. $D_{isr} = 0 \; \forall \; r$) to obtain the estimated municipality and cohort effects $\hat{\lambda}_s$ and $\hat{\gamma}_t$ while adding individual level controls X_{ist}. In a second step, the adjusted outcomes $Y_{ist} - \hat{\lambda}_s - \hat{\gamma}_t$ are regressed on $D_{isr} \; \forall \; r \in \{-9, ..., 16\}$ to identify the average effects $E(\beta_{isr} | D_{isr} = 1)$. By using that approach, I also control for potentially heterogeneous treatment effects (see Sun and Abraham, 2021; Roth et al., 2022).

Whether the estimates represent causal relationships depends on several identifying assumptions. The most important identifying assumption is the parallel trends assumption which states that the difference (in outcome) between the treatment and control group had to be constant over time in absence of the intervention. Or, in other words, the non-treated counterfactual of the treatment group is assumed to evolve in the same way as the control group. Here, it means assuming that cohorts graduating from high

school next to a new HEI would have made similar education and migration decisions, on average, as students graduating in control municipalities if no new college were opened. There are several arguments why this assumption is likely to hold here.

First, I argue that the new HEIs are as good as randomly assigned to locations in terms of my outcome variables.[17] As described in section 2.2, the locations of the new universities were only chosen according to geographical arguments. Descriptive evidence presented in Table 2.1 indicates that population density might have played a role as well. However, since population density does not vary a lot over time the unit fixed effects absorb these kinds of level differences between municipalities. A comprehensible concern could be that the higher distance to the closest HEI *before* the reform could be correlated to a lower level of college education or a lower population share of young adults since they were forced to move away to study at a university. However, "new uni" municipalities are not significantly different along those dimensions, including the average distance to the closest university as displayed in Table 2.1. The quasi-random choice of locations provides an argument that there are no systematic differences between treatment and control regions. Therefore, I conclude that the allocation of the intervention (i.e. the location of HEI) were not depending on any of my outcomes. Second, the rich data set allows the estimation of the so-called "pre-trend". An insignificant treatment effect *before* the intervention provides additional evidence that the parallel trend assumption is satisfied. As will be visualized in the next section, the estimates for cohorts before a new institution is opened are mostly 0 for all outcomes and specifications. Third, the fixed effects absorb all observed and unobserved time-constant differences at the municipality level and any kind of national trend that affects all municipalities similarly. In addition, I control for differences at the individual level like the parents' education and differences in migration costs by including family ties. However, my results do not depend on including these controls as shown in section 2.C in the Appendix.

Another important assumption is the Stable Unit Treatment Value As-

[17] Studies that use the same (Andersson et al., 2009; Nybom et al., 2022) or a similar (Suhonen and Karhunen, 2019; Berlingieri et al., 2022; Carneiro et al., 2023; Lehnert et al., 2020; Frenette, 2009) reform to obtain causal estimates use the same argument.

sumption (SUTVA) (Rubin, 1980). SUTVA is sometimes described as the assumption that each unit, including units from the control group, is only affected by its own treatment status. This implicitly rules out the relevance of general equilibrium and (spatial) spill-over effects. To minimize the influence of (local) general equilibrium effects I restrict the effect window to 15 years after a new institution was established. Since a new HEI might not only affect the access to education but also the local labor market, demographics, or local amenities of the location in the long run, the direct effect of the reform becomes harder to measure over time.[18] The exclusion of spill-over effects is traditionally problematic in spatial analysis (see Butts, 2021). A new HEI likely not only affects the location itself but also other regions close by since students could commute to the new college or consider moving closer to the location which they would not have done without the new university. For that reason, I exclude the catchment area of "new uni" municipalities from the control group. Instead, several specifications are estimating the treatment effect for the catchment area while excluding the "new uni" municipalities themselves. To identify the reach of spill-over effects, I vary the size of the catchment area in Appendix section 2.C. For "old uni" municipalities and their catchment areas, there is no change in the distance to the closest HEI, so it seems unlikely that there are spill-over effects from new universities opened further away than the one that was already there. However, students that would have enrolled at an "old" university may decide to apply at a new college, leaving an additional university place open at the "old" HEI. Also, academic staff like professors might relocate their workplace to a new college, which might also affect education and migration decisions of the local youth close to the "old" HEI. Hence, my main specification excludes the "old uni" municipalities (and their catchment area). To avoid the contamination of my control group by already treated units, I follow the latest development in the DiD / event-study literature by excluding treated observations from my control group (see Goodman-Bacon

[18] Before the intervention, the effect window consists of nine years, since that is the maximum number of years I can observe cohorts graduating from high school before the 1977-reform. Every (relative) year outside of that chosen effect window is included in the estimation by binning at the endpoints according to Schmidheiny and Siegloch (2022). This implies that there must be no university openings before or after my observation window, which is unproblematic as I can observe the full universe of universities in Sweden.

(2021) for an intuitive explanation of that issue). Finally, my control group consists of all high school graduates in municipalities that have a distance higher than the catchment area cutoff at the time of graduation. That includes both individuals in "never-treated" municipalities (treatment group 1) over the full observation period from 1968 to 2008 as well as high school graduates in treated municipalities (treatment groups 4 and 5) before the new university was opened. Note that, as shown in Figure 2.5, some of the so-called "never-treated" municipalities experience a drop in the distance to the closest HEI due to the reform as well. For the SUTVA to hold, I assume that a drop to a distance of more than the cutoff, for example, a drop from 120 km to 90 km, does not affect the migration and education decisions of young adults.[19]

If the above-mentioned assumptions hold, my estimates represent the Average Treatment Effect on the Treated (ATT) (Lechner, 2011). The estimated coefficients have to be interpreted relative to the reference period one year before a new HEI was opened and indicate the change in the difference between average outcome levels of the treatment and control group.

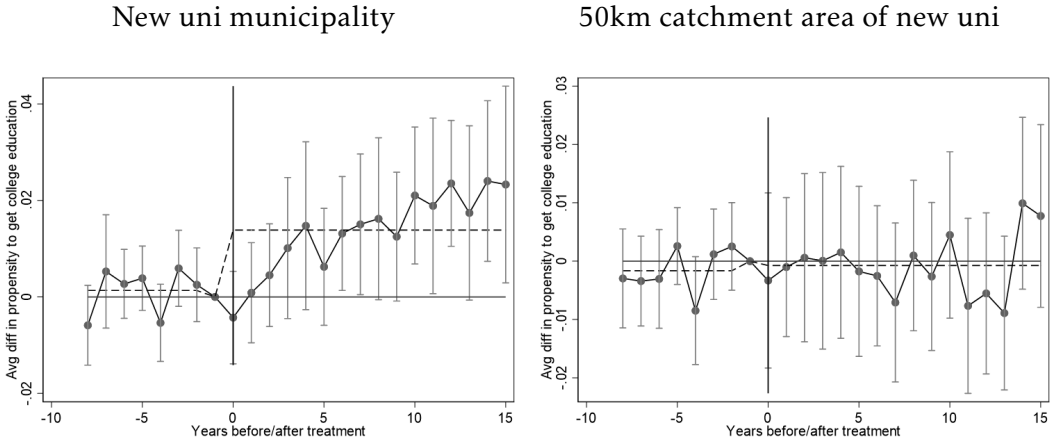
2.5 Main Results

2.5.1 College education

The first question that arises when investigating the effect of a higher education expansion reform is whether it had an effect on individuals' decisions to enroll at a HEI. Figure 2.7 shows the estimates of the main specification for "new uni" municipalities on the left-hand side. Individuals graduating from high school in "new uni" municipalities after a new college was opened have, on average, a 1.4%-points increased propensity to attain a college degree compared to the difference between treatment and control cohorts before the intervention. This corresponds to an average expansion in college education by 6.6%. The effect is increasing over time and becomes statistically significant on conventional levels five years after the treatment. The dynamics

[19] Varying the threshold of the catchment area definition provides evidence that there is no treatment effect on municipalities beyond 50 km. You find more details on that in section 2.C in the Appendix.

Figure 2.7: Propensity to obtain a college degree



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

can be explained by the growing number of enrolled students shown in Figure 2.2. The sign of the effect is in line with previous findings of Alm and Winters (2009), Frenette (2004), Sá et al. (2006), and Jepsen and Montgomery (2009) who documented a negative relationship between the distance to the closest university and tertiary education participation rates. Frenette (2009), who similarly investigates the effect of a reduction in the distance due to newly opened universities, finds comparable results also in magnitude.
However, I only find a negative relationship between geographical distance and participation rates in "new uni" municipalities and small and/or very close neighboring regions.[20] In contrast, the 50km catchment area of "new uni" municipalities shows no effect of a new HEI on college education as visualized on the right-hand side of Figure 2.7. It seems that the existence of the new universities close by did not change the decision of whether to apply for

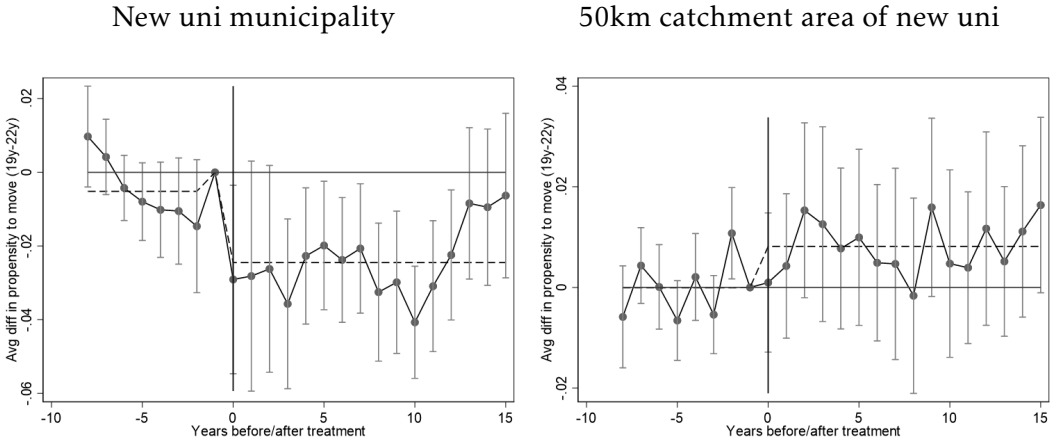
[20]As shown in section 2.C in the Appendix, there is a significant positive effect on college attendance rates in the 30km catchment area. That means that a drop in the distance to the closest university from above 30km to below 30km increases the likelihood of local high school graduates participating in tertiary education, even if the new institution was opened in the neighboring municipality.

college or not during the first 15 years after opening. That could be explained in two ways. On the one hand, the distance to the closest university might not have been an incremental part of the costs of attending college for high school graduates in the catchment area. This explanation seems unlikely given the fact that the new university has an effect on college education in "new uni" municipalities and smaller catchment areas. On the other hand, it might be the case that the improvement in access to tertiary education is not enough to overcome the prohibitive costs in the 50km catchment area, while it is sufficient for some in the "new uni" municipalities. Note that the result does not necessarily mean that there was no impact on high school graduates in the catchment area at all. First, results only indicate that the effect does not differ between graduates in treated and control municipalities. Second, it is still possible that the new HEI acted as a substitute for old universities, i.e. that high school graduates that would have applied for a place at an old institution chose the new local college instead to attain college education. Previous publications have already shown that geographic distance is an important determinant for institution choice (see for example Gibbons and Vignoles, 2012; Griffith and Rothstein, 2009).

2.5.2 Short-Term Mobility

The effects of the college expansion reform on the mobility of young adults at the age of 19-22 years are heterogeneous with regard to geographical distance to the new institution. I find negative and partially statistically significant effects for individuals graduating from high school in "new uni" municipalities as depicted in Figure 2.8 on the left-hand side. Compared to the control group, the new HEI makes young adults 2.4%-points less likely to move away, on average, which corresponds to a reduction of the propensity to move by around 10.1%. The sign is in line with the intuitive expectation. High school graduates that would have moved away to attend college have the opportunity to stay in their home region after the new HEI opened. However, the effect is not persistent, starts to decline ten years after the HEI opened and even becomes insignificant by the end of my observation period. For students finishing high school in the catchment area of a "new uni" municipality, results differ notably. For cohorts finishing high school

Figure 2.8: Propensity to move with 19y-22y

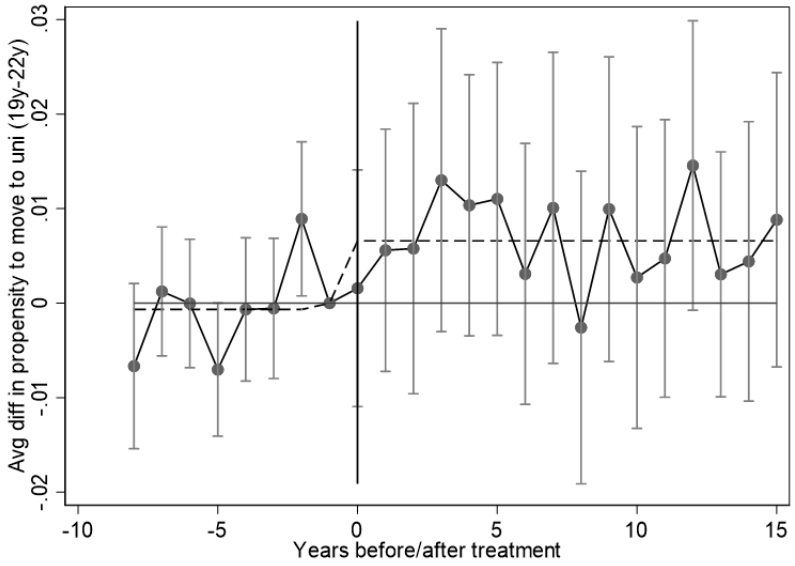


Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

within 15 years after a new HEI opened within a distance of 50km (but still in a different municipality), I find an increased probability to move out of the municipality at age 19 years to 22 years of 0,8%-points on average as visualized in the right graph of Figure 2.8. Even though the estimated effect is not statistically different from zero, the reverse sign of the coefficients compared to the "new uni" municipality specification is striking. While high school graduates from "new uni" municipalities are less likely to move away, young adults in the catchment area show, if anything, a higher propensity to move.

The results are similar when considering moves from the catchment area to the close municipality with the new college rather than any move out of the municipality of graduation as before, visualized in Figure 2.9. Young adults are more likely not just to leave their home region in the catchment area of a new local center of tertiary education, but they are indeed moving towards this new local center.

Figure 2.9: Propensity to move towards a university with 19y-22y in the catchment area



Notes: Treatment status assignment by treatment status of municipality of residence at age of 18 years. Treatment: drop in distance to closest higher education institution from above 50km to below 50km, excluding municipalities that got a new university. Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

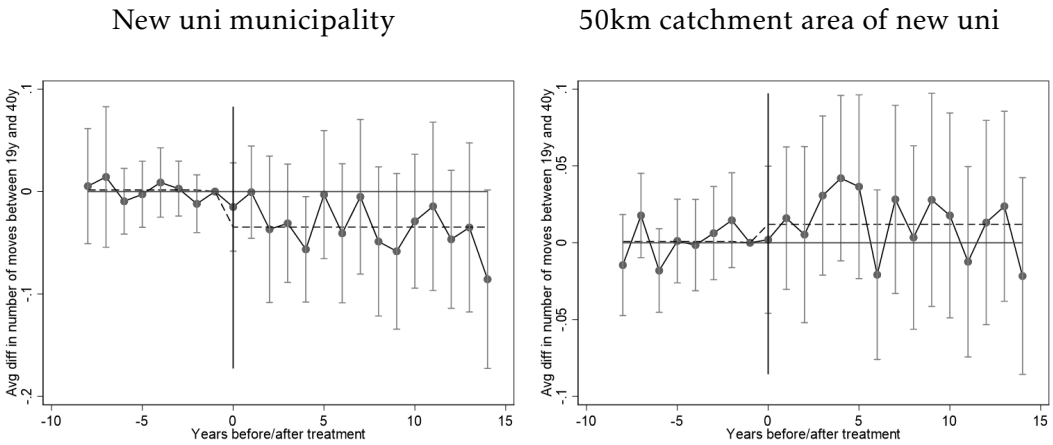
2.5.3 Long-Term Mobility

To investigate mobility at later stages of the life cycle, I need to observe individuals longer than for the short-term mobility outcome above. Therefore, I restrict my sample further to cohorts born between 1950 and 1972 to be able to track every individual's place of residence up to the age of 40 years.[21] Figure 2.10 shows the treatment effect on the number of moves between municipalities in the age of 19-40 years both for "new uni" municipalities and their 50km catchment area. While the sign is in line with results for short-term mobility presented above, the size of the estimates is very close to

[21]That means, the last cohorts of the restricted sample turns 18 years in 1990, which is 13 years after the majority of new institutions were opened in 1977. For that reason, the observation window is also reduced to 13 years, while still being binned at the endpoints.

zero and statistically insignificant. Results are similar when estimating the effects on the distance between the place of residence at age 18 and 30 years using the same non-restricted sample as in section 2.5.2 (see Figure 2.B.4 in the Appendix). Even if effects on individuals are only short-term with

Figure 2.10: Total number of moves with 19 - 40 years



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

regard to migration, it does not mean there is no long-term impact. Since the out-migration of young adults is reduced while the in-migration of young adults increases, the population share of that group increases in ”new uni” municipalities over time. Therefore, my results can explain the convergence of the share of young adults in ”new uni” municipalities to the level of ”old uni” municipalities described in Figure 2.6.

2.6 Conclusion

In this paper, I investigate the effects of access to tertiary education on the education and migration decisions of young adults in Sweden. To do so, I collect data on all universities and university colleges in Sweden, especially on those opened after 1968, and provide evidence that the change in access

to education induced by the opening of new higher education institutions (HEIs) can be exploited to obtain causal estimates. I can conduct an event study on individual-level outcomes by combining several administrative Swedish register data sources.

My results indicate that there is a positive impact of opening a new HEI on college attainment rates among the local youth. The positive effect is limited to high school graduates who finished secondary education in municipalities that received a new HEI. I do not find strong evidence for spatial spill-over effects to surrounding municipalities when it comes to educational outcomes, even though there was an improvement in access to tertiary education as well.

The effects on internal migration behavior differ between municipalities with a new HEI and their catchment areas, too. While a new university makes high school graduates in the same municipality less likely to move away, young adults become, if anything, more mobile after finishing high school in surrounding municipalities. The slight increase in the propensity to move is driven by migration toward universities.

For "new uni" municipalities, the results are in line with intuitive expectations: the new HEI reduced the costs of attending college and therefore increased the propensity of receiving tertiary education in that location. At the same time, high school graduates who would have moved away to another university had the opportunity to enroll at a HEI without leaving their home region, reducing mobility at the time of studying. However, I find evidence for long-term effects neither on location choices nor on overall labor-market mobility.

For the catchment area of "new uni" municipalities, effects are notably different. Although these regions experienced a drop in the distance to the closest HEI to below 50km, there was no effect on educational outcomes. Nevertheless, graduates showed a slight increase in the likelihood to move away to municipalities with universities. There are some candidates to explain how new universities attract young adults besides the direct educational channel. First, the new institution could have changed the local labor market. That could have happened either directly with the university as an employer but also indirectly via local multipliers (see Moretti, 2010).[22] Second, if friends

[22]First estimations of income effects show no statistically or economically significant

from the same cohort moved to the neighboring "new uni" municipality to study instead of moving to an institution relatively far away, peer effects can impact graduates that do not enroll at a university themselves. Third, the growing population share of young adults might increase the supply of local amenities like a more dynamic nightlife or a bigger marriage market (Shapiro, 2006).[23] In any case, further research has to investigate whether effects on the local economy, social factors, or changes in local amenities could explain the change in migration patterns.

My paper emphasizes the importance of geographical distance when evaluating place-based policies. The measured effects of the reform are highly localized when it comes to educational outcomes. In terms of migration, the targeted region seems to benefit from the existence of a HEI while neighboring municipalities are, if anything, worse off. Therefore, policymakers should be aware of geographical limits as well as potentially negative spatial spillover effects when using place-based policies to foster the attractiveness of a specific location.

treatment effect as shown in section Figure 2.B in the Appendix. However, further research is necessary for a well-founded conclusion.

[23] Results for the population of lower educated young adults, who are not eligible to enroll in a HEI, provide additional evidence that these indirect effects play a role, especially in the catchment area (see section 2.B in the Appendix).

Bibliography

Alm, J. and Winters, J. V. (2009). Distance and intrastate college student migration. *Economics of Education Review*, 28(6):728–738.

Andersson, R., Quigley, J. M., and Wilhelmson, M. (2004). University decentralization as regional policy: the Swedish experiment. *Journal of Economic Geography*, 4(4):371–388.

Andersson, R., Quigley, J. M., and Wilhelmsson, M. (2009). Urbanization, productivity, and innovation: Evidence from investment in higher education. *Journal of Urban Economics*, 66(1):2–15.

Anselin, L., Varga, A., and Acs, Z. (1997). Local Geographic Spillovers between University Research and High Technology Innovations. *Journal of Urban Economics*, 42(3):422–448.

Berlingieri, F., Gathmann, C., and Quinckhardt, M. (2022). College Openings and Local Economic Development. *IZA Discussion Paper*, 15364.

Butts, K. (2021). Difference-in-Differences Estimation with Spatial Spillovers. *arXiv working paper*, 2105.03737.

Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, 69(5):1127–1160.

Carneiro, P., Heckman, J. J., and Vytlacil, E. J. (2011). Estimating marginal returns to education. *American Economic Review*, 101(6):2754–2781.

Carneiro, P., Liu, K., and Salvanes, K. G. (2023). The Supply of Skill and Endogenous Technical Change: Evidence from a College Expansion Reform. *Journal of the European Economic Association*, 21(1):48–92.

Chaisemartin, C. D. and D 'Haultfoeuille, X. (2022). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey. *NBER working paper*, 29691.

Cooke, T. J. and Boyle, P. (2011). The Migration of High School Graduates to College. *Educational Evaluation and Policy Analysis*, 33(2):202–213.

Corcoran, J. and Faggian, A. (2017). Graduate migration and regional development: An international perspective. *Graduate Migration and Regional Development: An International Perspective*, pages 1–10.

DaVanzo, J. (1978). Does Unemployment Affect Migration ? Evidence from Micro Data. *The Review of Economics and Statistics*, 60(4):504–514.

Deen, J. (2007). Higher education in Sweden. Technical report, Enschede.

Edin, P.-A. and Fredriksson, P. (2000). LINDA-Longitudinal Individual Data for Sweden. Working paper, Uppsala University, Uppsala.

Fischer, M., Karlsson, M., Nilsson, T., and Schwarz, N. (2020). The Long-Term Effects of Long Terms – Compulsory Schooling Reforms in Sweden. *Journal of the European Economic Association*, 18(6):2776–2823.

Frenette, M. (2004). Access to college and university: Does distance to school matter? *Canadian Public Policy*, 30(4):427–442.

Frenette, M. (2006). Too far to go on? Distance to school and university participation. *Education Economics*, 14(1):31–58.

Frenette, M. (2009). Do universities benefit local youth? Evidence from the creation of new universities. *Economics of Education Review*, 28(3):318–328.

Gardner, J. (2022). Two-stage differences in differences. *arXiv working paper*, 2207.05943.

Gibbons, S. and Vignoles, A. (2012). Geography, choice and participation in higher education in England. *Regional Science and Urban Economics*, 42(1-2):98–113.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Griffith, A. L. and Rothstein, D. S. (2009). Can't Get Here from There: The Decision to Apply to a Selective Institution. *Economics of Education Review*, 28(5):620–628.

Groen, J. A. (2004). The effect of college location on migration of college-educated labor. *Journal of Econometrics*, 121(1-2):125–142.

Högskoleverket (National Agency for Higher Education) (2006). Högre utbildning och forskning 1945–2005 – en översikt. Technical report, Högskoleverket (National Agency for Higher Education), Stockholm.

Holmlund, H. (2021). A researcher's guide to the Swedish compulsory school reform. *Journal of the Finnish Economic Association*, 1(1):25–50.

Jepsen, C. and Montgomery, M. (2009). Miles to go before I learn: The effect of travel distance on the mature person's choice of a community college. *Journal of Urban Economics*, 65(1):64–73.

Kamhöfer, D. A. and Westphal, M. (2019). Fertility effects of college education: Evidence from the German educational expansion. DICE Discussion Paper 316, Düsseldorf.

Kyvik, S. (2009). Geographical and Institutional Decentralisation. In Kyvik, S., editor, *The Dynamics of Change in Higher Education. Higher Education Dynamics*, pages 61–80. Springer, Dordrecht, 27 edition.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, 4(3):165–224.

Lehnert, P., Pfister, C., and Backes-Geller, U. (2020). Employment of R&D personnel after an educational supply shock: Effects of the introduction of Universities of Applied Sciences in Switzerland. *Labor Economics*, 66(101883).

Liu, S. (2015). Spillovers from universities: Evidence from the land-grant program. *Journal of Urban Economics*, 87:25–41.

Lundholm, E., Garvill, J., Malmberg, G., and Westin, K. (2004). Forced or free movers? The motives, voluntariness and selectivity of interregional migration in the Nordic countries. *Population, Space and Place*, 10(1):59–72.

Molloy, R., Smith, C. L., and Wozniak, A. (2011). Internal migration in the United States. *Journal of Economic Perspectives*, 25(3):173–196.

Moretti, E. (2010). Local Multipliers. *American Economic Review: Papers & Proceedings*, 100(2):373–377.

Mulder, C. H. and Malmberg, G. (2014). Local ties and family migration. *Environment and Planning A*, 46(9):2195–2211.

Neumark, D. and Simpson, H. (2015). Place-Based Policies. In Duranton, G., Henderson, J. V., and Strange, W. C., editors, *Handbook of Regional and Urban Economics*, volume 5, chapter 18, pages 1197–1287. Elsevier.

Nybom, M., Plug, E., van der Klaauw, B., and Ziegler, L. (2022). Skills, Parental Sorting, and Child Inequality. *IZA Discussion Paper*, 15824.

Oreopoulos, P. and Salvanes, K. G. (2011). Priceless: The nonpecuniary benefits of schooling. *Journal of Economic Perspectives*, 25(1):159–184.

Plane, D. A. (1993). Demographic Influences on Migration. *Regional Studies*, 27(4):375–383.

Premfors, R. (1984). Analysis in politics: The regionalization of Swedish higher education. *Comparative Education Review*, 28(1):85–104.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2022). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. *arXiv working paper*, 2201.01194.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Sá, C., Florax, R. J. G. M., and Rietveld, P. (2006). Does Accessibility to Higher Education Matter? Choice Behaviour of High School Graduates in the Netherlands. *Spatial Economic Analysis*, 1(2):155–174.

Schmidheiny, K. and Siegloch, S. (2022). On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization. *ECONtribute Discussion Papers Series*, 201.

Shapiro, J. M. (2006). Smart Cities: Quality of Life, Productivity, and the Growth Effects of Human Capital. *The Review of Economics and Statistics*, 88(2):324–335.

Suhonen, T. and Karhunen, H. (2019). The intergenerational effects of parental higher education: Evidence from changes in university accessibility. *Journal of Public Economics*, 176:195–217.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

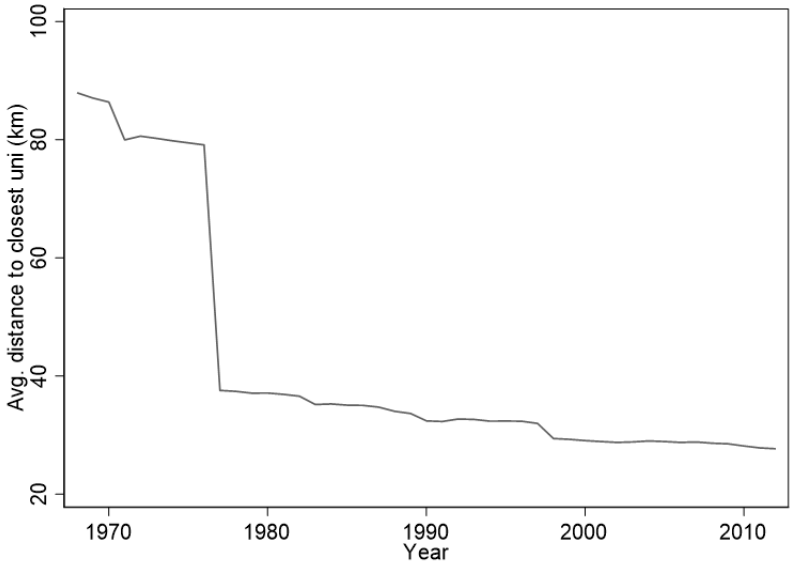
Swedish Government (1977). Regeringens Proposition 1976/77: 59. Technical report, Stockholm.

Varga, A. (1998). *University Research and Regional Innovation: A Spatial Econometric Analysis of Academic Technology Transfers*. Kluwer Academic Publishers, Dordrecht.

Appendix

2.A Distance to Closest HEI

Figure 2.A.1: Distance to the closest university



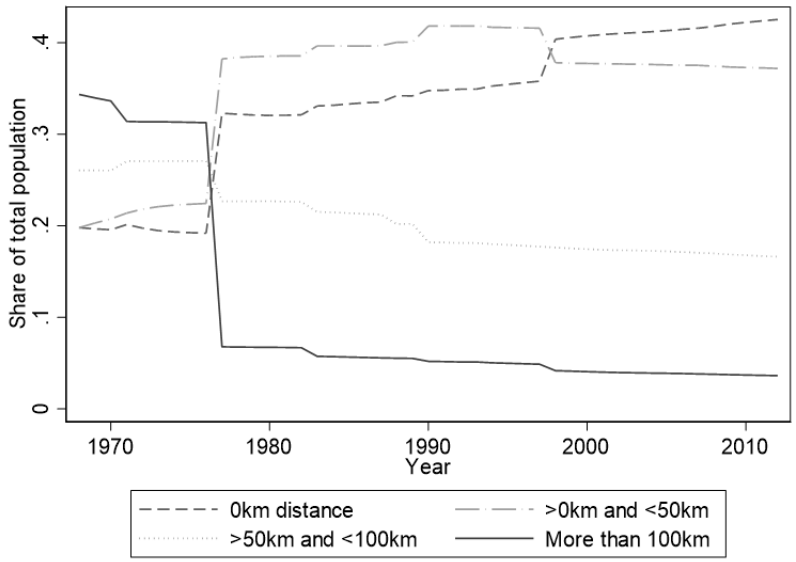
Notes: Population weighted average of the distance to the closest higher education institution. The population weights only use the 18-year-old population.

Figure 2.A.1 displays the aggregated effect of the university expansion reform on access to higher education for the population of 18 years old. While Figure 2.5 depicts the impact by treatment groups, Figure 2.A.1 shows that the reform had a substantial total effect. An average 18-year-old Swede lived more than 80 km away from the closest HEI in 1970 before the first new institutions opened. Already in 1977, the year of the main wave of the ex-

pansion reform, the average distance reduces to under 40 km, a distance that could theoretically be commuted. The distance reduces further afterward, but the main impact happened in 1977.

To learn more about how many residents were affected by the reform, Fig-

Figure 2.A.2: Share of total population by distance to closest university

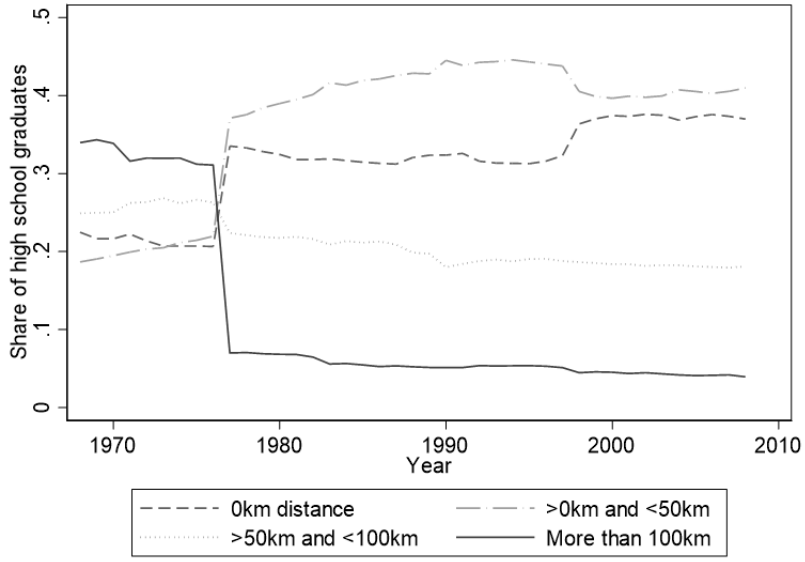


Notes: Share of the total population by distance to the closest higher education institution.

ure 2.A.2 plots the share of the total Swedish population by distance groups. In 1970, only 20% of the total population lived in municipalities with a university or in the 50km catchment area, respectively. The former share increased to more than 30% in 1977, the latter almost doubled to roughly 40%. By definition, the share of the population that lived relatively far away from the closest HEI decreased at the same time. Interestingly, the drop was much larger for the distance group of over 100km, emphasizing the stated goal of the reform to improve access to tertiary education, especially in areas where geographic access is low.

A similar pattern can be observed when only looking at the population of 18-year-olds that are about to finish high school as depicted in Figure 2.A.3. The similarity of Figure 2.A.2 and Figure 2.A.3 also shows that my study population of high school graduates is similarly distributed as the total population in terms of distance to the closest HEI.

Figure 2.A.3: Share of 18y old high school graduates by distance to closest university



Notes: Share of the 18-year-old population that will finish higher secondary education by distance to the closest higher education institution.

2.B Additional Results

Long-Term Mobility

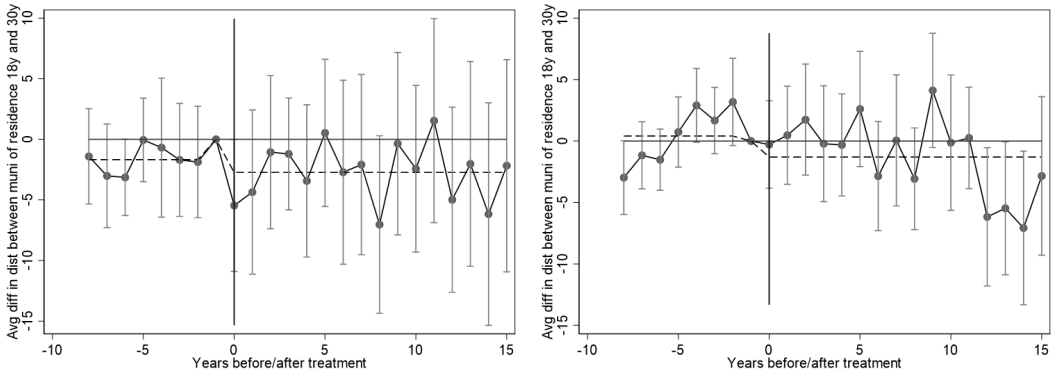
Figure 2.B.4 shows the effect on the distance between the place of residence at age 18 and age 30 years (in km). Although I measure a significant effect on mobility at age 19-22 years, there seems to be no measurable effect on location decisions after finishing a university education.

Mobility of Lower Educated

Figure 2.B.5 shows the treatment effect on the propensity to move at age 19-22 years for young adults with lower secondary education or less. Interestingly, the effects are similar in terms of the sign as for the higher educated of the main sample, although the population of this specification here is not eligible to enroll in a college. However, there are some differences in magnitude. In the "new uni" municipalities, the (negative) average treatment effect is closer to zero and becomes statistically insignificant. In contrast, the effect in

Figure 2.B.4: Distance between place of residence with 18y and 30y

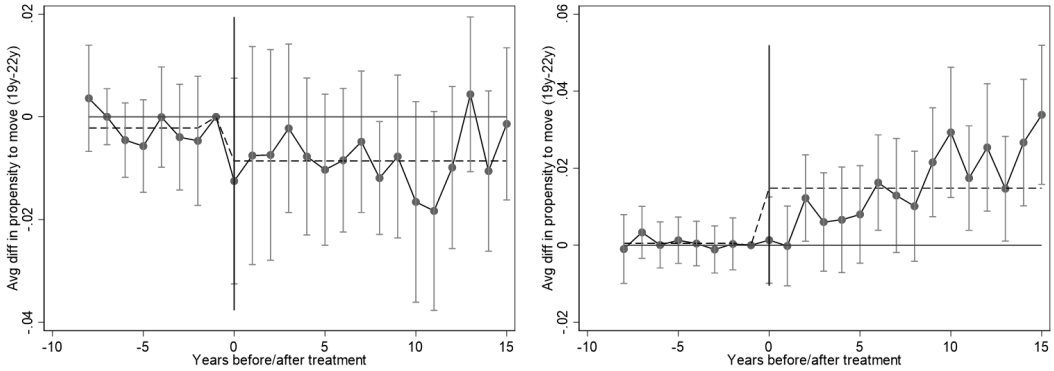
New uni municipalities 50km catchment area of new uni



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 2.B.5: Propensity to move with 19y-22y

New uni municipalities 50km catchment area of new uni

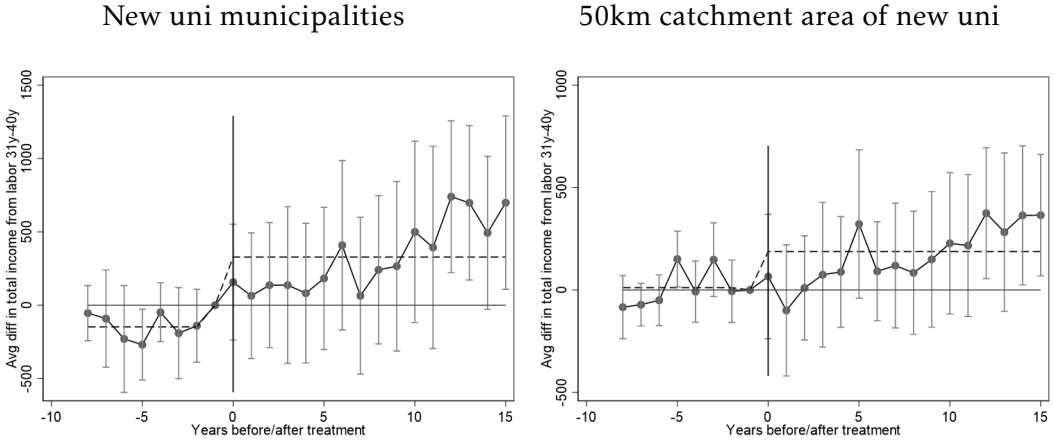


Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with lower secondary education or less. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

the 50km catchment area is more pronounced and statistically significant on a 5% level. These results are in line with the results of the main specification. In the "new uni" municipalities, the negative effect on the mobility of high school graduates is driven by an increased college participation rate. Since young adults with lower levels of education are not eligible to enroll in a university, there is no significant effect on mobility. In the catchment area, however, the slightly positive effect on the mobility of high school graduates could not be explained by the direct effect of high participation rates in tertiary education. For neighboring regions, labor market effects, peer effects, or changes in local amenities are (relatively) more relevant. Since these indirect effects can apply to all residents, not just those who have a high level of education, there is a measurable impact on the migration patterns of young adults with lower education in the catchment area.

Income

Figure 2.B.6: Change in the income of labor



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Controls: parental level of education and own level of education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 2.B.6 shows the effect on average yearly income from labor in SEK (adjusted to CPI of 2011) that individuals earn in their 30s (age 31-

40 years). When controlling for education there is a slightly positive but statistically and economically insignificant positive effect, both in "new uni" municipalities as well as in the 50km catchment area. Given these results, the indirect effect of new HEIs via the labor market seem to be low in the 15 years after the new institution opened.

2.C Robustness Checks

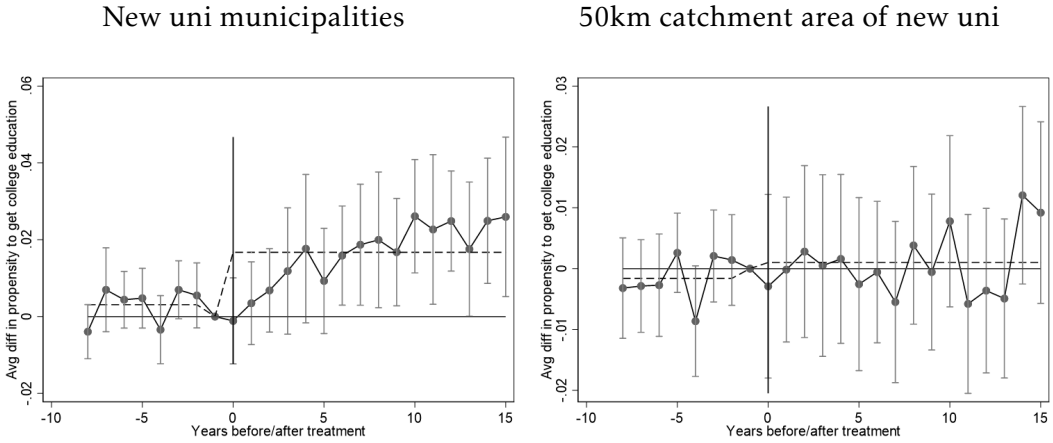
Controlling for Parental Education

One important determinant of participation in tertiary education is the educational background of parents. If parents have a college degree, children are more likely to enroll at a university, all other factors equal. High school graduates from treatment and control groups are systematically different in the level of education of their parents, results presented above would be biased. In this section, I show that results do not differ when controlling for parents' education. It is defined as the level of the father's education (primary and no education, secondary education, or tertiary education and higher) or the mother's education if the father's highest degree is unknown. Since the educational register of Sweden does not cover information on individuals that died before 1990, I use the census of 1970 to add information on the highest degree for older cohorts. Figure 2.C.7 plots the estimates for college education as the dependent variable, while Figure 2.C.8 presents the results for mobility between 19-22 years as an outcome. Comparing the estimates with results from my main specifications above, one can see that controlling for parents' education does not make a difference, presumably because the two-way-fixed-effects framework deals with potential (time-consistent) differences between treatment and control groups already.

Excluding Malmö and Huddinge from Treatment Group

The municipalities of Malmö and Huddinge both received a new university in 1998. However, both municipalities were close to an old university even before that. While Lund is close to Malmö, Huddinge belongs to the greater area of Stockholm, although being its own municipality. Therefore, they

Figure 2.C.7: Propensity to obtain a college degree



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Controls: parental level of education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

could be assigned both to the group of treated as well as to the group of always-treated regions. Here, I choose the latter option, dropping them from my sample entirely. Figure 2.C.9 and Figure 2.C.10 show that results do not change compared to the main specification, where both municipalities belong to the treatment group.

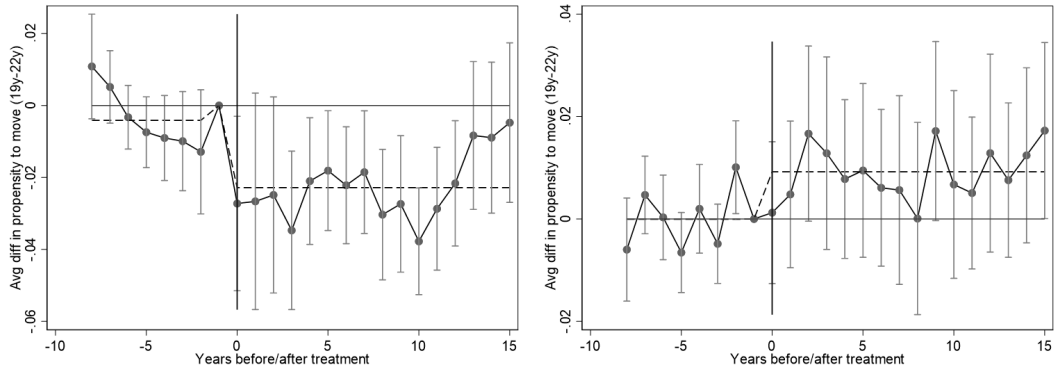
Different Definitions of the Catchment Area

Which municipality belongs to which of the five treatment groups defined in subsection 2.3.3 depends on the threshold of catchment areas. A higher threshold includes more municipalities in the catchment areas of both, "new uni" municipalities (treatment group 4) and always-treated "old uni" municipalities (treatment group 2). The question is: how far does a university's (both old and new) effect on education and migration decisions reach in terms of geographical distance?

To make sure that my results do not depend on the choice of the size of the catchment areas, this section provides results for alternative definitions of 30km (Figure 2.C.11 and Figure 2.C.12) as well as 75km (Figure 2.C.13 and

Figure 2.C.8: Propensity to move with 19y-22y

New uni municipalities 50km catchment area of new uni

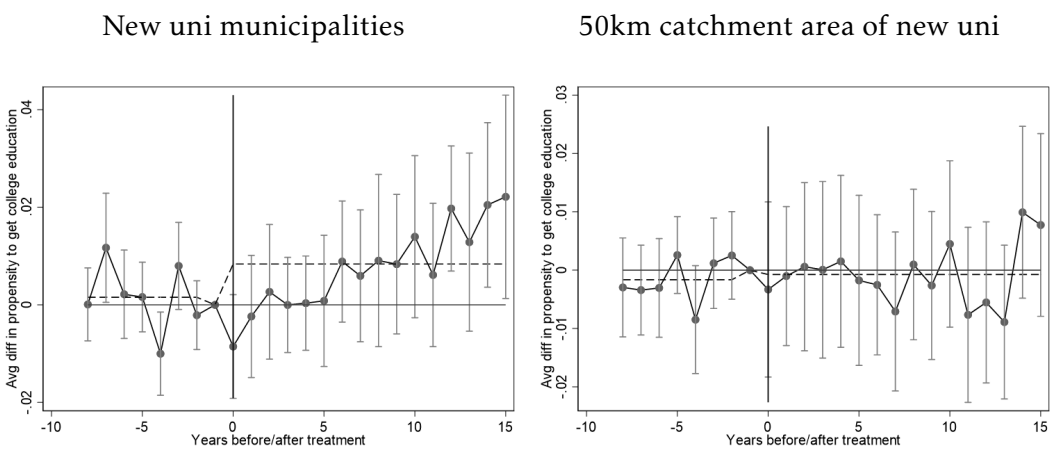


Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Controls: parental level of education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 2.C.14). Comparing the estimates to the results of my main specification with a catchment area threshold of 50km, one can see that results for "new uni" municipalities do not change, even though the control group varies.

Looking at results for different catchment areas, where both treatment and control groups are different, there are little differences. The propensity to obtain a college degree is higher in the 30km specification (Figure 2.C.11), indicating that there are some positive spill-overs in that area. As with 50km, the estimates are also zero with a 75km catchment area (Figure 2.C.13). It seems that positive spatial spill-overs decay with distance and disappear somewhere between 30km and 50km, making municipalities beyond 50km a suitable control group. The propensity to move between 19-22 years shows a decay in the treatment effect, too. The average effect in the main specification with 50km is a little bit lower compared to the 30km specification (Figure 2.C.12), but higher than estimates in the 75km variant (Figure 2.C.14). For 75km, the estimated treatment effect is not significant even at the 10% level. I conclude that the SUTVA, the assumption requiring the control

Figure 2.C.9: Propensity to obtain a college degree

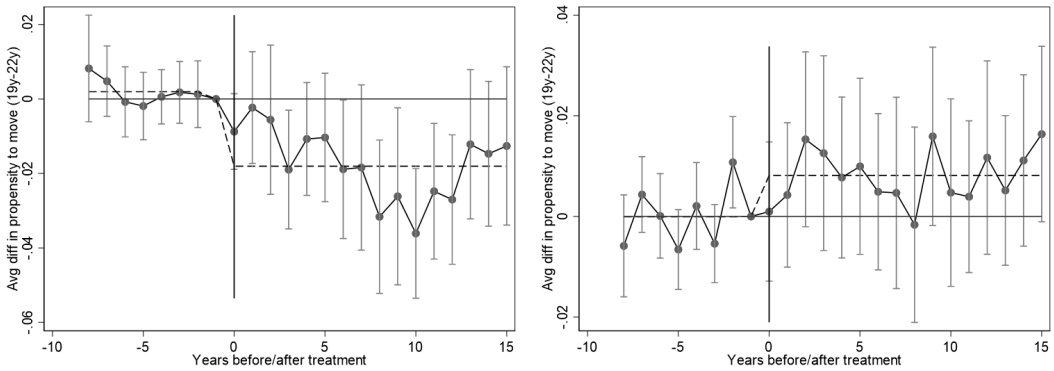


Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Municipalities that were close to an "old uni" municipality and got their own university are dropped as well. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

group to be unaffected by the treatment, is satisfied. However, even if we conclude that spatial spill-overs are not entirely zero for municipalities that are more than 50km away from the university, I argue that a 50km catchment area definition is sufficient. Using the 75km specification makes the control groups relatively small which results in the loss of precision and statistical power. In addition, estimates are, if biased at all, biased towards zero when a positive treatment effect is incorporated in the control group.

Figure 2.C.10: Propensity to move with 19y-22y

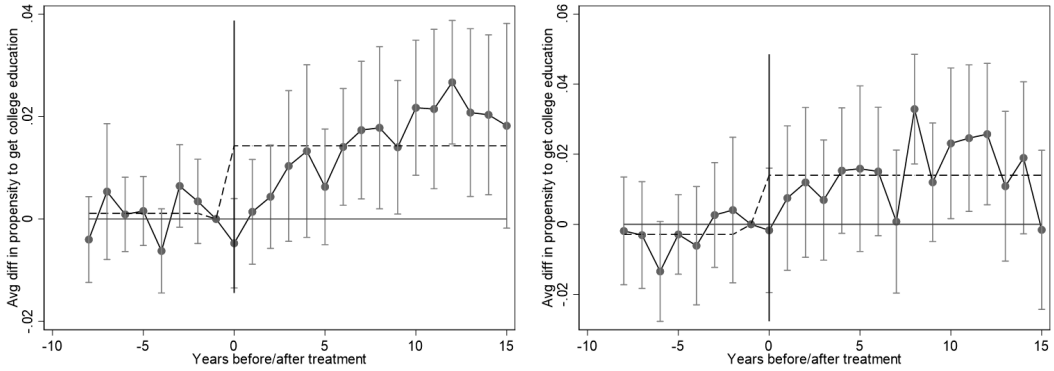
New uni municipalities 50km catchment area of new uni



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 50km to below 50km (right). Always treated municipalities excluded. Municipalities that were close to an "old uni" municipality and got their own university are dropped as well. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 2.C.11: Propensity to obtain a college degree; 30km

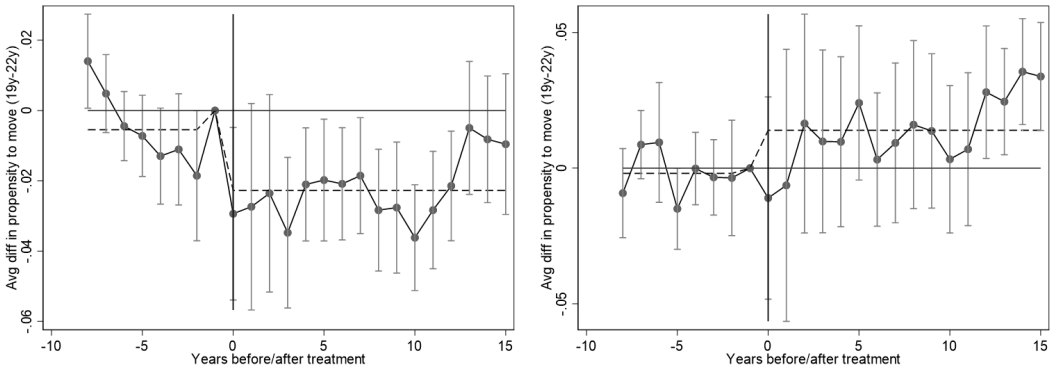
New uni municipalities 30km catchment area of new uni



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 30km to below 30km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 2.C.12: Propensity to move with 19y-22y; 30km

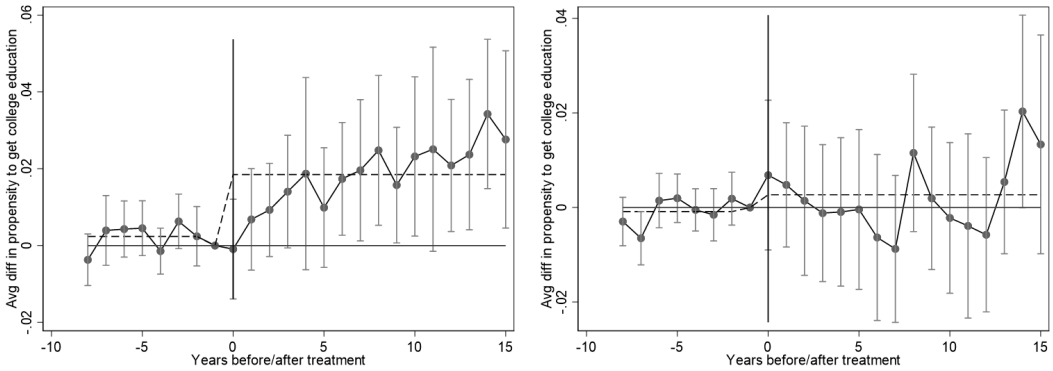
New uni municipalities 30km catchment area of new uni



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 30km to below 30km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 2.C.13: Propensity to obtain a college degree; 75km

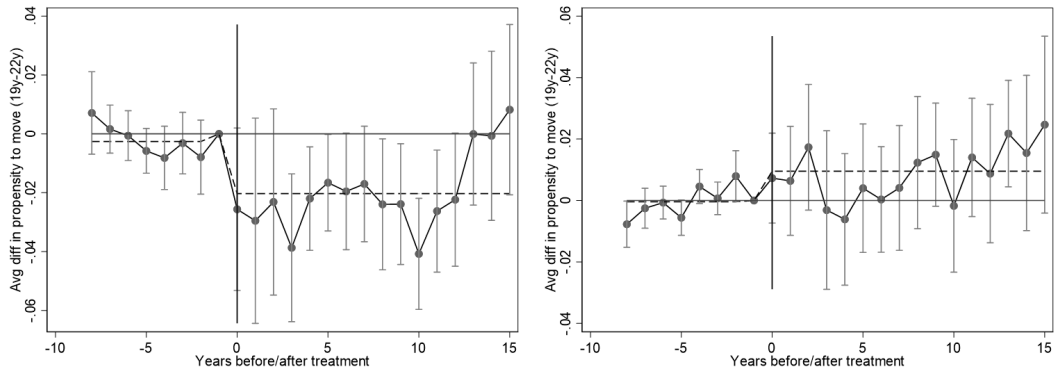
New uni municipalities 75km catchment area of new uni



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 75km to below 75km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 2.C.14: Propensity to move with 19y-22y; 75km

New uni municipalities 75km catchment area of new uni



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution (left) or drop in distance to closest higher education institution from above 75km to below 75km (right). Always treated municipalities excluded. Only including individuals with at least higher secondary education. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

CHAPTER 3

Assortative Mating and the Access to Higher Education

Parts of this chapter are published in Markus, Philipp (2023): "Effects of Access to Universities on Education and Migration Decisions", Ruhr Economic Papers No. 996.

Chapter Abstract

College education expansions are associated with increased marital sorting, especially at the top of the educational distribution. In this paper, I document that educational assortative mating is stronger for highly educated in Sweden. However, I do not find any causal impact of a geographic tertiary education expansion reform beginning in the 1970s on educational homogamy. Moreover, I estimate that opening a new higher education institution increases the likelihood of local high school graduates marrying someone born in the same municipality by 11.62%. The results suggest that the high degree of assortative mating among college-educated is partly driven by higher mobility rather than the level of qualification.

3.1 Introduction

The research on marriage and family formation became a permanent topic in economics with Gary S. Becker's seminal but controversial introduction of the notion of the marriage market (Becker, 1973, 1974).[1] A significant part of the subsequent literature has focused on answering the question of who married whom (see Chiappori (2020) for an overview). Educational assortative mating (AM), the tendency to marry someone with a similar level of education, has attracted growing attention for being a potential driver of inequality both within and between generations (Burtless, 1999; Fernández and Rogerson, 2001; Greenwood et al., 2014; Frémeaux and Lefranc, 2020; Nybom et al., 2022; Holmlund, 2022). Several studies have documented that assortativity is especially strong at the top of the educational distribution (Blossfeld and Timm, 2003; Eika et al., 2019; Chiappori et al., 2020a; Gihleb and Lang, 2020), while papers on potential explanations for this pattern are relatively rare.

In this paper, I estimate the effect of changes in migration and education behavior induced by a Swedish tertiary education expansion reform beginning in the 1970s on marital patterns of highly educated young adults. More specifically, I answer the question of whether mobility is a determinant of the high degree of martial sorting among college graduates.

My first contribution to the above-mentioned literature on educational AM is to confirm that marital sorting is stronger for higher levels of education in Sweden, as documented for other countries already (Blossfeld and Timm, 2003; Eika et al., 2019; Chiappori et al., 2020a; Gihleb and Lang, 2020). I do so by extending the perfect-random normalization measure based on Liu and Lu (2006) in such a way that I am able to calculate the degree of AM for each birth cohort separately in a way that is meaningful in the presence of a changing education distribution. This approach offers future researchers a way to derive the degree of AM by the cohort of birth in a more generalized way, which is necessary when investigating the treatment effect of an intervention that affects only certain cohorts.

When it comes to causal mechanisms rationalizing the relationship between education and the degree of AM, the current state of research offers

[1] See Pollak (2003) for an overview over the critique of Becker's approach.

only few explanations. Related papers have revealed that martial sorting among college-educated is even stronger if the degree is in the same field of study (Eika et al., 2019; Artmann et al., 2021; Pestel, 2021; Bicakova and Jurajda, 2016; Han and Qian, 2021) or from the same institution (Nielsen and Svarer, 2009). This is in line with two mechanisms explaining strong educational assortativity among college-educated (see Schwartz (2013) and Luo (2017) for overviews). First, graduates of the same field could be a better match because they are more likely to share common interests or characteristics that are correlated with choosing to study in a specific field or at a given institution. Second, having lectures together or ending up in the same company after graduating makes it more likely to meet. Exploiting discontinuities in the college admission process created by unpredictable cutoffs, Kirkebøen et al. (2021) compare the importance of a common institution and a similar field of study for marriage market outcomes of college students in Norway. They show that studying at the same institution significantly increases the chance of getting married, especially when being enrolled at the same time. Furthermore, enrolling in a specific field makes students more likely to marry someone within the same field only at the same institution, while there is no increased propensity to couple with someone of the same field attending a different institution. Hence, the authors conclude that the opportunity to meet is more relevant than sharing a common interest, providing causal evidence that the college is a relevant local marriage market for highly educated. This is in line with previous descriptive evidence: In a survey on couples in France in 1984, a majority of the highest skill groups state that they met their partner while studying (Bozon and Heran, 1989). Building on the idea of colleges as marriage markets, Blossfeld and Timm (2003) and Blossfeld (2009) argue that the selectivity of the education system is key in explaining higher levels of homogamy at the top of the educational distribution. Selection barriers make social networks more homogeneous in the later stages of the educational career, increasing the propensity to meet people with similar characteristics and levels of qualification.[2] According

[2]Blossfeld (2009) also discusses two additional channels of how the selectivity could impact differences in mating patterns that both focus on the timing of marriage. First, marriages are usually formed after the completion of education, which happens later for highly educated. Hence, they have a higher chance to be in a relationship with someone they met during the period of education when typically considering marriage. Second, vice

to that, college is a local marriage market for highly skilled, while lower levels of education do not have an equivalent institution that provides the opportunity to exclusively meet potential partners with the same degree.

I argue that university students differ in the degree they are exposed to these selective marriage markets. Psychological studies have documented that social network formation is especially important after changing the place of residence (Oishi, 2010; Magdol and Bessel, 2003; Seder and Oishi, 2008), suggesting that students who move between stages of the education system participate more in the social network at college.[3] Hence, the local marriage market has a larger impact on them. On an individual level, students that do not move to a different location in order to attend college should be less likely to sort into a homogamous union compared to more mobile students. On the aggregated level this relation could explain the correlation of assortativity and education since college-educated are known to be more mobile (see for example Corcoran and Faggian, 2017).

Identifying the causal effect of migration on marital patterns is challenging due to potential endogeneity problems. There might be selection into migration, education and certain marriages based on unobserved characteristics. For example, more curious and open-minded individuals might be more likely to leave their parent's home early, to achieve a higher level of qualification and to marry someone who is also open-minded. Additionally, migration decisions might depend on the civil status or preferences of partners, e.g. when moving is more costly in a relationship, raising the concern of reversed causality. To cope with these issues, I use a quasi-exogeneous expansion of the tertiary education landscape in Sweden to estimate the causal chain between education, migration, and marital sorting. The opening of 14 new universities and university colleges between 1968 and 2012 generated massive changes in access to higher education in Sweden. Markus (2023) estimated that young adults graduating from high schools in municipalities with a new university are, on average, 6.6% more likely to attend college

versa, lower educated leave the education system earlier and diversify their social network before becoming ready for marriage, leading to more heterogeneous unions. However, the treatment investigated in this paper seems to have no effect on the timing of marriage. Hence, I am not able to test these hypotheses.

[3] See Marmaros and Sacerdote (2006) for an overview of the literature on social network formation on campus.

and 10.1% less likely to move away in the years after finishing secondary education. Theoretically, the education effect should lead to higher rates of homogamy due to the reasons already discussed above. In contrast, the sign of the effect of reduced mobility of college students on the propensity to marry someone alike is ambiguous. On the one hand, the opportunity to couple with a college student might change as studying in a different municipality compared to the counterfactual situation of moving away exposes students to a different local marriage market. On the other hand, immobile high school graduates might be less integrated into social networks on the campus in absence of migration, attenuating the positive opportunity effect of education.

By using a two-way fixed effects approach I control for time-constant differences between municipalities as well as the national time trend. Since the new institutions were founded in different years I use a dynamic Difference-in-Difference (DiD) estimation method, usually referred to as event study (see Roth et al. (2022) and Chaisemartin and D 'Haultfoeuille (2022) for a review of the latest development). It compares individuals from "treated regions" (i.e. municipalities where a university or university college was newly opened within a certain radius) with individuals from "control regions" (i.e. municipalities that never had a higher education institution close by), relative to the difference between those two groups that existed already before the new institution opened.

My results indicate that the higher education expansion reform had no effect on the likelihood of marrying someone with a college degree conditional on having college education and getting married, i.e. the rate of homogamy among highly educated individuals does not increase. I provide evidence that the reduced mobility counteracts potentially positive effects of education by increasing the propensity to marry someone born in the same municipality by 11.62%. My results suggest that mobility and geographical sorting play an important role in explaining heterogeneity in educational AM, complementing the educational selectivity hypothesis of Blossfeld and Timm (2003) and Blossfeld (2009) as well as the idea of colleges as local marriage markets (Kirkebøen et al., 2021; Pestel, 2021; Nielsen and Svarer, 2009). It implies that college students are more prone to marital sorting not only because they have access to the college as an exclusive marriage market

but also because their social network becomes more homogeneous due to geographic mobility.

In a broader sense, my paper adds to the scarce literature on spatial homogamy (see Haandrikman et al., 2008; Haandrikman, 2019; Nielsen and Svarer, 2009). Haandrikman (2019), for example, finds that even today half of all partners in Sweden did not live more than 9 kilometers away from each other before moving in together, despite the developments in educational participation, mobility, and technology. My results emphasize the importance of migration in early adulthood for the geography of marriage.

Another related body of research analyzes the effect of higher education expansion reforms as place-based policies.[4] Some studies document effects on economic outcomes like regional income and supply of high-skilled labor (Carneiro et al., 2023; Berlingieri et al., 2022; Liu, 2015) or local innovative activities (Andersson et al., 2009; Lehnert et al., 2020). Likewise, other studies find evidence of positive effects on the education of the local youth (Frenette, 2009; Gibbons and Vignoles, 2012; Markus, 2023) while Suhonen and Karhunen (2019) show that a Finnish higher education expansion reform increases spillover effects from parental to children's education. Kamhöfer and Westphal (2019) provide causal evidence that the university expansion reform in Germany decreased the fertility of college-educated women. Most similar to my study, Nybom et al. (2022) use the same Swedish tertiary education expansion reform as an instrument that exogenously increases tertiary education to provide causal evidence for growing inter-generational inequality through educational AM. My results put doubt on the validity of using the geographical expansion reform as an instrument since the exclusion restriction is violated by the mobility channel.

This paper is structured as follows. The next section summarizes the institutional background of the Swedish tertiary education landscape and the expansion reform beginning in the 70s. Section 3.3 describes the data before I document heterogeneity in the educational assortativity in section 3.4. After discussing the empirical method, I present my main empirical results on the causal effects of mobility in section 3.6. The final section concludes.

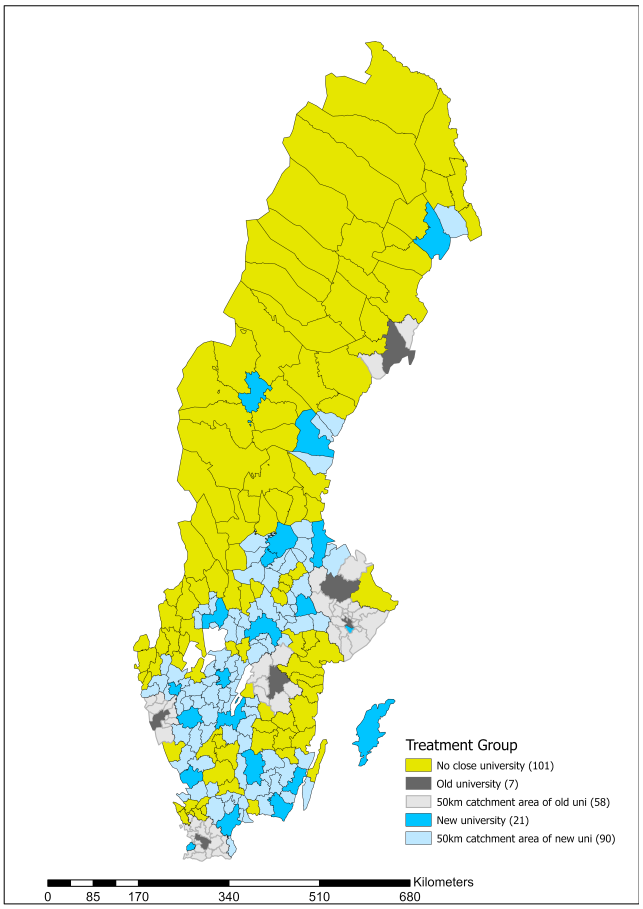
[4]See Kyvik (2009) for an overview of higher education expansion reforms.

3.2 Institutional Background

3.2.1 The University Expansion Reform

In 1970, 19 higher education institutions (HEIs) operated in seven different municipalities in Sweden. The locations are depicted in dark gray in Figure 3.1. Only six of the 19 institutions were universities providing general

Figure 3.1: Access to tertiary education in Sweden



Notes: Sweden's municipalities with boundaries of 1977 grouped by treatment status defined in section 3.3.3. The number of municipalities of each group is in parentheses.

tertiary education at that time.[5] The other HEIs were more specialized and affiliated with a university. Therefore, they were located in the same munic-

[5]Uppsala, Lund, Göteborg, Stockholm, and Umeå universities. Linköping university already offered a wide range of programs but got the official status of a university in 1975.

ipalities as the general universities. The only exception is the Karolinska Institute in Solna, which is a part of the greater area of Stockholm. In Sweden, tertiary education is provided by universities (*Universitet*) and university colleges (*Högskolan*). In contrast to universities, university colleges do not provide doctoral education. As this difference is not relevant to my paper, the terms university, college, and HEI are used interchangeably. I also do not distinguish between specialized universities or those providing education in all academic fields throughout this paper, unless stated otherwise.

Due to the size of the country, having only seven locations offering access to tertiary education means that a substantial part of the population lives relatively far away from a university. In 1970, the (population-weighted) average distance to the next college was over 80 km.[6] Since schooling including post-secondary education is traditionally free of tuition fees in Sweden (Deen, 2007), the (lack of) geographical access was seen as a major college education friction. Politicians feared that such distances imposed a prohibitive high economic, social or psychological cost to attend college for some (Premfors, 1984). Therefore and due to the generally increasing number of students, the government decided to establish a significant number of new HEIs in the 1970s.[7] The Luleå University of Technology was already founded in 1971. But the substantial change in the higher education landscape in Sweden happened in 1977 when 14 new HEIs were established in 14 different locations where no university was operating before.[8] From 1977 on, there was a total of 22 municipalities with at least one HEI offering access to tertiary education. As intended by the government, this massive expansion more than halved the average distance to the closest college to below 40 km (see Figure 3.A.1 in the Appendix). Additionally, roughly 40% lived under 50 km away from the next university in 1970 before that share jumped up to almost 70% in 1977.[9]

After 1977, six more universities were established, again in municipali-

[6]The average distance to the closest HEI over time is also visualized in Figure 3.A.1 in the Appendix.

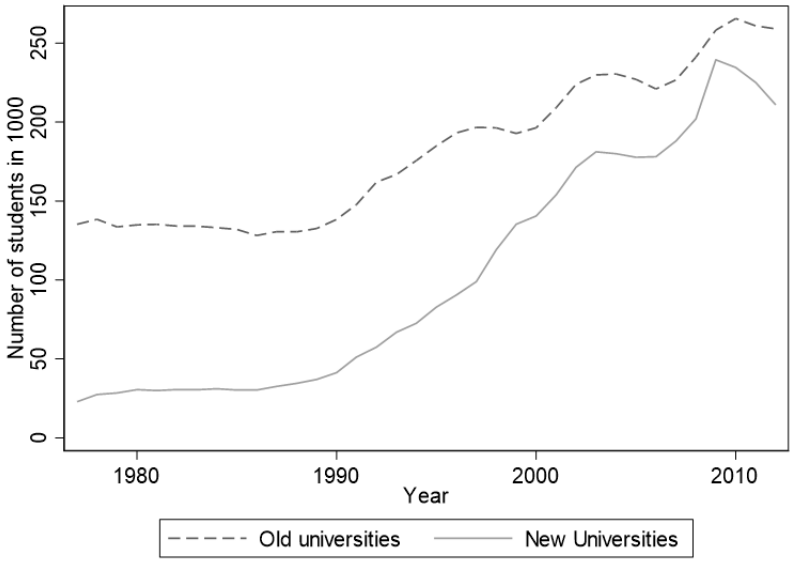
[7]See e.g. Varga (1998) and Anselin et al. (1997) for a review on the growing number of students.

[8]The 14 municipalities receiving a HEI in 1977 were Jönköping; Vaxjö; Kalmar; Kristianstad; Borås; Skövde; Karlstad; Örebro; Västerås; Falum; Borlänge; Gävle; Sundvall and Östersund.

[9]The full dynamic is visualized in Figure 3.A.2 in the Appendix.

ties without any HEI until then.[10] Figure 3.1 shows the "new" structure of tertiary education in Sweden where municipalities with new colleges are represented in dark blue. Notation-wise, I will refer to universities that existed already before 1970 as "old" universities and the according municipalities as "old uni" municipalities. Equivalently, colleges established after 1970 are termed "new" universities and their municipalities "new uni" municipalities. The "new" colleges started with relatively low numbers of enrolled students as shown in Figure 3.2 and did not start to catch up until the 1990s.

Figure 3.2: Total number of students enrolled in universities



Notes: "Old" universities were founded before 1968 and "new" universities after 1968.

As mentioned above, one of the political goals of the Swedish college expansion reform was to make post-secondary education more accessible, geographically as well as financially (Andersson et al., 2004). The motivation was to reduce education frictions by lowering the costs of migration or commuting, something that might affect those from lower social classes stronger. The unofficial slogan of the enabling legislation for the initial expansion in 1977 was "En hogskoleenhet i varje ort", which roughly translates to "a unit

[10]The universities opened after 1977 are Halmstad University in 1983, Blekinge Institute of Technology in Karlskrona in 1988, University of Trollhättan/Uddevalla in Trollhättan in 1990, Södertörn University in Huddinge, again a part of the larger Stockholm area, and Malmö University as well as Gotland University in 1998.

of higher education in every locality" (Swedish Government, 1977). It can not be ruled out entirely that other factors like regional economic or demographic characteristics have played a role in the location choices of the new institutions. However, interviews with responsible policymakers of that time (Andersson et al., 2009), as well as the reports of the responsible commissions (Premfors, 1984), confirm the hypothesis that geographic dispersion of access to higher education was the primary objective when choosing the locations. I also provide descriptive evidence that socio-demographic characteristics of the municipalities did not play a role in choosing the location for the new college after introducing the data in section 3.3.

3.2.2 Schooling

To investigate the effect of changes in access to higher education, I need to observe individuals before they decide to go to college. In Sweden, children start school in the year they turn seven. Since a reform in 1950, nine years of schooling are compulsory (Holmlund, 2021).[11] Hence, students are not allowed to leave school before the year they turn 16. After completing the 9th grade, students have to choose to follow either a vocational track, a general track, or a theoretical track if they decide to continue schooling. Each track consists of several national programs where programs of the theoretical track aim to prepare students for university. Until 1993, the vocationally orientated programs (lower secondary education) had a duration of two years while the theoretical track (higher secondary education) took three years, making students eligible to enroll at a university. However, the duration of the vocational programs was extended to three years as well in 1993 such that they also qualify students for accessing higher education (Deen, 2007). Summing up, it has always been necessary to finish 12 years of schooling before entering a university, which is usually the case in the year students turn 19.

[11] See Fischer et al. (2020) for a more detailed assessment of the Swedish education system before 1950.

3.3 Data and Descriptives

3.3.1 University Data

To analyze the effect of university openings, the time and location of these openings are crucial. I mainly follow a report of the Swedish National Agency for Higher Education (*Högskoleverket*) on the Swedish higher education landscape from 2006 (Högskoleverket (National Agency for Higher Education), 2006). However, the official founding year is not always the year in which a higher education institution (HEI) becomes a notable provider of tertiary education. For that reason, information from the Higher Education Register (*Högskoleregistret*) on the number of enrolled students and hand-collected data was added to determine the de-facto start of a new college. Throughout this paper, a HEI is considered as such if it is labeled as a university or university college in the report of the National Agency for Higher Education, offers (at least) undergraduate education in more than one field, or has more than 500 students enrolled. That excludes specialized HEIs like nursing, military, or theater schools.
The location of the main campus is linked to the respective municipality, where I use centroids to determine the distance to other municipalities and their residents. All municipalities are defined in the administrative borders of 1977 to keep the borders constant over time. That leaves 277 municipalities of today's 290. In case a university has numerous branches in different locations or more than one campus, only the location of the main campus is considered. The only exception is secondary campuses which were independent universities and continued to host a sizable number of students after a merger. Given that rule, there are no university closures between 1968 and 2012.

3.3.2 Individual-Level Data

To investigate the effect on the martial decisions of individuals, I combine several Swedish administrative records. All individual-level data in this paper comes from the Swedish Interdisciplinary Panel (SIP), administered by the Centre for Economic Demography, Lund University, Sweden.

The birth register contains information on the year of birth, gender, and place of birth of the full population.[12] The register of the total population, available from 1968 onward, includes yearly information on the municipality of residence and links to spouses and parents. The place of residence is a central variable in my analysis. It is defined as the municipality in which the individual was registered by the end of the year and allows me to identify migration patterns and local marriage markets since I know who lived together in a region each year. The municipality of residence is also used to calculate the distance to the closest HEI and, therefore, is the determinant of each individual's treatment status (see section 3.3.3). One notable concern is that people might not live at the place where they are registered. For instance, young adults might move out of their parent's homes without registering the new address. That would result in an underestimation of mobility. However, it should be noted that the law requires citizens to register their place of living and non-compliance is penalized.

The linked spouse is crucial information for assessing assortative mating since it allows for linking characteristics of the spouse to every person in a partnership. This includes marriages as well as registered partnerships that allow same-sex couples to officially register a partnership since 1995.[13] In addition, two adults registered at the same address and having a child together are also linked in the data, even if they are not married or in a registered partnership. Hence, all partnerships that are officially registered (as registered partnerships or marriages) and cohabiting couples with a common child are identified. Since the differences between these forms of partnerships are beyond the scope of the paper, terms like partner and spouse are used interchangeably. Non-married couples that either do not live together or do not have a child together cannot be taken into account. If a person is married but the spouse is not part of the register of the total

[12] Although the birth register contains information on the district of birth, I only use the municipality of birth. As noted in section 3.3.1, I consider the administrative borders of 1977. For municipalities, this involves only a few approximations for reforms between 1952 and 1967. In contrast, there are many splits and mergers of districts before 1977 which makes assigning them to the state of 1977 inaccurate. Still, when using the information on the municipality of birth, I have to exclude cohorts born before 1952.

[13] In- or excluding same-sex partnerships does not change the results, as less than 0.6% of the sample population is linked to a spouse of the same sex (see Figure 3.B.5 in the Appendix).

population (e.g. because the spouse never lived in Sweden), the civil status is still known, even without a linked partner.

Given that the treatment status is assigned according to the place of residence at age 18, my base study population consists of the full Swedish population born between 1950 and 1972 who finished at least 12 years of schooling, i.e. are eligible to attend college, and married at least once by the age of 40. The cohort born in 1950 turned 18 in 1968, which is the first year where I can observe (among others) the municipality of residence. I cannot use older cohorts, since the first time I observe their place of living is after they (potentially) moved out of their place of schooling. Stopping at 1972 makes the panel almost perfectly balanced, as I can observe every included cohort at least until the age of 40.[14] As in Kirkebøen et al. (2021), selection in or out of the sample based on the treatment effect on the decision whether to marry at all is not an issue as the university reform does not affect the chance to get married (see Figure 3.C.12 in the Appendix).

I construct a proxy for migration costs with the provided link to parents and grandparents. Previous research has shown that individuals who live in the same location as their parents and grandparents have a lower probability of moving away, all other factors constant, since the social costs of moving are higher (see for example Mulder and Malmberg (2014) for Sweden). Controlling for the presence of local family ties ensures that my estimated effect is not driven by different structures and past migration patterns of the family.

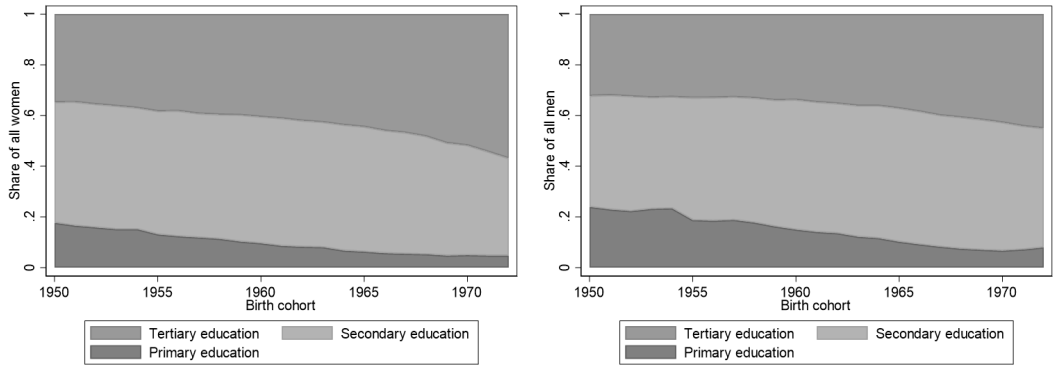
The educational register provides information on the highest achieved level of education. It allows me to distinguish between only primary education without a high school degree (less than 11 years of schooling), secondary education with a high school degree (11-12 years of schooling), and (some)

[14]As depicted in Figure 3.B.4 in the Appendix, the share of each cohort that gets married before turning 41 is declining over time. At the same time, the average age at the time of the first marriage increases for younger cohorts. That raises the question of whether a specific group drops out of the sample of married individuals due to delaying the first marriage to an age larger than 40. That might be especially problematic for the highest education class, as college-educated get married at a later stage of their life. I argue that this is unlikely to be a significant issue for two reasons. First, even for the youngest cohort considered in the paper, the average age when getting married is 31 for tertiary educated, which is still way below 40. Second, the likelihood to be married by the age of 40 does not decline faster for highly educated individuals, indicating that there is no disproportional selection out of the sample.

tertiary education, i.e. college education (more than 12 years of schooling). As described in section 3.2.2, not every high school degree makes students eligible to enroll in a college. Hence, the group of secondary education is separated into "eligible for college" (12 years of schooling) and "not eligible for college" (11 years of schooling) when estimating the treatment effect of the tertiary education reform, where only those eligible for college are considered relevant. Meanwhile, all three groups are used to investigate assortative matching for each education class. The education register was recorded in 1990 for the first time and includes all degrees obtained until 2019. Therefore, individuals who died before 1990 are not covered. This is primarily a problem for the parents of the study population. To complement information on parents' education as well as economical background the Swedish Census of 1970 was added whenever information from the educational register is missing. In contrast to the education register, information in the Census is based on self-enumeration and refers to October 1970 instead of the status in 1990, which might explain little differences. It should be emphasized that the register data provides some information on the time when the highest degree was obtained, but only for a relatively small sub-sample of the total population. Hence, all information on education received from the educational register is time-invariant and represents only the highest educational degree. The Census of 1970 does not necessarily contain the highest educational degree, but the highest degree obtained by October 1970. Nevertheless, information from the census is mainly used to complement information for individuals who died before 1990 and therefore can be expected to have completed their educational careers at the time of the census. Figure 3.3 plots the change in the educational distribution for cohorts born between 1950 and 1972. For both genders, the average level of education is increasing over time. While around 20% of the cohort born in 1950 does not have a high school degree, this holds for only 10% of the 1972 cohort. The share of secondary educated decreases slightly for women, while it remains fairly constant at roughly 40% for men. The expansion of tertiary education, a well-documented trend in many developed countries, is represented by an increase in college education for both genders. Almost every second woman born in 1972 attended college, while this was the case for less than 30% of women born in 1950. For the same cohorts, the share of college-educated

Figure 3.3: Changes in educational attainment

(a) Women (b) Men



Notes: Education shares of individuals born between 1950 and 1972. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education. The left panel (a) displays the share of each level of education for women. The right panel refers to the educational distribution of men.

men increased less, from about 30% to more than 40%.

Combining the spousal link with data on education, place of birth and other characteristics allows me to investigate various rates of homogamy. As standard in the literature, I only consider the first spouse. My main outcome variable is a dummy variable equalling one if the person is in a partnership with a person with the same level of education. To investigate the effects on the geographical sorting of spouses, I additionally construct an indicator that equals one if both spouses were born in the same municipality.

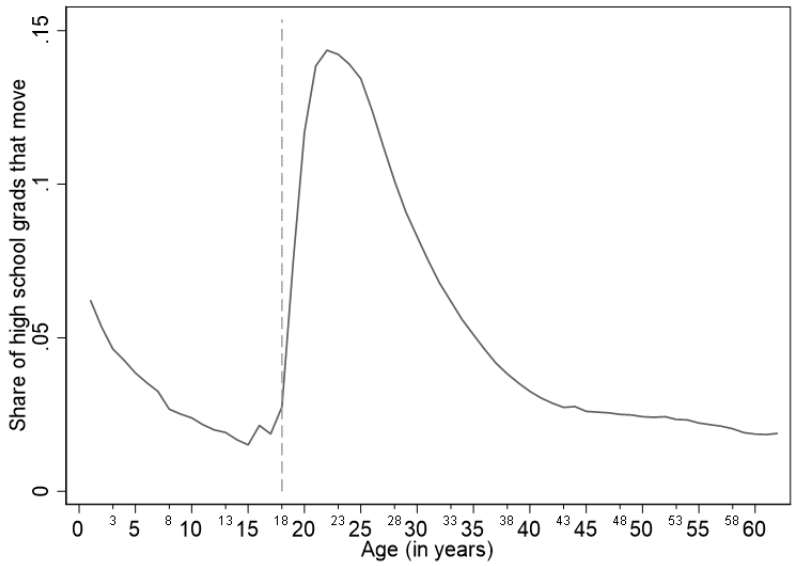
3.3.3 Assigning Treatment Status

An individual is part of the treatment group if she lived in a treated municipality **before** finishing high school. As described in section 3.2.2, students usually finish higher secondary education in the year they turn 19. For that reason, individuals are assigned to the treatment or control group according to the treatment status of their place of residence in the year they turn 18 in the main specification. Figure 3.4 visualizes the likelihood to move by age within my sample. Although mobility starts to increase with the age of 15 years already, the big jump is exactly after turning 18 years, supporting my

approach of assigning individuals to municipalities in the year of turning 18 years.

The treatment status of a municipality depends on the distance to the closest

Figure 3.4: Propensity to move by age



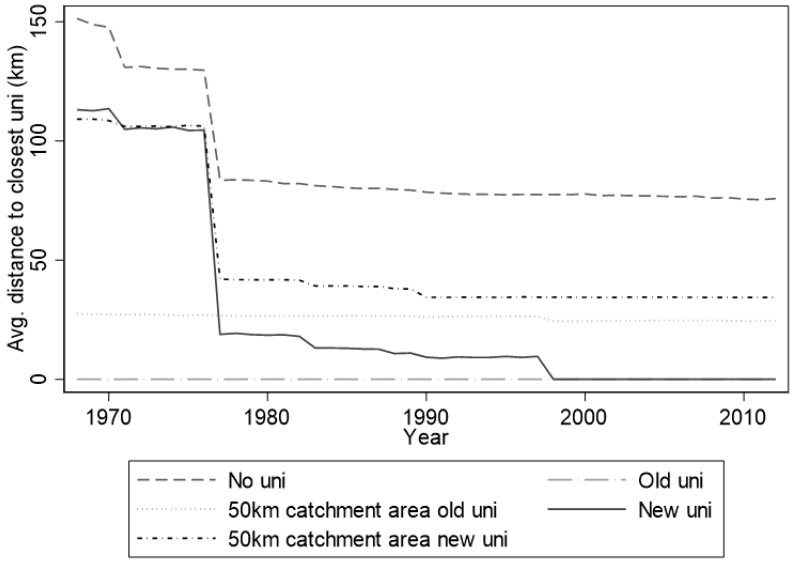
Notes: Propensity to move between municipalities by age. All individuals born between 1950 and 1990 with (at least) secondary education are included.

HEI. In general, there are five different groups of municipalities: (1) "Old uni" municipalities already have a HEI at the beginning of the observation period in 1968 and therefore always have a distance of 0 km to the closest HEI. (2) The catchment area of these "old uni" municipalities includes municipalities that have a distance below a certain threshold throughout the whole observation period. These two groups are usually referred to as always-treated municipalities. (3) "New uni" municipalities are municipalities without a university in 1968 but where a new HEI has opened afterward (i.e. the distance to the closest HEI dropped from above 0 km to 0 km).[15] (4) Equivalently to the "old uni" municipalities, the "new uni" municipalities

[15]Depending on the size of the catchment area, there are two municipalities (Malmö and Huddinge) that belong, by definition, to group 2 *and* 3. They are classified as part of group 3 in the main specification. Treating them as always-treated (group 2) and dropping them from the sample does not change the results, as shown in section 3.D.

have their catchment area as well. Regions from group 3 make up the main treatment group while municipalities from group 4 are used to identify potential spatial spillover effects. (5) Finally, all municipalities that always have a distance above the catchment area threshold are considered never-treated municipalities. Figure 3.1 provides a geographical overview of the five groups for a catchment area threshold of 50 km. There might be changes in the distance to the closest university for some municipalities of the never-treated group as can be seen in Figure 3.5. However, these changes are

Figure 3.5: Distance to the closest higher education institution



Notes: Population weighted average of the distance to the closest higher education institution, by treatment group. The population weights only use the 18 years old population.

considered non-relevant when it comes to education, migration, and mating decisions, either because the changes are very small in magnitude or because the distance is very sizable even after the drop. Individuals graduating from high school in these regions are making up the control group, together with not-yet-treated observations from treatment groups 3 or 4, depending on the specification. Section 3.D in the Appendix shows that the main results do not depend on the definition of the catchment area threshold. However, using graduates from treatment group 5 as well as from not-yet-treated municipalities of treatment groups 3 or 4 requires that these groups are similar

Table 3.1: Population characteristics by treatment group in 1970

	Old (1)	Old Catchment (2)	New (3)	New Catchment (4)	Never uni (5)
Distance to uni (km)	0	27	112	109	146
	(0)	(13)	(71)	(43)	(93)
Pop density	2,142	313	327	39	42
	(1,828)	(530)	(601)	(27)	(74)
Remoteness	310.76	313.91	354.31	307.81	425.56
	(84.7)	(65.97)	(121.55)	(80.71)	(204.79)
Age	31.42	28.6	30.5	31.4	31.93
	(1.47)	(2.8)	(1.11)	(1.12)	(1.57)
Pop share 18y	0.016	0.015	0.017	0.017	0.017
	(0.0008)	(0.0021)	(0.0011)	(0.0013)	(0.0017)
Pop share 19y-29y	0.24	0.22	0.21	0.19	0.18
	(0.02)	(0.03)	(0.01)	(0.01)	(0.02)
Mobility (past 2y)	0.042	0.073	0.041	0.038	0.035
	(0.199)	(0.259)	(0.2)	(0.192)	(0.183)
Primary educ	0.31	0.34	0.37	0.43	0.42
	(0.03)	(0.08)	(0.03)	(0.04)	(0.04)
Secondary educ	0.39	0.41	0.4	0.38	0.39
	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)
Tertiary educ	0.29	0.25	0.24	0.19	0.2
	(0.04)	(0.07)	(0.02)	(0.03)	(0.03)
Married 40y	0.69	0.78	0.75	0.76	0.74
	(0.05)	(0.04)	(0.03)	(0.02)	(0.05)
Individuals	1,318,525	1,129,965	1,409,444	1,273,169	1,605,757
Municipalities	7	58	21	90	101

Notes: Old: Municipalities with universities established before 1968; Old Catchment: Municipalities without university but where a university was established before 1968 within a 50 km radius; New: Municipalities with universities established between 1968-2012; New Catchment: Municipalities without university but where a university was established between 1968-2012 within a 50 km radius; Never uni: No university until 2012. Remoteness is the population-weighted sum of all distances to all other municipalities. High remoteness means that a region is relatively far away from the rest of Sweden's population. Mobility measures the share of people that moved at least once between municipalities in 1968 and 1969. The education variables represent the population share of residents with a respective level of education and those who will attain such a degree at a later point in time. Therefore, it also includes future education outcomes. All variables were calculated on the municipality level and aggregated to the group level as population-weighted averages. Standard deviations in parentheses.

in absence of the university. I already provided anecdotal evidence that the locations of the new HEI were selected quasi-randomly with respect to socio-demographic factors in section 3.2.1. Using the individual-level data introduced above, Table 3.1 compares "new uni" municipalities (column 3) in 1970 before the first new HEI was opened in 1971, with other potential candidates for a new college, i.e. municipalities without any HEI (columns 4 and 5), as well as with "old uni" municipalities (column 1) and their catchment area (column 2) to support that claim. In line with the above-described objective of dispersion, the new colleges were established in municipalities that had, on average, a high distance to the closest "old" university.[16] When focusing on comparing "new uni" municipalities (column 3) to other municipalities outside of the catchment area of "old" colleges in columns 4 and 5, the distance is not significantly different. In contrast, the "new uni" municipalities' population density is higher and remoteness is lower than in other municipalities without a HEI in 1970. A higher value in remoteness, defined as the sum of population-weighted distances to all municipalities, means that people in that region live relatively far away from the rest of the Swedish population. That emphasizes the importance of municipality-level fixed effect to control for level differences between municipalities, as will be discussed later.

Indicators regarding the population show only little differences when comparing municipalities without proper access to tertiary education in columns 3-5. People living in "new uni" municipalities were similar in terms of age, both on average and in the population share of young adults.[17] The probability to move at least once during the two years before 1970 was only slightly higher in "new uni" municipalities. Inhabitants of "new uni" municipalities show higher levels of education compared to those in other municipalities that did not have a university. However, it should be noted that the measure includes future education outcomes, which might include outcomes of the reform already. When looking at the population of 40-year-old residents there are no notable differences in the likelihood to be married. In contrast, "new uni" municipalities and their catchment area in columns 1

[16] Distance is measured between the centroids or mid-points of the municipalities.

[17] Comparing Figure 3.A.2 and Figure 3.A.3 in the Appendix provides additional evidence that my study population of high school graduates is similarly distributed as the total population when it comes to distance to the closest HEI.

and 2, i.e. always-treated municipalities, are less similar to municipalities without access to close by colleges. Hence, they will be dropped from the estimation sample. The exact empirical strategy including the underlying assumptions is discussed in section 3.5.

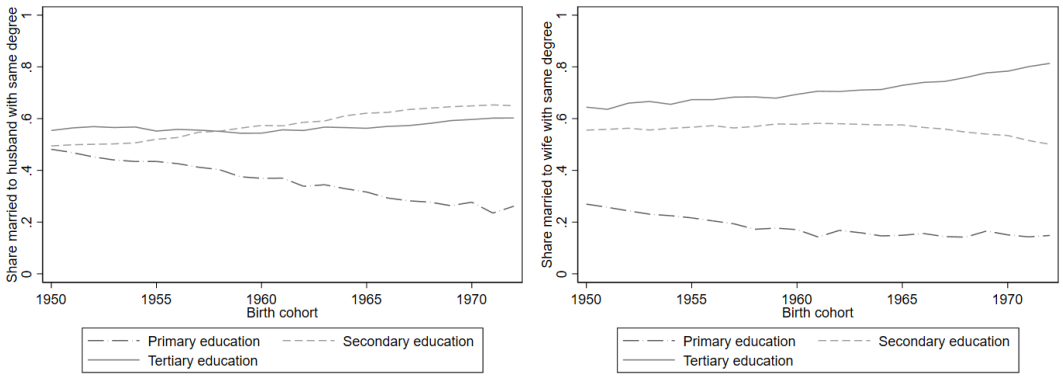
3.4 Marriage Market and Assortative mating

3.4.1 Observed Homogamy

The degree of sorting of individuals with similar traits in the marriage market can be measured as the share of couples that match with regard to assorted traits, e.g. level of education, which is called the rate of observed homogamy. It is one of my main outcome variables and was used to measure the degree of marital sporting in the literature in the early 2000s (Fernández and Rogerson, 2001). Conditional on being married, Figure 3.6 visualizes the share of individuals with a spouse of the same of three education groups: primary education, secondary education (high school degree), and tertiary education. The graphs show that the share of couples with the same level of

Figure 3.6: Observed homogamy rates by level of education

(a) Women (b) Men



Notes: Share of married individuals born between 1950 and 1972 with a spouse with the same level of education. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education. The left panel (a) displays the share of wives married to a husband with the same level of education. The right panel (b) refers to the share of husbands, respectively.

education is higher for post-primary-educated. For the cohort born in 1950, this relation is more pronounced for men in panel (b) than for women in panel (a). While men with secondary and tertiary education are roughly twice as likely to marry a wife of the same education group than primary-educated men, college-educated women have only a 10% higher degree of observed homogamy compared lower educated women. Looking at the changes over time reveals that the patterns between the genders become more similar. The share of women with post-primary education that are married to a husband with the same level of education increased to above 60% for the cohort born in 1972, especially for secondary-educated women, while the rate of observed homogamy for primary-educated women born in 1972 dropped to below 30%. For men, the gap between the education groups widened as well, but less than for women. Tertiary-educated men born in 1972 were four times more likely to marry a wife of the same education group than men of the same cohort with only primary education. The patterns are comparable to those observed in other countries, with few minor exceptions. For example, spouses without a high-school degree as well as with college education are more likely to be married to one alike for both genders in the US. In addition, the decline in sorting in the group of lower educated is not present in the US (for more details, see Figure 3.B.6 in the Appendix, based on results from Chiappori et al. (2020b)). Nevertheless, the fact that homogamy is stronger for higher levels of education without any sign of conversion for cohorts born between 1950 and 1972 is documented in many other countries (see also Blossfeld and Timm, 2003; Eika et al., 2019; Chiappori et al., 2020a; Gihleb and Lang, 2020).

3.4.2 Educational Assortative Mating

Nowadays, the degree of marital sorting is measured by the degree of assortative mating (AM) rather than by the rate of observed homogamy. AM is defined as the degree of homogamy compared to what would be expected under a random mating pattern. The reason is that observed homogamy depends on the underlying distribution of traits (e.g. Liu and Lu, 2006). For example, if the number of college-educated women is smaller than the number of college-educated men, the share of couples where both partners

have a college degree can never be 100%, as some men will not be able to match with a corresponding woman. This is especially relevant when using this measure to make comparisons between marriage markets with different distributions of traits, both cross-sectional as well as over time. As depicted in Figure 3.3 already, the share of educational groups is indeed suspect to significant changes in my study sample.

To account for changes in the distribution of education over time, I follow the approach of Liu and Lu (2006) by constructing a measure of assortative mating that is rescaled using a benchmark where individuals are perfectly randomly matched. In contrast to other methods of measuring assortative mating, the so-called perfect-random normalization allows me to calculate values for each educational group and gender separately.[18] Since calculating the randomly matching benchmark involves the distribution of traits at both sides of the marriage market, measures of AM require a definition of the pool of potential candidates. Most established measures, including the approach of Liu and Lu (2006), do this implicitly when restricting the sample to couples with at least one of the spouses belonging to a certain age group (see also Eika et al. (2019); Chiappori et al. (2020b)). By focusing on age instead of the year of birth, couples with a considerable age gap could be included twice: first when the older spouse belongs to the chosen age group and second time when the younger spouse crosses the lower bound of the age spell. Therefore, the result can only be interpreted as the degree of AM at a point in time rather than allowing any conclusion about changes in the degree of AM of birth cohorts. In addition, focusing on couples exclusively ignores endogenous effects of changes in the distribution of traits on the age differences between spouses, i.e. selection into the sample. Kirkebøen et al. (2021) improve on that by not restricting the sample to married couples and considering the distribution of traits in the total population. Still, they define the pool of candidates to consist of individuals of the same age group, implicitly assuming that there are no unions with a notable age difference. As the authors admit, that assumption might be problematic when the distribution of traits changes over time, as the true composition of the pool of candidates

[18] Chiappori et al. (2020b) discuss various measures of AM, including the perfect-random normalization based on the ideas of Liu and Lu (2006). Among other things, they show that the perfect-random normalization measure is equivalent to the minimum distance approach used, for example, by Abbott et al. (2019).

might differ from what is observed in the same age group. To investigate the development of AM in the presence of a shock to the distribution of traits for certain cohorts, I extend the AM measure of Kirkebøen et al. (2021) by explicitly modeling the pool of marriage candidates, accounting for cross-cohort effects of the change in the composition of traits.

Let the degree of assortative mating for given trait $e \in$ (primary education, secondary education, tertiary education), year of birth $c \in [1950, 1972]$ and gender $g \in$ (female, male) be described by $R_{e,c,g}$, scaling from 0 (no assortativity) to 1 (maximum of attainable assortativity given the underlying distributions of traits for men and women).[19] Similar to the approach by Kirkebøen et al. (2021),

$$R_{e,c,g} = \frac{h_{e,c,g} - h^r_{e,c,g}}{h^m_{e,c,g} - h^r_{e,c,g}} \tag{3.1}$$

rescales the observed homogamy rate $h_{e,c,g}$ plotted in Figure 3.6 by the range of potential values $h^m_{e,c,g} - h^r_{e,c,g}$. $h^m_{e,c,g}$ is the maximal attainable rate of homogamy if everyone tries to match to a spouse with the same trait e and $h^r_{e,c,g}$ is the level of homogamy one would expect when matching on traits is completely random. As highlighted above, I model both sides of the marriage market explicitly. Notation-wise, all variables with an apostrophe represent the opposite sex $g' \neq g$, i.e. the other side of the market also called the opportunity structure or supply side of the marriage market. Given that, let

$$s'_{e,c,g} \equiv \frac{P'_{e,c,g}}{P'_{c,g}} \tag{3.2}$$

denote the share with the same trait e in the pool of potential partners from the perspective of an individual with trait e, born in c with gender g. The (size of the) pool of mating candidates $P'_{c,g}$ for an individual of gender g born in c is calculated using cohort weights $w_{c',c,g}$. The cohort weight $w_{c',c,g}$ represents the probability of a (married) person of birth cohort c and gender g to be married to an (opposite-sex) spouse born in year c', derived from the sample averages of the universe of individuals born between 1950 and 1972 that got married before turning 41.[20] Hence, the cohort-adjusted size of the

[19] The measure can also turn negative if there is a preference for avoiding matches with similar traits. However, this is not the case with the traits of interest in the paper.

[20] Note that the weights are independent of any trait like the level of education. Technically, one could calculate weights for each education-cohort group as well. However,

pool of candidates for a person of gender g born in c is defined as

$$P'_{c,g} = \sum_{c'} w_{c',c,g} \times N_{c',g}, \tag{3.3}$$

where $N'_{c',g'}$ is the total number of individuals of the opposite sex $g' \neq g$ born in year c'.[21] Note that the restriction to my main sample of individuals born between 1950 and 1972 does not apply to the birth year of the spouse. Instead, all cohorts with observed unions with a person born between 1950 and 1972 are considered. The number of individuals with the same trait e in the pool of potential partners, $P'_{e,c,g}$, is defined equivalently applying the same weights on the sub-population with the respective trait e, $N_{c',e,g}$. Thus, $s'_{e,c,g}$ provides information on the proportion of potential mates with the same trait for any individual of cohort c and gender g with trait e. The development of the education shares in the pool of mating candidates is visualized in Figure 3.B.8 in the Appendix. Together with corresponding shares of traits at the supply side of the marriage market, $s_{e,c,g} \equiv \frac{P_{e,c,g}}{P_{c,g}}$ depicted above in Figure 3.3, I can derive the boundaries of homogamy. The level of homogamy under random matching is defined as

$$h^r_{e,c,g} = s_{e,c,g} \times s'_{e,c,g} \tag{3.4}$$

and the maximal attainable rate of homogamy as

$$h^m_{e,c,g} = \frac{\min(s_{e,c,g}, s'_{e,c,g})}{s_{e,c,g}}. \tag{3.5}$$

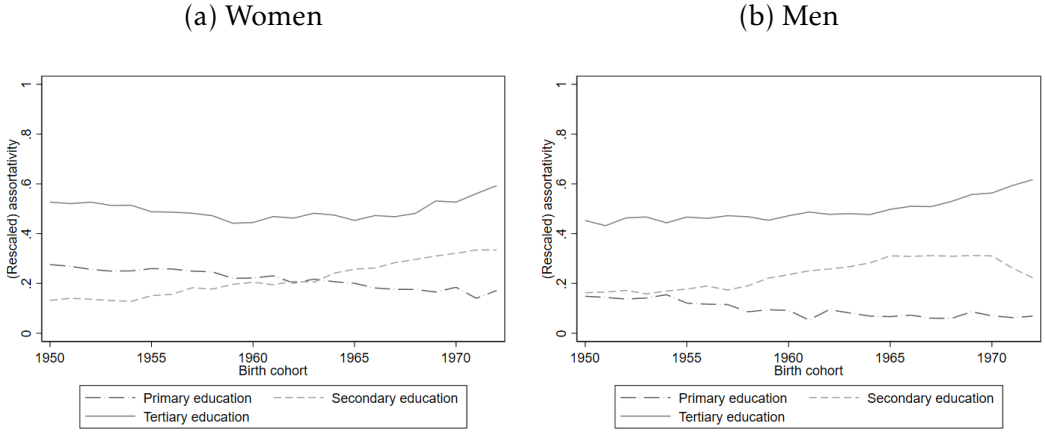
The level of homogamy under random matching, $h^r_{e,c,g}$, is nothing else than the fraction of the cohort-adjusted pool of mating candidates for an individual with gender g born in c with the same trait e. The higher the share of

that would incorporate changes in assortativity preferences which might be an outcome of the investigated education reform. Additionally, the distribution of age differences between spouses does not differ significantly between the education groups as depicted in Figure 3.B.7 in the Appendix.

[21] For consistency, I only used individuals married until the age of 40 here. However, the resulting assortative matching measure is qualitatively robust to in- or excluding singles when calculating the distribution of traits in the pool of candidates with $P'_{c,g}$ and $P'_{e,c,g}$, as the educational distribution is similar among non-married and married individuals (see Figure 3.B.9 in the Appendix).

individuals with trait e at the supply side of the marriage market, the higher the chance to match to someone with the same trait randomly. In contrast, a growing share of candidates with trait e increases the maximal attainable rate of homogamy, $h^m_{e,c,g}$, only if the trait is relatively underrepresented in cohort c for gender g compared to the respective supply pool of potential partners. Plugging $h^r_{e,c,g}$ and $h^m_{e,c,g}$ into Eq. 3.1 let me calculate the degree of assortative matching for all possible combinations of e, c, and g, visualized in Figure 3.7. Comparing educational assortativity with the rate of observed homogamy

Figure 3.7: Educational assortativity



Notes: Educational assortativity measure for individuals born between 1950 and 1972 by education and gender. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education. The left panel (a) displays $R_{e,c,\text{female}}$ of each level of education for women. The right panel refers to the educational assortativity of men, $R_{e,c,\text{male}}$.

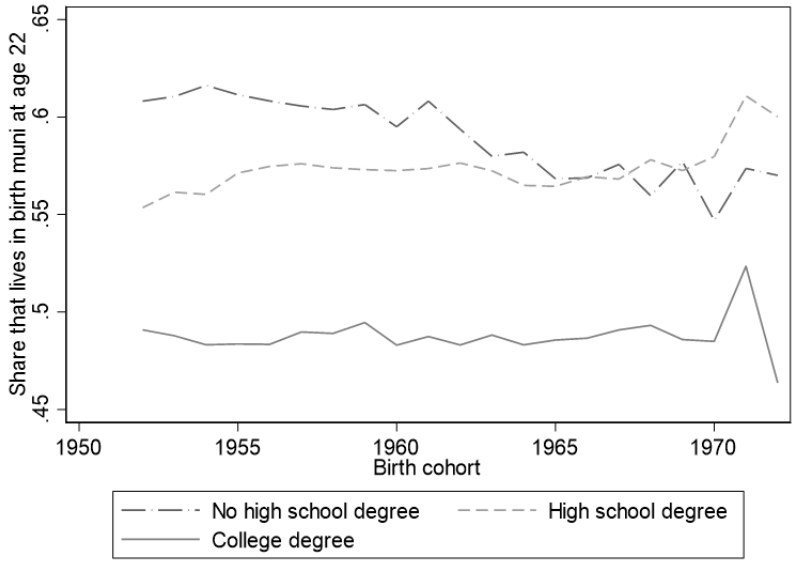
depicted in Figure 3.6, the degree of AM for tertiary-educated spouses is lower when accounting for the underlying educational distribution both for women (panel a) and men (panel b). However, the rescaling has an even stronger effect on the assortativity in the group with secondary education, making the positive relationship between educational assortativity and the level of education more noticeable at the top of the educational distribution. In contrast, the decline in educational assortativity of primary-educated is less pronounced when accounting for the relatively strong drop in primary education for both genders. Still, marital sorting on education among tertiary-educated individuals is 50%-100% stronger compared to secondary-

educated and up to 5 times higher compared to individuals with primary education throughout my main sample born between 1950 and 1972.

3.4.3 Marriage and Mobility

It is a stylized fact that mobility is positively correlated with education (Corcoran and Faggian, 2017). College students are more likely to move not only after attending college but already before (Markus, 2023). Figure 3.8 confirms that relation for the study population. A person who receives tertiary

Figure 3.8: Geographical immobility by level of education



Notes: Share of each cohort that lives in the municipality of birth at age 22 by the highest level of education. Education is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education.

education at some point in her life is roughly 10% less likely to live in the municipality of birth at age 22 compared to a secondary-educated individual who does not attend college and even 10-20% more mobile compared to those without secondary education. Most likely, this difference is driven by high school graduates moving closer to a university for receiving tertiary education, while most lower educated do not have to move for schooling or to enter the labor market. Psychological studies have documented that

existing social ties are often replaced by friendships at the new place of residence by moving (Magdol and Bessel, 2003). That also seems to be the case for mobility towards universities, as Seder and Oishi (2008) show that more mobile first-year college students have more new friends on campus compared to non-movers. In addition, Oishi (2010) emphasizes that mobile individuals have more freedom to choose their friends, e.g. based on preferences over traits, whereas rooted, non-mobile individuals tend to form friendships with members of a shared group, e.g. in the class at school or the local community. Concluding, stronger geographical mobility might be one of the culprits of the more homogeneous social network of college students proposed by Blossfeld and Timm (2003) and Blossfeld (2009). In

Figure 3.9: Educational homogamy and mobility

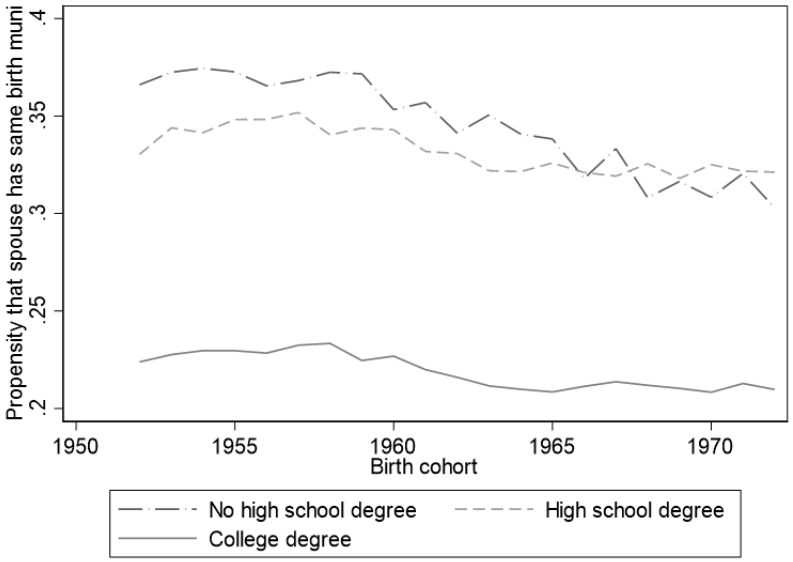
(a) With tertiary education (b) Without tertiary education

Notes: Share of married individuals born between 1950 and 1972 with a spouse with the same level of education by early adulthood mobility. The group of movers includes individuals that do not live in their municipality of birth at age 22, i.e. four years after completing high school. Stayers live in the same municipality they were born in at age 22. The left panel (a) contains only tertiary-educated spouses and the right figure (b), conversely, only those without (any) tertiary education.

line with that, Figure 3.9 shows that highly educated spouses have higher rates of observed homogamy if they do not live in their municipality of birth at age 22 anymore, while there is no such difference for less educated groups. That provides descriptive evidence that mobility in early adulthood might be a driver of educational sorting among tertiary-educated. The fact that differences in mobility are especially relevant before getting married rather than after forming a union supports that hypothesis (see Figure 3.B.11 in the

Appendix). Another way to provide evidence for the relationship between

Figure 3.10: Spatial homogamy



Notes: Share of married individuals born between 1950 and 1972 with a spouse born in the same municipality by level of education. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education.

education, mobility, and the choice of a partner is to look at rates of spatial homogamy in Figure 3.10. Spatial homogamy, i.e. the share of spouses that marry someone born in the same municipality, is more than 50% higher for primary and secondary-educated compared to tertiary-educated individuals. While roughly 1 out of 3 without tertiary education marry someone born in the same municipality, this is the case for only a little bit more than 20% of the most mobile education group of tertiary-educated. In line with that Haandrikman (2019) documents geographic mobility to explain a large part of the variation in spatial homogamy. However, the effect of mobility on spatial homogamy should not be interpreted causally. Among other factors, the documented correlation might be driven by sorting into migration. The empirical strategy to identify causal effects is discussed in the next section.

3.5 Empirical Strategy

The causal effect of variation in access to higher education induced by the Swedish expansion reform on the education and migration decisions of young adults has been documented in Markus (2023). Building on these results, I estimate the effects of a new university on marital outcomes. As highlighted above, the higher education expansion led to a massive drop in the average geographical distance to tertiary education.[22] Figure 3.5 documents a notable variation in the change of the distance to the closest HEI between municipalities, which is exploited in this paper. Since outcome variables as well as the treatment status are constant over time, my main sample is a repeated cross-section consisting of cohorts of 19-year-old high school graduates born between 1950 and 1972.[23] As the treatment (the opening of a new college) happened in different years, I estimate the dynamic treatment effect using a staggered DiD approach, also called event-study design. In line with the recent development in the two-way-fixed-effect literature (see Roth et al. (2022) and Chaisemartin and D 'Haultfoeuille (2022) for an overview), I exclude treated observations from the control group. However, not-yet-treated observations are part of the control group to increase the statistical power. I follow the approach of Gardner (2022) with the event-study regression equation

$$Y_{ist} = \lambda_s + \gamma_t + \sum_{r \neq -1} \beta_r D_{isr} + X_{ist} + \epsilon_{ist}, \tag{3.6}$$

[22] See also Figure 3.A.1 in the Appendix.

[23] My period of observation is 1968-2012. For the assignment of the treatment status, I need to observe individuals in the year they turn 18. Therefore, the first cohort I included in the sample was born in 1950. The upper bound of my sample is limited by outcome variables that require me to observe individuals up to the year they turn 40.

where individual i graduates from high school in municipality s in year t.[24] Y_{ist} is the outcome of interest, for example, an indicator that equals one for individuals who marry someone from the same education group or a dummy indicating whether an individual married someone born in the same municipality. $r \in \{-9, ..., -2, 0, ..., 15\}$ indexes the relative time-wise distance from the treatment and D_{isr} are indicators of treatment adoption.[25] D_{isr} equals one if person i graduates in a treatment municipality r years **after** the new university was opened for $r \geq 0$. Equivalently for the case $r < 0$, D_{isr} is zero for individuals completing secondary education while living in a treatment municipality $|r|$ years **before** the new university is established. For young adults that graduate from high school in control municipalities, $D_{isr} = 0 \; \forall \; r$. The coefficients β_r capture the dynamic treatment effect of interest for $r \geq 0$. They can be used as evidence for the plausibility of the parallel trend assumption before the treatment (i.e. when $r < 0$). λ_s and γ_t are municipality and time or cohort fixed effects. The former controls for all time-invariant differences between municipalities, while the latter controls

[24] It should be emphasized here that there is a difference between the year where I assign individuals to the place of residence and the year of (potential) treatment. As noted in section 3.3.3, I use the place of living in the year individuals turn 18 to ensure to not pick up a move right after finishing high school. Nevertheless, the actual year of potential treatment is the year t when graduating individuals turn 19, which is the year they finish secondary education. Therefore, the place of residence in the year of turning 18 is a proxy for the place of high school graduation. This is based on the assumption that there is no movement in the year of graduation before secondary education is actually finished. Taking the year of high school graduation as the period of potential treatment is especially important for cohorts turning 19 around the time the HEI is opened and, therefore, for the reference cohort. Take a cohort of high school graduates born in 1968, who turn 18 in 1976 in a municipality that is treated one year later, in 1977. If I used the year where observations turn 18 as the year of (potential) treatment, these individuals would be considered as not-yet-treated since they live in a treatment municipality in the year before the treatment happens, i.e. the university opens. However, that is only true for the year where I assign individuals to the place of residence, not the year where the young adults actually graduate from high school, which is 1977 when the new institution was already operating.

[25] The lower bound of the observation window is limited by the time of the majority of variation, which happens in 1977. The first observed cohort, born in 1950, graduates in 1969, which is eight years before 1977. Coefficients for the relative years before that can be estimated as well, but they rely on a small number of treated observations only. This is an unavoidable problem of event studies that have an unbalanced panel by construction. For that reason, the endpoints of the interval are binned following Schmidheiny and Siegloch (2022): Let l index the non-binned relative time, then $r = -9$ if $l \leq -9$ and $r = 15$ if $l \geq 15$. In contrast to the lower bound, the upper bound of the observation window is limited due to economic reasons. To limit the problem caused by potential general equilibrium effects, the observation window ends after 15 years after the university was opened.

for a national trend. Therefore, the coefficients are estimated by exploiting the variation between individuals *and* time only. Intuitively, coefficient β_r captures the change in the level-difference of the outcome between treatment and control group r years after (or before if $r < 0$) the intervention, relative to the difference in a reference period. Here, the reference period is $r = -1$, i.e. one year before a new college is opened. Or more precisely, the cohort graduating from high school one year before a new university opens in their municipality is the reference cohort.

Following the logic of Gardner (2022), I obtain the coefficients by applying a two-stage approach. First, I estimate the model

$$Y_{ist} = \lambda_s + \gamma_t + X_{ist} + \epsilon_{ist} \tag{3.7}$$

on the sample of individuals that lived in control municipalities at the age of 18 (i.e. $D_{isr} = 0 \; \forall \; r$) to obtain the estimated municipality and cohort effects $\hat{\lambda}_s$ and $\hat{\gamma}_t$ while adding individual level controls X_{ist}. In a second step, the adjusted outcomes $Y_{ist} - \hat{\lambda}_s - \hat{\gamma}_t$ are regressed on $D_{isr} \; \forall \; r \in \{-9, ..., 15\}$ to identify the average effects $E(\beta_{isr} | D_{isr} = 1)$. By using that approach, I also control for potentially heterogeneous treatment effects between different treatment cohorts (see Sun and Abraham, 2021; Roth et al., 2022).

Whether the estimates represent causal relationships depends on several identifying assumptions. The most important identifying assumption is the parallel trends assumption which states that the difference (in outcome) between the treatment and control group has to be constant over time in absence of the intervention. Or, in other words, the non-treated counterfactual of the treatment group is assumed to evolve in the same way as the control group. This means that cohorts graduating from high school in a municipality with a university opening would have made similar education and migration decisions in the hypothetical, counterfactual case of no university opening, as students graduating in municipalities without university openings. There are several arguments why this assumption is likely to hold here.

First, I argue that the new HEIs are as good as randomly assigned to locations in terms of my outcome variables.[26] As described in section 3.2.1, the

[26]Studies that use the same (Andersson et al., 2009; Nybom et al., 2022) or a similar

locations of the new universities were only chosen according to geographical arguments. Descriptive evidence presented in Table 3.1 indicates that population density might have played a role as well. However, since population density does not vary a lot over time the unit fixed effects absorb these kinds of level differences between municipalities. A comprehensible concern could be that the higher distance to the closest HEI *before* the reform could be correlated to a lower level of college education or a lower population share of young adults since they were forced to move away to study at a university. However, "new uni" municipalities are not notably different to control municipalities along those dimensions, including the average distance to the closest university. The quasi-random choice of locations provides an argument that there are no systematic differences between treatment and control regions. Therefore, I conclude that the allocation of the intervention (i.e. the location of HEI) was not depending on any of my outcomes. Second, the rich data set allows the estimation of the so-called "pre-trend". An insignificant treatment effect *before* the intervention provides additional evidence that the parallel trend assumption is satisfied. As visualized in the next section, the estimates for cohorts before a new institution is opened are mostly 0 for all outcomes and specifications. Third, the fixed effects absorb all observed and unobserved time-constant differences between municipalities and any kind of national trend that affects all municipalities similarly. In addition, I control for differences at the individual level like the parents' education and differences in migration costs by including family ties. However, my results are not sensitive to including these controls as shown in section 3.D in the Appendix.

Another important assumption is the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980). SUTVA is sometimes described as the assumption that each unit, including units from the control group, is only affected by its own treatment status. This implicitly rules out the relevance of general equilibrium and (spatial) spill-over effects. To minimize the influence of (local) general equilibrium effects I restrict the effect window to 15 years after a new institution was established. Since a new HEI might not only affect the access to education but also the local labor market, demographics,

(Suhonen and Karhunen, 2019; Berlingieri et al., 2022; Carneiro et al., 2023; Lehnert et al., 2020; Frenette, 2009) reform to obtain causal estimates use the same argument.

or local amenities of the location in the long run, the direct effect of the reform becomes harder to measure over time.[27] The exclusion of spill-over effects is traditionally problematic in spatial analysis (see Butts, 2021). A new HEI likely not only affects the location itself but also other regions close by since students could commute or move to the new college and meet their future spouses there which they would not have done without the new university. For that reason, I exclude the catchment area of "new uni" municipalities from the control group. Instead, I test for the existence of spatial spill-over effects by estimating the treatment effect for the catchment area while excluding the "new uni" municipalities themselves. To identify the reach of spill-over effects, I vary the size of the catchment area in Appendix section 3.D. For "old uni" municipalities and their catchment areas, there is no change in the distance to the closest HEI, so it seems unlikely that there are spill-over effects from new universities opened further away than the one that was already there. However, students that would have enrolled at an "old" university may decide to apply at a new college, leaving an additional university place open at the "old" HEI. Also, academic staff might relocate their workplace to a new college, which might also indirectly affect the local youth close to the "old" HEI. Hence, my main specification excludes the "old uni" municipalities (and their catchment area).[28] To avoid the contamination of my control group by already treated units, I follow the latest development in the DiD / event-study literature by excluding treated observations from my control group (see Goodman-Bacon (2021) for an intuitive explanation of that issue). Finally, my control group consists of all high school graduates in municipalities that have a distance higher than the catchment area cutoff at the time of graduation. That includes both individuals in "never-treated" municipalities (treatment group 1) over the full observation period as well as high school graduates in treated municipalities (treatment groups 4 and 5) before the new university was opened. Note that, as shown in Figure 3.5,

[27] Before the intervention, the effect window consists of eight years, since that is the maximum number of years I can observe cohorts graduating from high school before the 1977-reform. Every (relative) year outside of that chosen effect window is included in the estimation by binning at the endpoints according to Schmidheiny and Siegloch (2022). This implies that there must be no university openings before or after my observation window, which is unproblematic as I can observe the full universe of universities in Sweden.

[28] As a robustness check I show that my main results do not depend on excluding always-treated municipalities in section 3.D.

some of the so-called "never-treated" municipalities experience a drop in the distance to the closest HEI due to the reform as well. For the SUTVA to hold, I assume that a drop to a distance of more than the cutoff, for example, a drop from 120 km to 90 km, does not affect the migration, education, and marriage decisions of young adults.[29]

If the above-mentioned assumptions hold, my estimates represent the Average Treatment Effect on the Treated (ATT) (Lechner, 2011). The estimated coefficients have to be interpreted relative to the reference period one year before a new HEI was opened and indicate the change in the difference between average outcome levels of the treatment and control group.

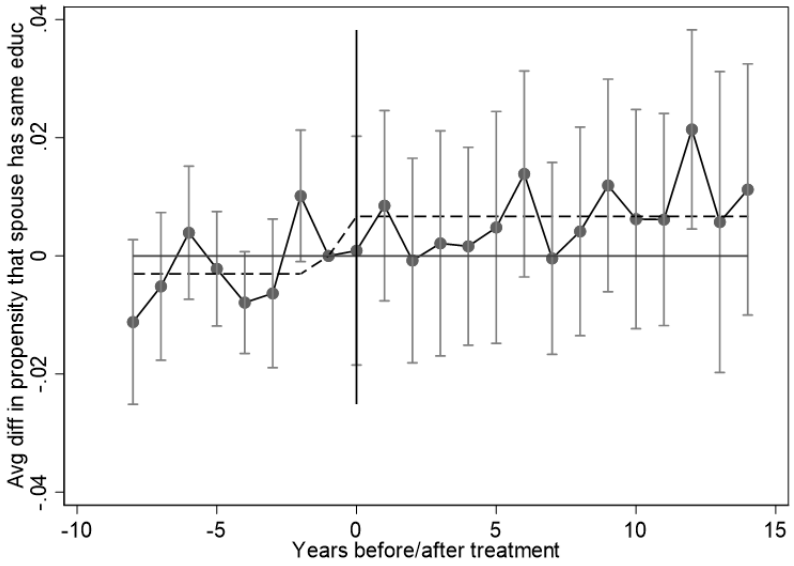
3.6 Main Results

3.6.1 Educational homogamy

As highlighted before, the effect of opening a new HEI on the marriage patterns of high school graduates is theoretically ambiguous. Figure 3.11 plots the estimated treatment effect of a new HEI on the propensity to marry someone with the same level of education, i.e. the effect on the rate of homogamy. It shows that, in contrast to the prediction based on Blossfeld and Timm (2003) and Blossfeld (2009), high school graduates finishing high school in a municipality where a new university opened are not more likely to marry someone from the same education class later in life, conditional on getting married before turning 41. As some of the observed graduates continue with their educational career by enrolling at a university the sample used here contains observations of two groups of education, namely (higher) secondary and tertiary education. Therefore, the null effect on homogamy could be a result of opposite effects on the two groups that cancel each other out. To be more precise, I distinguish between three categories of individuals concerning their response to a change in the treatment status:

[29] In Appendix section 3.C I test for the relevance of spatial spillover effects of the treatment. There is no evidence of an impact on the catchment area in terms of marital behavior. Varying the threshold of the catchment area definition of treatment group 2, i.e. the catchment area of "old uni" municipalities, separately provides evidence that there is also no spillover effects of old HEI on surrounding municipalities beyond 50 km. You find more details on that in section 3.D in the Appendix.

Figure 3.11: Educational homogamy



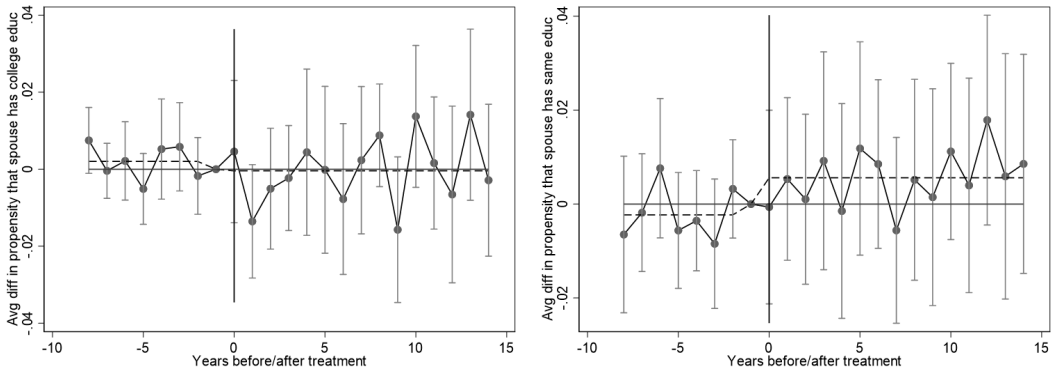
Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. Sample: population of all individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education), conditional on being married by the age of 40. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

always-takers always choose to receive college education, independent of the existence of a nearby HEI; compliers in "new uni" municipalities enroll at a HEI only when treated, i.e. after the new university opened in that region; never-takers stop their educational career after graduating from high school, regardless of the existence of a close-by college. The homogamy rate increases when the category of compliers is large enough and higher educated individuals are more likely to sort into homogeneous unions. The homogamy rate decreases, ceteris paribus, if always- and/or never-takers become less likely to marry a college-educated person. Always-taker might be more likely to marry someone from a lower education class because they did not move to a university town with a larger number of students (see Figure 3.2). Instead, they stayed in the municipality of high school graduation, being

more likely to keep social ties to peers without tertiary education (Seder and Oishi, 2008; Oishi, 2010). This also implies that never-takers become more likely to marry someone with a higher level of education than themselves, as their potential partners do not have to move to a different municipality to obtain tertiary education. In addition, the growing number of students in the home region increases the propensity to marry up, independent of preserved social ties. However, Figure 3.12 shows that there is no positive treatment

Figure 3.12: Marrying a college-educated spouse

(a) High school only (b) College only



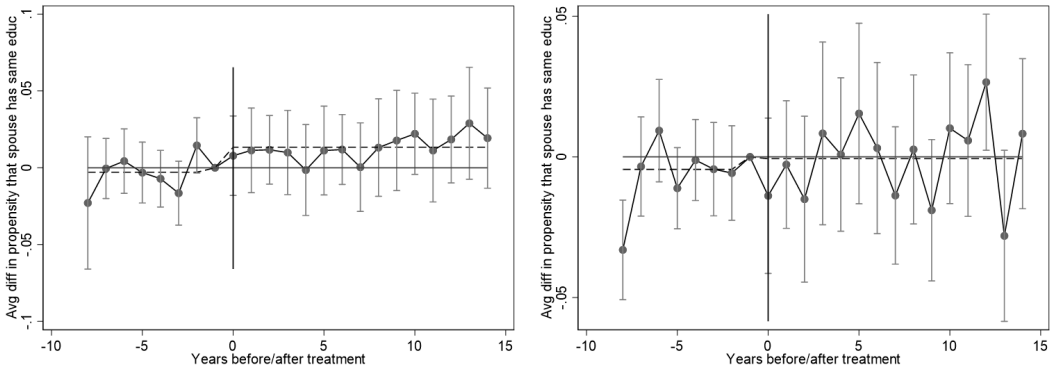
Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. Panel (a) uses the population of individuals born between 1950 and 1972 with a high school degree as the highest degree achieved, conditional on being married by the age of 40. Panel (b) includes college-educated only, conditional on being married by the age of 40. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

effect on the propensity to mate a college-educated person, neither for high school graduates who do not enroll at a HEI (i.e. never-taker) in panel (a) nor for higher educated (i.e. compliers and always-takers) in panel (b). To further investigate the treatment effect on the latter categories, compliers and always-takers, I separate the sample into movers and stayers. Stayers are defined as still living in the municipality of birth at the age of 22, while movers have a different municipality of residence at that age. As visualized in Figure 3.13 there is no effect heterogeneity by mobility. College students who still live in the home region by the age of 22 in panel (a) are not affected

differently by the new HEI with regard to the likelihood of being wedded to a spouse of the highest education class. Based on the idea that compliers

Figure 3.13: College homogamy and mobility

(a) Stayers (b) Movers



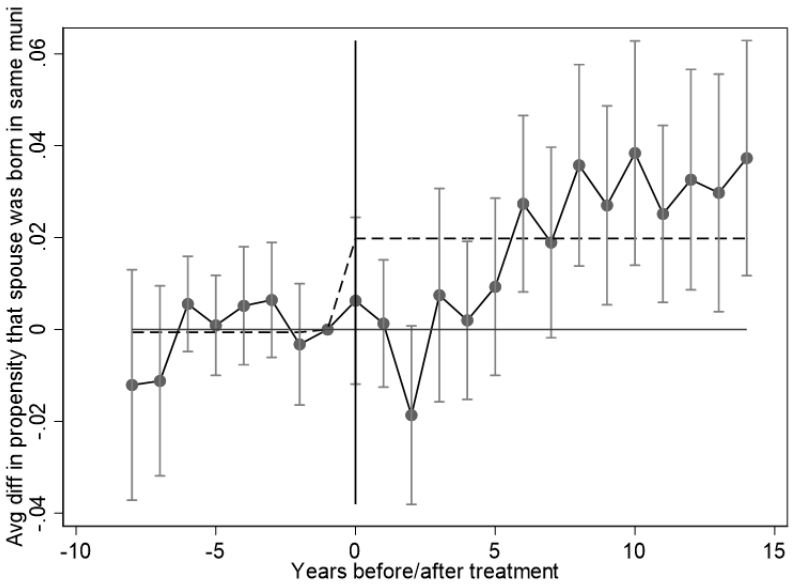
Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. Panel (a) uses the population of all individuals born between 1950 and 1972 who (will) have some college education and live in the municipality of birth at age 22, conditional on being married by the age of 40. Panel (b) includes all college-educated who do not live in the municipality of birth at age 22, conditional on being married by the age of 40. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

are mostly stayers, while always-takers could be both, these results leave only two possible explanations for not finding any effect of the education expansion reform on educational homogamy. Either, the positive effect on compliers is offset by a negative effect on always-taker who decide to stay in their home region. Or, there is no sizeable effect for both of these groups. The latter explanation, however, raises the question of why high-school graduates that are "pushed" into college education by the reform are not increasing the overall rate of homogamy. In the next section, I provide evidence that reduced mobility could explain this null effect.

3.6.2 Spatial homogamy

As discussed above, Oishi (2010) argues that less mobile individuals are more likely to keep social ties, e.g. from earlier stages of education, which

Figure 3.14: Spatial homogamy



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. Sample: population of all individuals born between 1952 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education), conditional on being married by the age of 40. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

makes their social network more heterogeneous. If a new HEI decreases the mobility of high school graduates, the limited integration into social networks of university students could explain the missing effect on educational homogamy documented above. Figure 3.14 presents the treatment effect on the propensity to marry someone who was born in the same municipality. On average, high school graduates finishing secondary education in a "new uni" municipality after the new HEI has opened are 2.77 percentage points more likely to mate with someone born in the same municipality, compared to graduates from "never uni" municipalities and "new uni" municipalities before the treatment. Given the sample average of spatial homogamy of 23.83%, this corresponds to a notable increase of 11.62%. The result is statistically significant at the 1% level. In line with expectations, the positive

impact on spatial assortativity is driven by less mobile individuals, as shown in Figure 3.C.13 in the Appendix. This provides additional evidence that the opening of a new HEI increases marital sorting on geographical aspects through the reduction in mobility of high school graduates.

3.7 Conclusion

In this paper, I address the question of whether migration plays a role in marriage patterns and how mobility interacts with educational homogamy. To overcome potential endogeneity issues, I exploit the exogenous effect of a university expansion reform on changes in access to tertiary education.

First, I document that the positive relationship between marital sorting and education in Sweden is even higher when considering the underlying distributions in the population compared to only observing the rate of educational homogamy. For that, I develop a measure that allows calculating the degree of assortative mating (AM) separately for each cohort and each education class by defining a pool of mating candidates.

In the second step, I use data on all universities and university colleges in Sweden and Swedish administrative data on the full population born between 1950 and 1972 to conduct an event study on individual-level marital outcomes. I provide suggestive evidence that mobility can explain higher rates of educational homogamy among college-educated spouses. While opening a new higher education institution does not affect the likelihood of local high school graduates mating with someone of the same education class later in life, I find a positive and significant effect on spatial homogamy of 11.62%. When a college offers access to tertiary education locally, high school graduates are more likely to stay in their municipality of birth and marry someone from the same region, compared to individuals in comparable locations without a close-by university who are forced to move to get higher education. My results indicate that an expected positive effect of the reform on educational homogamy might be offset by a negative effect induced by lower levels of mobility of students.

The results emphasize the importance of mobility for social outcomes like the choice of a partner. As mobility and education interact, especially for university students, this raises the question of whether effects that are

assumed to be caused by education are rather partially driven by differences in migration patterns for future research. This includes the role of increasing mobility for assortative mating and therefore ultimately for growing levels of inequality.

Bibliography

Abbott, B., Gallipoli, G., Meghir, C., and Violante, G. L. (2019). Education Policy and Intergenerational Transfers in Equilibrium. *Journal of Political Economy*, 127(6):2569–2624.

Andersson, R., Quigley, J. M., and Wilhelmson, M. (2004). University decentralization as regional policy: the Swedish experiment. *Journal of Economic Geography*, 4(4):371–388.

Andersson, R., Quigley, J. M., and Wilhelmsson, M. (2009). Urbanization, productivity, and innovation: Evidence from investment in higher education. *Journal of Urban Economics*, 66(1):2–15.

Anselin, L., Varga, A., and Acs, Z. (1997). Local Geographic Spillovers between University Research and High Technology Innovations. *Journal of Urban Economics*, 42(3):422–448.

Artmann, E., Ketel, N., Oosterbeek, H., and van der Klaauw, B. (2021). Field of study and partner choice. *Economics of Education Review*, 84(102149).

Becker, G. S. (1973). A Theory of Marriage: Part I. *Journal of Political Economy*, 81(4):813–846.

Becker, G. S. (1974). A Theory of Marriage: Part II. *Journal of Political Economy*, 82(2, Part 2):S11–S26.

Berlingieri, F., Gathmann, C., and Quinckhardt, M. (2022). College Openings and Local Economic Development. *IZA Discussion Paper*, 15364.

Bicakova, A. and Jurajda, S. (2016). Field-of-Study Homogamy. *CERGE-EI Working Paper Series*, 561.

Blossfeld, H.-P. (2009). Educational Assortative Marriage in Comparative Perspective. *Annual Review of Sociology*, 35(1):513–530.

Blossfeld, H.-P. and Timm, A. (2003). *Who Marries Whom?: Educational Systems as Marriage Markets in Modern Societies*, volume 12. Springer Science & Business Media, Dordrecht.

Bozon, M. and Heran, F. (1989). Finding a Spouse: A Survey of how French Couples Meet. *Population: An English Selection*, 44(1):91–121.

Burtless, G. (1999). Effects of growing wage disparities and changing family composition on the U.S. income distribution. *European Economic Review*, 43(4-6):853–865.

Butts, K. (2021). Difference-in-Differences Estimation with Spatial Spillovers. *arXiv working paper*, 2105.03737.

Carneiro, P., Liu, K., and Salvanes, K. G. (2023). The Supply of Skill and Endogenous Technical Change: Evidence from a College Expansion Reform. *Journal of the European Economic Association*, 21(1):48–92.

Chaisemartin, C. D. and D 'Haultfoeuille, X. (2022). Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey. *NBER working paper*, 29691.

Chiappori, P., Costa-Dias, M., Crossman, S., and Meghir, C. (2020a). Changes in Assortative Matching and Inequality in Income: Evidence for the UK. *Fiscal Studies*, 41(1):39–63.

Chiappori, P.-A. (2020). The Theory and Empirics of the Marriage Market. *Annual Review of Economics*, 12(1):547–578.

Chiappori, P.-A., Costa Dias, M., and Meghir, C. (2020b). Changes in Assortative Matching: Theory and Evidence for the Us. *NBER working paper*, 26932.

Corcoran, J. and Faggian, A. (2017). Graduate migration and regional development: An international perspective. *Graduate Migration and Regional Development: An International Perspective*, pages 1–10.

Deen, J. (2007). Higher education in Sweden. Technical report, Enschede.

Eika, L., Mogstad, M., and Zafar, B. (2019). Educational Assortative Mating and Household Income Inequality. *Journal of Political Economy*, 127(6):2795–2835.

Fernández, R. and Rogerson, R. (2001). Sorting and Long-Run Inequality. *The Quarterly Journal of Economics*, 116(4):1305–1341.

Fischer, M., Karlsson, M., Nilsson, T., and Schwarz, N. (2020). The Long-Term Effects of Long Terms – Compulsory Schooling Reforms in Sweden. *Journal of the European Economic Association*, 18(6):2776–2823.

Frémeaux, N. and Lefranc, A. (2020). Assortative Mating and Earnings Inequality in France. *Review of Income and Wealth*, 66(4):757–783.

Frenette, M. (2009). Do universities benefit local youth? Evidence from the creation of new universities. *Economics of Education Review*, 28(3):318–328.

Gardner, J. (2022). Two-stage differences in differences. *arXiv working paper*, 2207.05943.

Gibbons, S. and Vignoles, A. (2012). Geography, choice and participation in higher education in England. *Regional Science and Urban Economics*, 42(1-2):98–113.

Gihleb, R. and Lang, K. (2020). Educational Homogamy and Assortative Mating Have Not Increased. In Polachek, S. W. and Tatsiramos, K., editors, *Change at Home, in the Labor Market, and On the Job*, volume 48 of *Research in Labor Economics*, pages 1–26. Emerald Publishing Limited.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Greenwood, J., Guner, N., Kocharkov, G., and Santos, C. (2014). Marry Your Like: Assortative Mating and Income Inequality. *American Economic Review*, 104(5):348–353.

Haandrikman, K. (2019). Partner choice in Sweden: How distance still matters. *Environment and Planning A: Economy and Space*, 51(2):440–460.

Haandrikman, K., Harmsen, C., van Wissen, L. J. G., and Hutter, I. (2008). Geography Matters: Patterns of Spatial Homogamy in the Netherlands. *Population, Space and Place*, 14(5):387–405.

Han, S. and Qian, Y. (2021). Concentration and dispersion: school-to-work linkages and their impact on occupational assortative mating. *The Social Science Journal*, pages 1–18.

Högskoleverket (National Agency for Higher Education) (2006). Högre utbildning och forskning 1945–2005 – en översikt. Technical report, Högskoleverket (National Agency for Higher Education), Stockholm.

Holmlund, H. (2021). A researcher's guide to the Swedish compulsory school reform. *Journal of the Finnish Economic Association*, 1(1):25–50.

Holmlund, H. (2022). How Much Does Marital Sorting Contribute to Intergenerational Socioeconomic Persistence? *Journal of Human Resources*, 57(2):372 – 399.

Kamhöfer, D. A. and Westphal, M. (2019). Fertility effects of college education: Evidence from the German educational expansion. DICE Discussion Paper 316, Düsseldorf.

Kirkebøen, L., Leuven, E., and Mogstad, M. (2021). College as a Marriage Market. *NBER working paper*, 28688.

Kyvik, S. (2009). Geographical and Institutional Decentralisation. In Kyvik, S., editor, *The Dynamics of Change in Higher Education. Higher Education Dynamics*, pages 61–80. Springer, Dordrecht, 27 edition.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics*, 4(3):165–224.

Lehnert, P., Pfister, C., and Backes-Geller, U. (2020). Employment of R&D personnel after an educational supply shock: Effects of the introduction of Universities of Applied Sciences in Switzerland. *Labor Economics*, 66(101883).

Liu, H. and Lu, J. (2006). Measuring the degree of assortative mating. *Economics Letters*, 92(3):317–322.

Liu, S. (2015). Spillovers from universities: Evidence from the land-grant program. *Journal of Urban Economics*, 87:25–41.

Luo, S. (2017). Assortative mating and couple similarity: Patterns, mechanisms, and consequences. *Social and Personality Psychology Compass*, 11(8):1–14.

Magdol, L. and Bessel, D. R. (2003). Social capital, social currency, and portable assets: The impact of residential mobility on exchanges of social support. *Personal Relationships*, 10(2):149–170.

Markus, P. (2023). Effects of Access to Universities on Education and Migration Decisions. *Ruhr Economic Papers*, 996.

Marmaros, D. and Sacerdote, B. (2006). How Do Friendships Form? *The Quarterly Journal of Economics*, 121(1):79–119.

Mulder, C. H. and Malmberg, G. (2014). Local ties and family migration. *Environment and Planning A*, 46(9):2195–2211.

Nielsen, H. S. and Svarer, M. (2009). Educational Homogamy: How Much is Opportunities? *Journal of Human Resources*, 44(4):1066–1086.

Nybom, M., Plug, E., van der Klaauw, B., and Ziegler, L. (2022). Skills, Parental Sorting, and Child Inequality. *IZA Discussion Paper*, 15824.

Oishi, S. (2010). The Psychology of Residential Mobility: Implications for the Self, Social Relationships, and Well-Being. *Perspectives on Psychological Science*, 5(1):5–21.

Pestel, N. (2021). Searching on campus? The marriage market effects of changing student sex ratios. *Review of Economics of the Household*, 19(4):1175–1207.

Pollak, R. A. (2003). Gary Becker's Contributions to Family and Household Economics. *Review of Economics of the Household*, 1(1):111–141.

Premfors, R. (1984). Analysis in politics: The regionalization of Swedish higher education. *Comparative Education Review*, 28(1):85–104.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2022). What's Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. *arXiv working paper*, 2201.01194.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Schmidheiny, K. and Siegloch, S. (2022). On Event Studies and Distributed-Lags in Two-Way Fixed Effects Models: Identification, Equivalence, and Generalization. *ECONtribute Discussion Papers Series*, 201.

Schwartz, C. R. (2013). Trends and Variation in Assortative Mating: Causes and Consequences. *Annual Review of Sociology*, 39(1):451–470.

Seder, J. P. and Oishi, S. (2008). Friendculture: Predictors of Diversity in the Social Networks of College Students. In *Poster session presented at the annual meeting of the Society for Personality and Social Psychology, Albuquerque, NM*.

Suhonen, T. and Karhunen, H. (2019). The intergenerational effects of parental higher education: Evidence from changes in university accessibility. *Journal of Public Economics*, 176:195–217.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

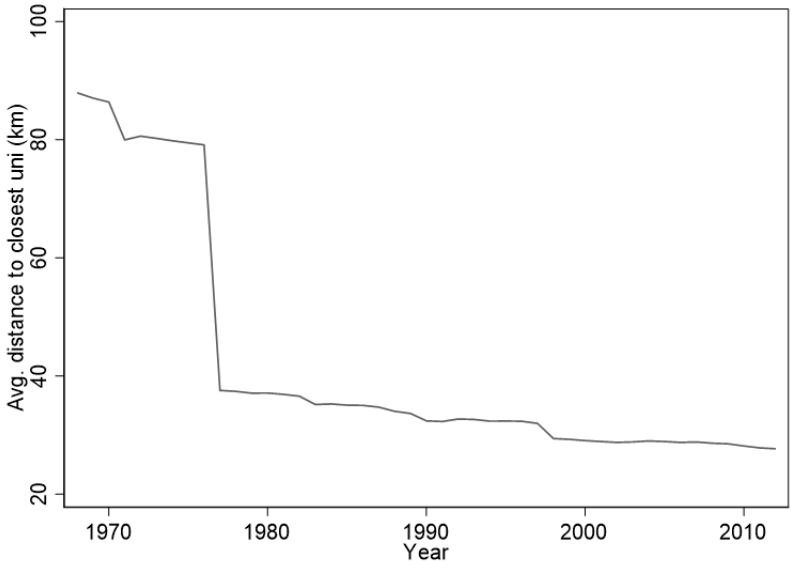
Swedish Government (1977). Regeringens Proposition 1976/77: 59. Technical report, Stockholm.

Varga, A. (1998). *University Research and Regional Innovation: A Spatial Econometric Analysis of Academic Technology Transfers*. Kluwer Academic Publishers, Dordrecht.

Appendix

3.A Distance to Closest HEI

Figure 3.A.1: Distance to the closest university



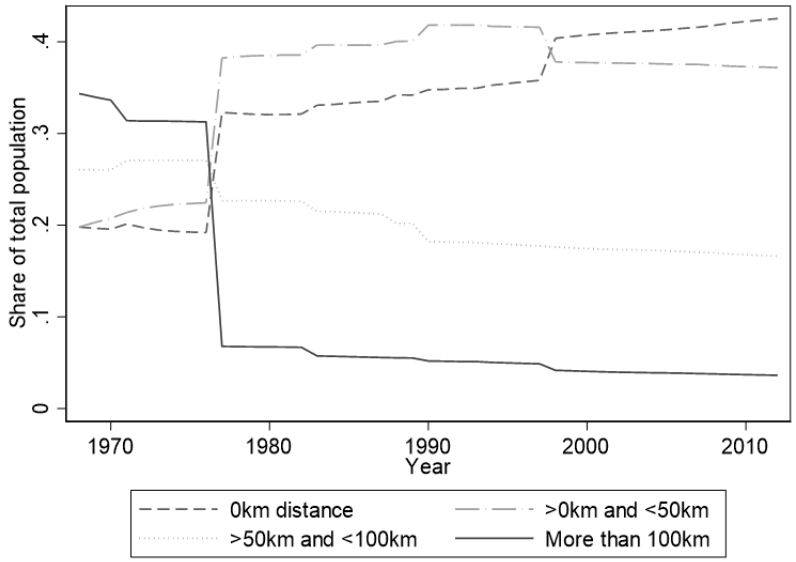
Notes: Population weighted average of the distance to the closest higher education institution. The population weights only use the 18-year-old population.

Figure 3.A.1 displays the aggregated effect of the university expansion reform on access to higher education for the population of 18 years old. While Figure 3.5 depicts the impact by treatment groups, Figure 3.A.1 shows that the reform had a substantial total effect. An average 18-year-old Swede lived more than 80 km away from the closest HEI in 1970 before the first new institutions opened. Already in 1977, the year of the main wave of the ex-

pansion reform, the average distance reduces to under 40 km, a distance that could theoretically be commuted. The distance reduces further afterward, but the main impact happened in 1977.

To learn more about how many residents were affected by the reform, Fig-

Figure 3.A.2: Share of total population by distance to closest university

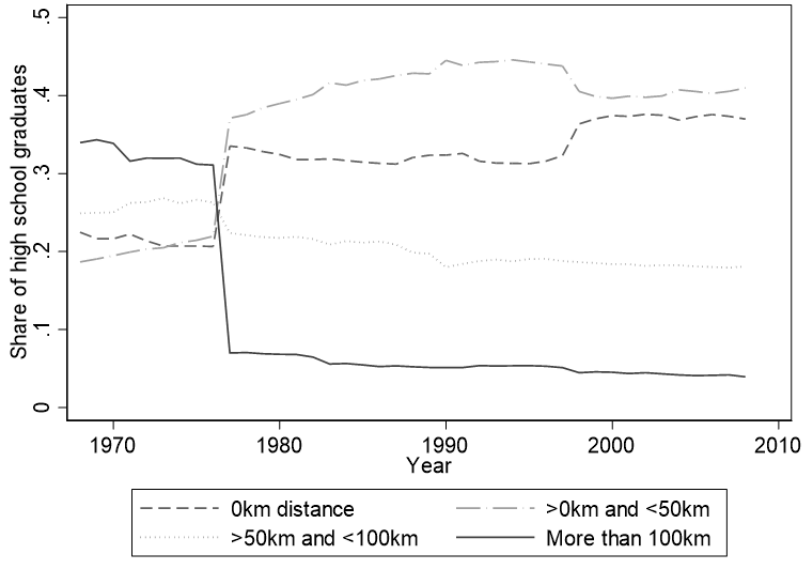


Notes: Share of the total population by distance to the closest higher education institution.

ure 3.A.2 plots the share of the total Swedish population by distance groups. In 1970, only 20% of the total population lived in municipalities with a university or in the 50 km catchment area, respectively. The former share increased to more than 30% in 1977, the latter almost doubled to roughly 40%. By definition, the share of the population that lived relatively far away from the closest HEI decreased at the same time. Interestingly, the drop was much larger for the distance group of over 100 km, emphasizing the stated goal of the reform to improve access to tertiary education, especially in areas where geographic access is low.

A similar pattern can be observed when only looking at the population of 18-year-olds that are about to finish high school as depicted in Figure 3.A.3. The similarity of Figure 3.A.2 and Figure 3.A.3 also shows that my study population of high school graduates is similarly distributed as the total population in terms of distance to the closest HEI.

Figure 3.A.3: Share of 18y old high school graduates by distance to closest university



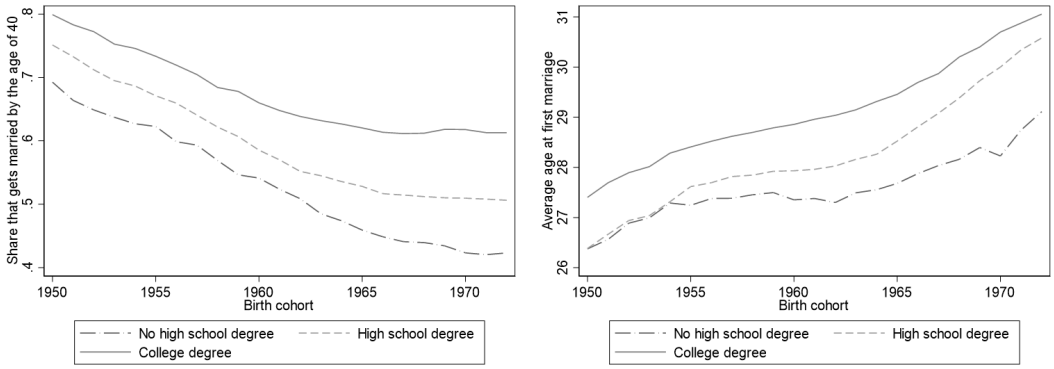
Notes: Share of the 18-year-old population that will finish higher secondary education by distance to the closest higher education institution.

3.B Marriage and Assortative Mating

My study sample consists of individuals with at least secondary education who marry at least once until the age of 40. There might be a concern that changes in the marital patterns that differ between education groups drive my results by sample selection. To address this concern, Figure 3.B.4 visualizes the dynamics of marriage habits. Two trends are potentially problematic. On the one hand, the left figure shows that, on average, the chance of getting married before turning 41 is larger for higher-educated individuals. Younger cohorts are less likely to marry and the decline is more pronounced for lower education classes. Among the 1972 cohort, a person who received tertiary education is more than 50% more likely to wed until reaching the age of 41. On the other hand, those born in 1972 who get married marry roughly three years later in their life compared the cohort. Again, the dynamic is stronger for higher levels of education. As expected, the more years of schooling the older individuals are at the time of getting married for the first time. Born in 1972, those with primary education marry at 29 years, while higher-educated

Figure 3.B.4: Marriage habits

(a) Getting married (b) Age at marriage



Notes: Panel (a) depicts the share of each cohort that married before turning 41. Panel (b) is the average age at the time of the first marriage, conditional on getting married at all before turning 41. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education.

groups are, on average, two years older. The concern that my main study sample becomes less representative for each education group, e.g. because disproportional many singles from the lower education classes or those who get married later in life are excluded is allayed by the fact that the changes over time counteract each other. The tertiary education group has the highest share of married individuals at age 40 despite being the group that marries the latest. While restricting the sample to wedded individuals excludes increasingly disproportionately primary-educated, limiting the sample to at most 40 years old spouses can be expected to drop more highly than lower-educated individuals.

Same-sex partnerships were legalized in 1995 in Sweden. Hence, it is not surprising that the share of spouses in same-sex partnerships among all individuals in a union is higher for younger cohorts, as depicted in Figure 3.B.5. Although increasing, the small absolute number limits the relevance of same-sex partnerships for my analysis. For all cohorts included in the sample, the share of same-sex couples never exceeds 0.6%. Consequently, excluding spouses in same-sex partnerships does not change any of the results in this paper.

To compare marital sorting tendencies with other countries, I reproduce

Figure 3.B.5: Share of same-sex marriages

Notes: Share of spouses that are married to a spouse of the same sex in their first marriage before turning 41.

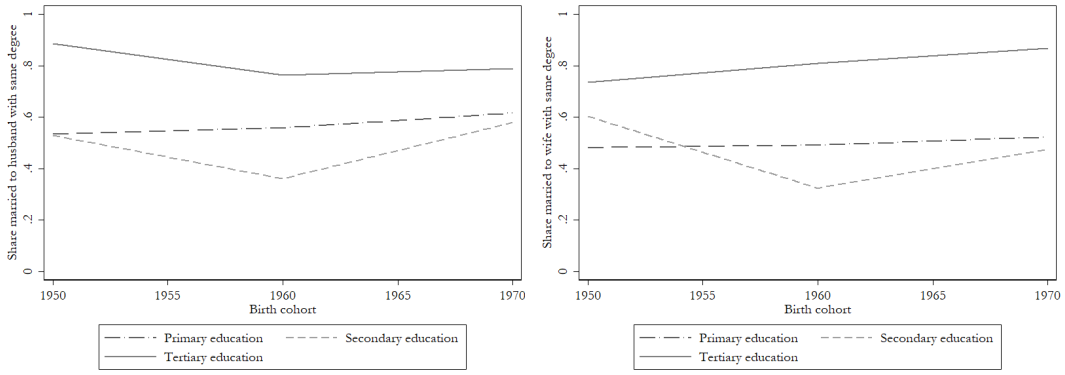
Figure 3.6 for the US based on data presented in Chiappori et al. (2020b), visualized in Figure 3.B.6. Similar to Sweden, the rate of educational homogamy is highest for those with more than a high school degree. Additionally, the numbers are comparable in magnitude.

Figure 3.B.8 visualizes the education shares in the pool of opposite-sex mating candidates for each cohort. As the supply side of the marriage market is calculated based on sample averages of unions between a given cohort and all other cohorts, the education shares of the candidates are essentially a smoothed version of the "real" education shares presented in Figure 3.3. That has some important implications. Changes in educational attainment are less relevant when comparing two cohorts with a small age difference. Consequently, reforms that are introduced based on the year of birth might be a bad source of variation to be studied in a regression discontinuity design as the supply side of the marriage market is affected similarly for individuals in the neighborhood of the cut-off.

Figure 3.B.9 reproduces Figure 3.7 while incorporating singles. As differences in the education share between genders are not significantly different

Figure 3.B.6: Observed homogamy in the US

(a) Women (b) Men



Notes: Share of married individuals born between 1950 and 1972 with a spouse with the same level of education. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education. The left panel (a) displays the share of wives married to a husband with the same level of education. The right panel (b) refers to husbands, respectively. The data is calculated based on results by Chiappori et al. (2020b), table 16.

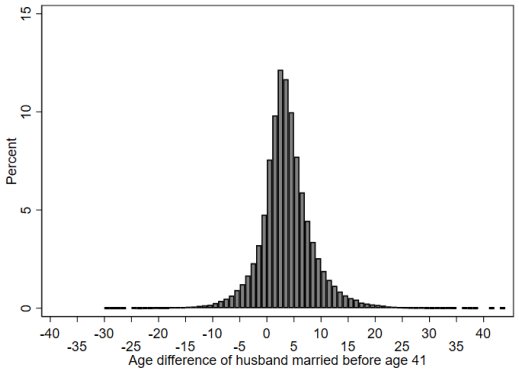
for the population of singles, the resulting degree of assortative mating does not change notably.

Figure 3.B.10 shows that there always were more college-education women than men in my observation period. The excess of females in the group of high highly educated even increases over time, as the tertiary education expansion affects women more stronger than men.

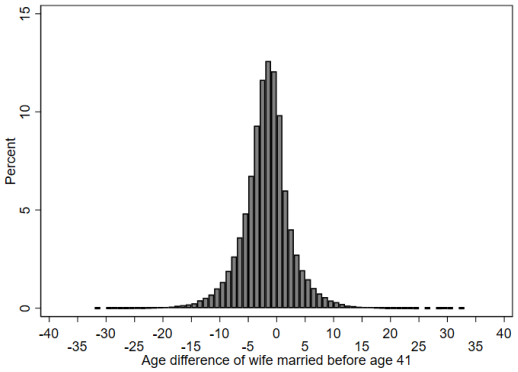
Figure 3.B.11 plots the propensity to move before and after the time of the first marriage by the level of education. Overall, mobility is higher in the years before the marriage and even increases towards the year of the wedding. Once in a union, the migration between municipalities drops sharply and continues to decrease, which is not surprising given the fact that many couples become parents at that time. Looking at the differences between education classes, it is visible that college-educated individuals are again more mobile than those with lower levels of education. The gap is especially large before forming the union, which is exactly the period I focus on in this paper.

Figure 3.B.7: Marital age differences by education and gender

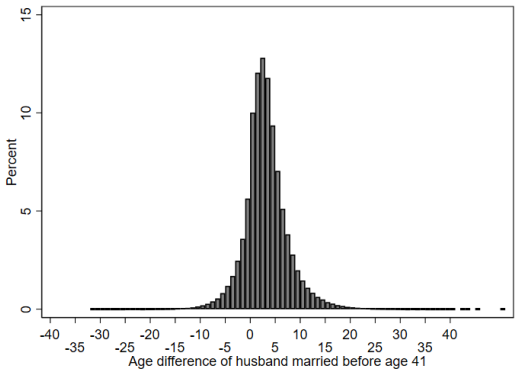
(a) Women, primary education



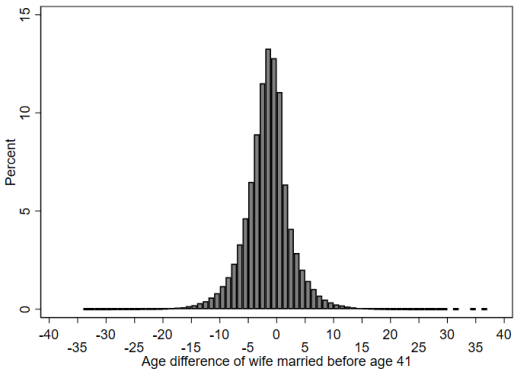
(b) Men, primary education



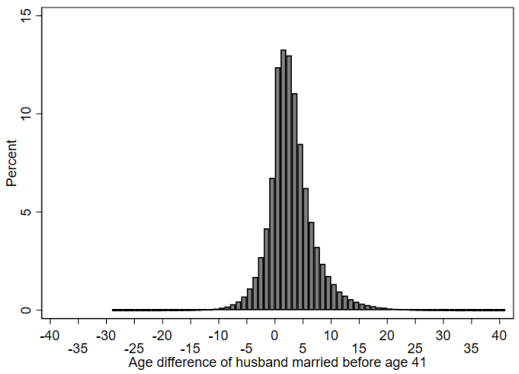
(a) Women, secondary education



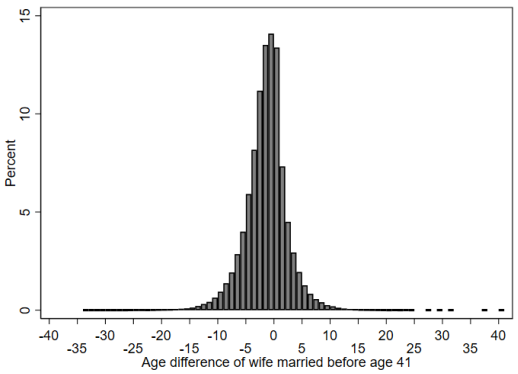
(b) Men, secondary education



(a) Women, tertiary education



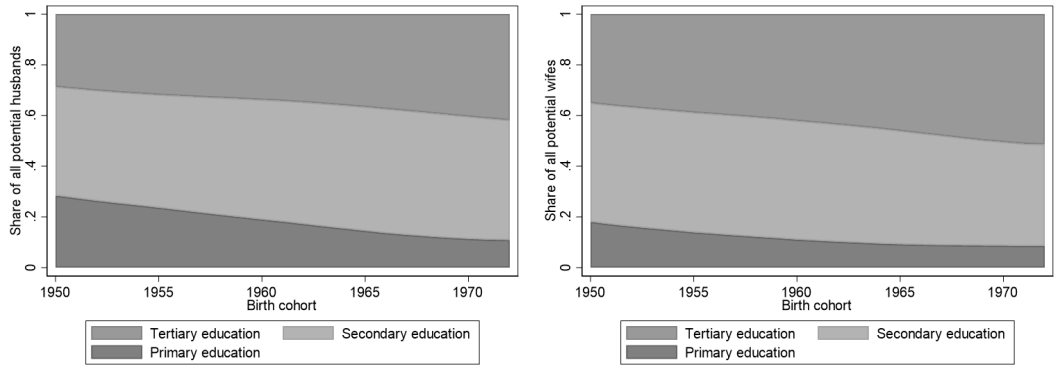
(b) Men, tertiary education



Notes: Distribution of age differences of spouses born between 1950 and 1972. The age difference is calculated as *age of spouse - own age*, such that a positive value indicates an older spouse. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma) on top, secondary education (high school degree) in the middle, and (some) tertiary education, i.e. college education at the bottom in line 3. The left panel (a) displays the age difference of husbands. The right panel (b) refers to the relative age of wives, respectively.

Figure 3.B.8: Education shares in the marriage market

(a) Candidates for women (b) Candidates for men



Notes: Relative supply of education in the marriage market of individuals born between 1950 and 1972. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education. The left panel (a) displays the share of each education class in the pool of potential candidates for women. The right panel (b) refers to the educational distribution of candidates for men. The pool of candidates is a cohort-weighted average of education shares of the opposite gender.

3.C Additional Results

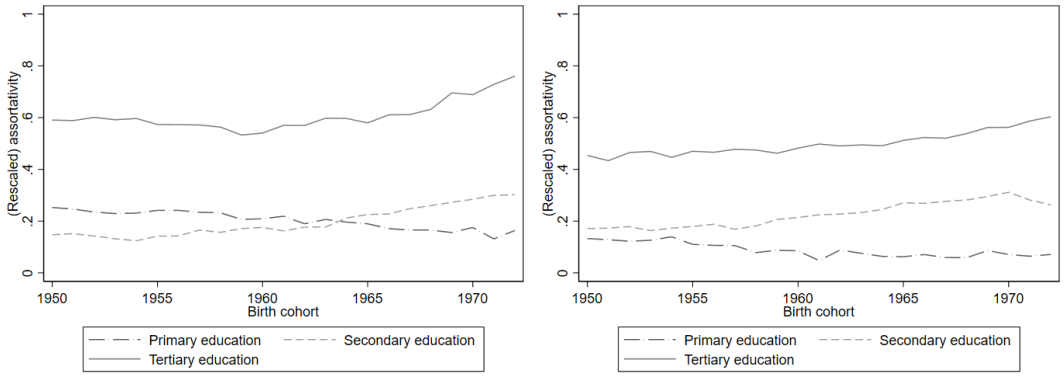
Figure 3.C.12 plots the treatment effect of a new HEI on the propensity to marry before turning 41. The reform has no sizable effect on whether individuals get married or not. Therefore, selection in or out of my main study sample, which includes only those who get married by the age of 40, due to the treatment status is unlikely to be an issue.

Spatial Spillover Effects

Figure 3.C.14 compares the treatment effect on educational homogamy in "new uni" municipalities in panel (a) with the effect on high school graduates in surrounding regions. Independent of the size of the threshold, there is no impact of the reform in the catchment area of "new uni" municipalities.

Figure 3.B.9: Educational assortativity including singles

(a) Women (b) Men



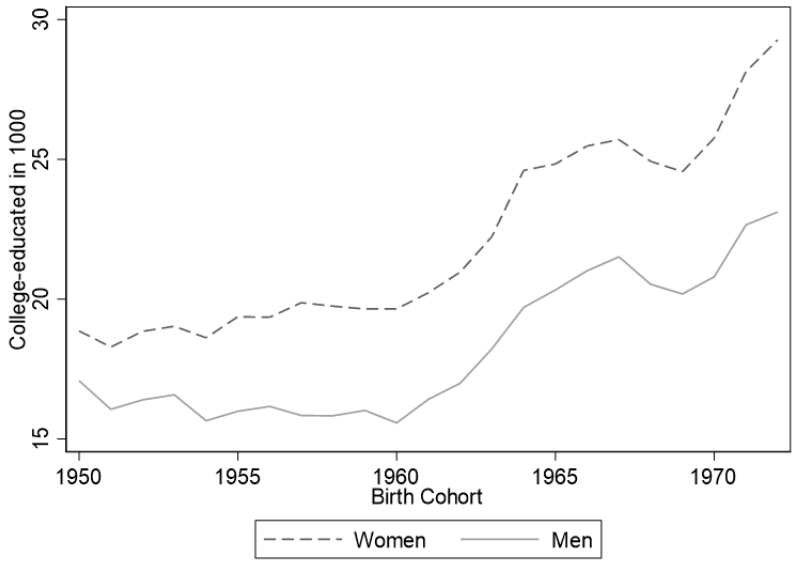
Notes: Educational assortativity measure for individuals born between 1950 and 1972 by education and gender. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education. The left panel (a) displays the $R_{e,c,\text{female}}$ of each level of education for women. The right panel refers to the educational assortativity of men, $R_{e,c,\text{male}}$.

3.D Robustness Checks

Controlling for Family Characteristics

One important determinant of participation in tertiary education is the educational background of parents. If parents have a college degree, children are more likely to enroll at a university, all other factors equal (Nybom et al., 2022). If high school graduates from treatment and control groups are systematically different in the level of education of their parents, the results presented above would be biased. In this section, I show that results do not differ when controlling for parents' education. It is defined as the level of the father's education (primary and no education, secondary education, or tertiary education and higher) or the mother's education if the father's highest degree is unknown. Since the educational register of Sweden does not cover information on individuals that died before 1990, I use the census of 1970 to add information on the highest degree for older cohorts. Additionally, Haandrikman (2019) results indicate that family ties play an important role in mating decisions. Hence, I control for the existence of a local family

Figure 3.B.10: Absolute number of college-educated



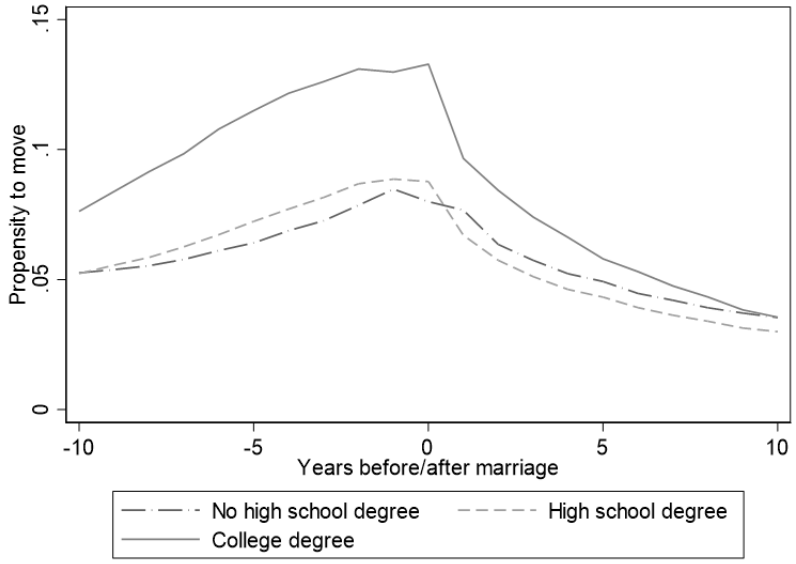
Notes: Total number of individuals born between 1950 and 1972 who attended college at some point.

network by adding an indicator equal to one if the individual lives in the same municipality as at least one of the grandparents at the age of 18. Young adults are assumed to have more local family ties if the family history shows a low level of mobility. Rerunning the main specification for educational homogamy including these controls gives similar estimated coefficients as plotted in Figure 3.D.15. Controlling for parents' education and the size of the local family network does not make a difference, presumably because the two-way-fixed-effects framework deals with potential (time-consistent) differences between treatment and control groups already (see Figure 3.11 for a comparison).

Including Always Treated Municipalities

Table 3.1 indicates that always-treated municipalities, i.e. regions that always had a university during my observation period, are not as comparable to my treated locations in terms of education and marriage decisions as the "never uni" municipalities. Therefore, "old uni" municipalities have been excluded, together with their catchment area, from the main specification. In this

Figure 3.B.11: Mobility around marriage



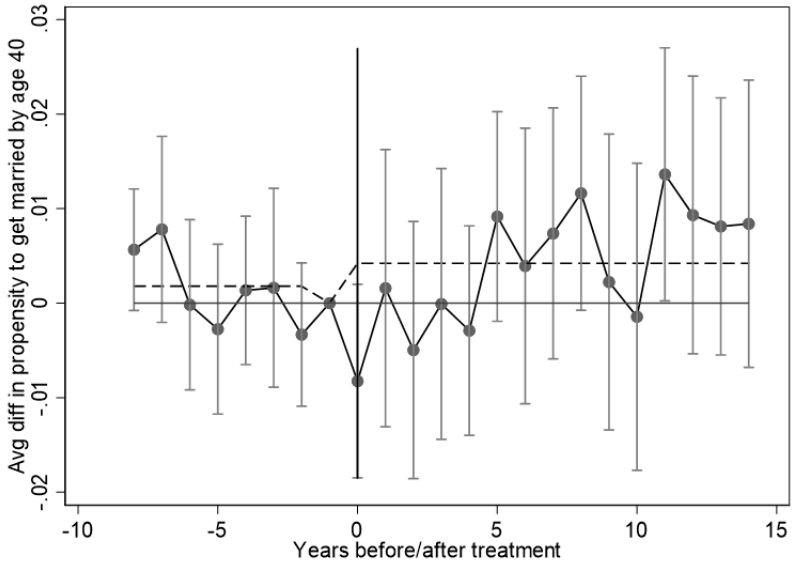
Notes: Likelihood to move between municipalities relative to the year of first marriage. The level of education refers to the highest level of education achieved and is grouped into primary education (no high school diploma), secondary education (high school degree), and (some) tertiary education, i.e. college education. All individuals born between 1950 and 1972 that got married before turning 41 are considered.

section, I show that including always-treated units in the control group does not change my results. Apparently, the municipality fixed-effect controlling for level differences between treatment and control group is sufficient to make ”old uni” regions and those who get a new HEI comparable in terms of educational homogamy, as indicated by Figure 3.D.16.

Excluding Malmö and Huddinge from Treatment Group

The municipalities of Malmö and Huddinge both received a new university in 1998. However, both municipalities were close to an old university even before that. While Lund is close to Malmö, Huddinge belongs to the greater area of Stockholm, although being its own municipality. Therefore, they could be assigned both to the group of treated as well as to the group of always-treated regions. As both locations received a new university in 1998, seven years after my youngest cohorts turned 19, high school graduates

Figure 3.C.12: Effect on getting married



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. All individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education) are included, independent of the marriage status.

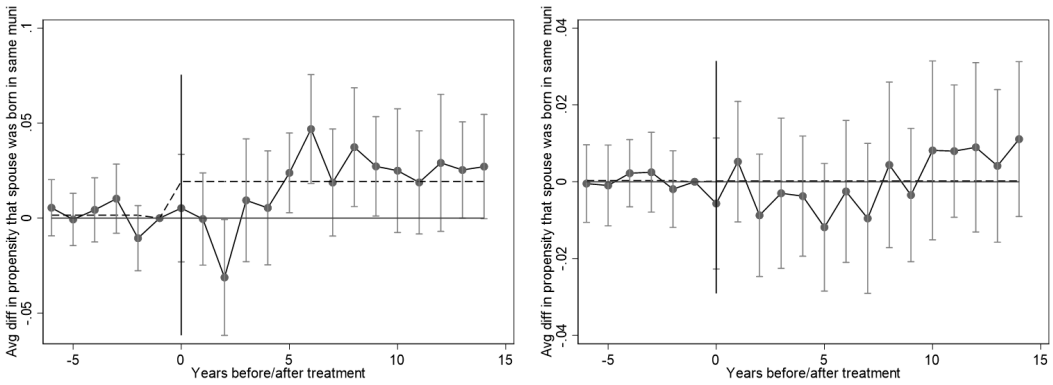
of these regions are part of the control group only. Here, I test whether dropping them from the sample makes a difference. As can be seen in Figure 3.D.17, the results are no sensitive to excluding the above-mentioned municipalities from the sample.

Different Definitions of the Catchment Area

Which municipality belongs to which of the five treatment groups as defined in section 3.3.3 depends on the threshold of catchment areas. A higher threshold includes more municipalities in the catchment areas of both, "new uni" municipalities (treatment group 4) and always-treated "old uni" municipalities (treatment group 2). The question is: how far does a university's (both old and new) effect on outcomes reach in terms of geographical distance? While results presented in section 3.C already indicate that there are no spillover effects of the treatment, i.e. on treatment group 4, I address a

Figure 3.C.13: Spatial homogamy and mobility

(a) Stayers (b) Movers

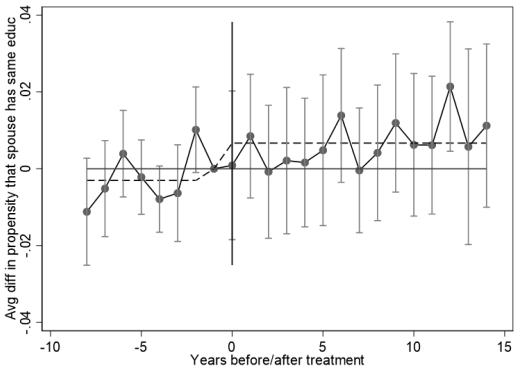


Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. Panel (a) uses the population of all individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education) and live in the municipality of birth at age 22, conditional on being married by the age of 40. Panel (b) includes all individuals of the same cohorts and education group that do not live in the municipality of birth at age 22, conditional on being married by the age of 40. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

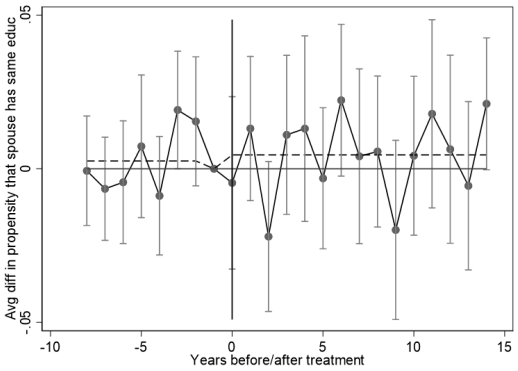
similar concern for treatment group 2 here. To do so, I vary the cutoff of the catchment area definition only for municipalities close to an old HEI while keeping the sample of regions in treatment group 2 fixed. Figure 3.D.18 allows comparing the estimated treatment effect on educational homogamy with different definitions of the catchment area of always-treated municipalities to results of the main specification, where all regions with a distance to an "old" HEI of below 50km are excluded from the sample. Independently of excluding more (75km catchment area) or less (30 km catchment area) locations from the control group, the estimated treatment effect on educational homogamy remains zero.

Figure 3.C.14: Spatial spillover effects on educational homogamy

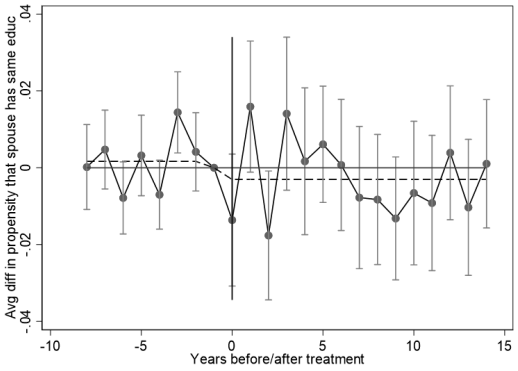
(a) "New uni" municipalities



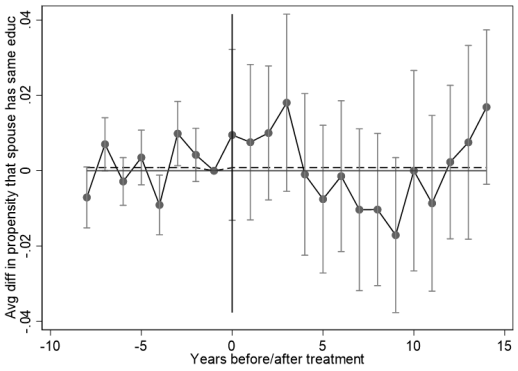
(b) 30km catchment area



(c) 50km catchment area

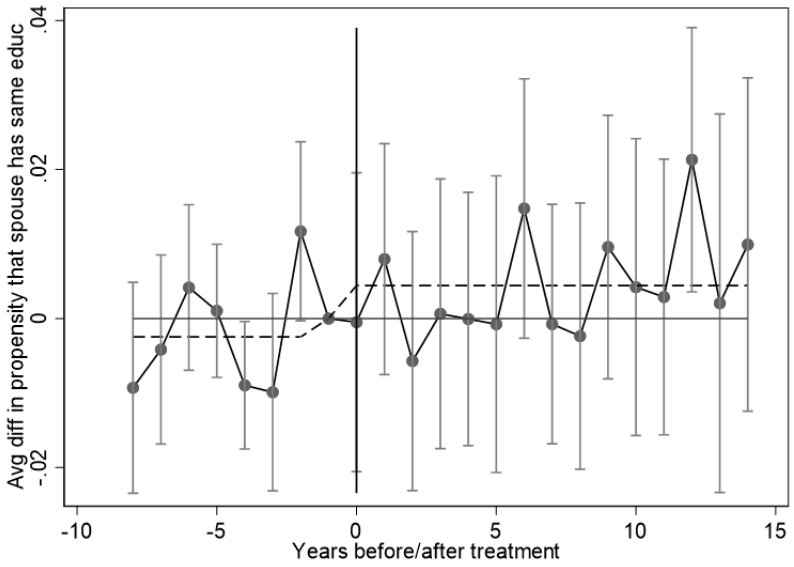


(d) 75km catchment area



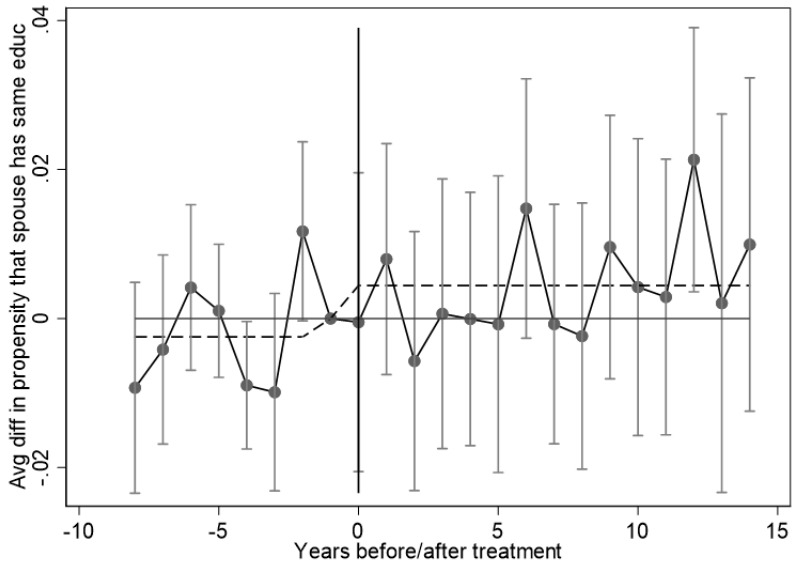
Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment of panel (a): new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. Treatment of panel (b)/(c)/(d): new higher education institution in a municipality at most 30km/50km/75km away. Excluding the 30km/50km/75km catchment area of treated municipalities and always treated municipalities. All individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education) and got married by the age of 40 are included. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 3.D.15: Educational homogamy with controls



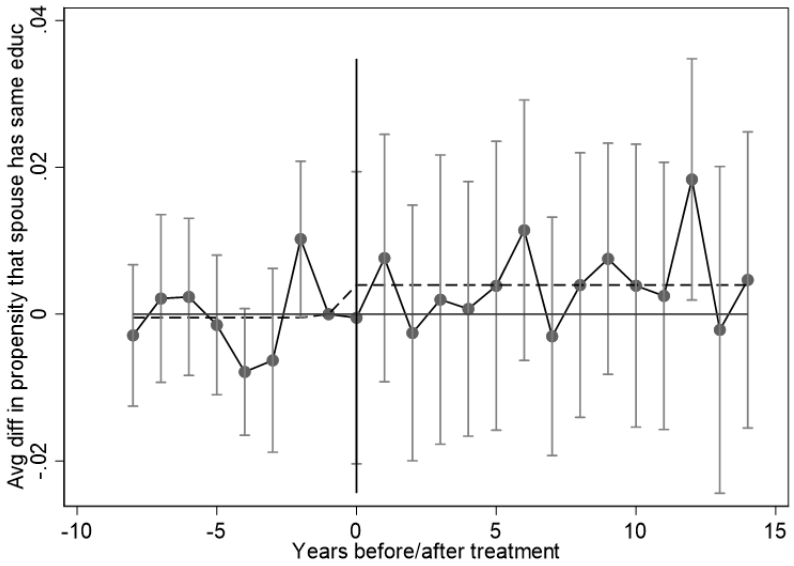
Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities. Sample: population of all individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education), conditional on being married by the age of 40. Controls: parental level of education and local family ties. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 3.D.16: Educational homogamy including always-treated



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Including the 50km catchment area of treated municipalities and always treated municipalities in the control group. Sample: population of all individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education), conditional on being married by the age of 40. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

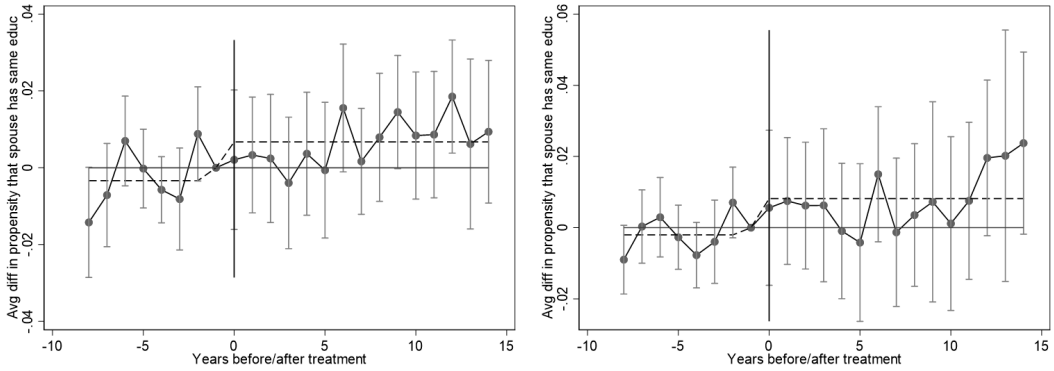
Figure 3.D.17: Educational homogamy including always-treated



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Excluding the 50km catchment area of treated municipalities and always treated municipalities, where Malmö and Huddinge are considered as being part of the latter. Sample: population of all individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education), conditional on being married by the age of 40. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Figure 3.D.18: Educational homogamy: varying catchment area size in control group

(a) 30km catchment area (b) 75km catchment area



Notes: Treatment status assignment by treatment status of the municipality of residence at age of 18 years. Treatment: new higher education institution. Panel (a) excludes the 30km catchment area of treated municipalities and always treated municipalities. Panel (b) excludes the 75km catchment area of treated municipalities and always treated municipalities. All individuals born between 1950 and 1972 who finished 12 years of schooling (higher secondary education, i.e. eligible for tertiary education) and got married by the age of 40 are included. Standard errors are clustered on the municipality level. The dashed line represents the average treatment effect before and after the reference period. 95% confidence intervals are displayed.

Concluding Remarks

In this thesis, I investigate the role of location characteristics in migration decisions. In a collection of three Chapters, I provide (I) an overview of potentially important measures of local amenities, (II) causal evidence for the importance of a certain location characteristic, the access to tertiary education, to be relevant for migration decisions of young adults and (III) show how such migration decisions at a relatively early stage of life can have far-reaching consequences for decision at later stages of life, for example, the choice of a spouse.

In the first Chapter, I review the literature on amenities and quality of life with a focus on providing a comprehensive collection of indicators used in previous academic literature and non-academic region and city rankings to represent regional attractiveness. It shows that the lack of a systematic approach makes selecting amenity measures very subjective and arbitrary, with the category of social indicators being widely ignored in urban economics. To provide a more objective way of selecting indicators, I use the statistical learning methods Lasso regression and random forest to identify the most relevant of over 100 measures for predicting quality of life differences between German counties. In line with my expectation, a social indicator, election turnout, was proved to be the best predictor of quality of life. However, other novel proxies of social participation, sports club membership rates and social media connectedness, are less relevant and require further, more targeted research in the future.

The second Chapter examines the location and education decisions of young adults. My results indicate that improved local access to tertiary education through a university opening increases college participation rates and makes high school graduates less likely to move away. Surprisingly, I

do not find any effect on the education decisions of young adults graduating more than 30 km away from a university. In contrast, effects on mobility are slightly positive, indicating that the effects of access to education differ between regions with a university and their catchment area. Future research could investigate why young adults from surrounding municipalities move toward the new university without participating in tertiary education. For example, peer effects and social ties could have an impact on the migration decisions of individuals who are not enrolled at the university themselves.

Hence, the third Chapter tests whether the opening of a new university impacts the decision of whom to marry. Despite the positive correlation between the level of education and the propensity to marry someone from the same education class, I do not find any evidence of access to universities affecting the level of education of spouses. However, local high school graduates become more likely to marry someone born in the same municipality after the new university opened. Given the results in Chapter 2, I am not able to fully disentangle the potential channels of education and mobility. Hence, further research is needed to examine the causal impact of mobility on marital sorting by education.

Overall, the results of my thesis emphasize the relevance of location characteristics for individual well-being and migration. I show that place-based policies like the opening of a new university can have a substantial and probably unintended effect on individuals' location decisions which interacts with many other aspects of life.

Erklärung

gemäß §10 Abs. 6 der Promotionsordnung der Mercator School of Management, Fakultät für Betriebswirtschaftslehre der Universität Duisburg-Essen, vom 11. Juni 2012.

Hiermit versichere ich, dass ich die vorliegende Dissertation selbständig und ohne unerlaubte Hilfe angefertigt und andere als die in der Dissertation angegebenen Hilfsmittel nicht benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht.

Duisburg, 11. April 2023

Philipp Markus

DOI: 10.17185/duepublico/78985
URN: urn:nbn:de:hbz:465-20230907-074631-5