Comparative Population Studies
Vol. 45 (2020): 143-178 (Date of release: 09.04.2020)

How Well Can the Migration Component of Regional Population Change be Predicted? A Machine Learning Approach Applied to German Municipalities

Hannes Weber

Abstract: For several decades, demographic forecasts had predicted that the majority of Germany's cities and rural areas would experience population decline in the early 21st century. Instead, recent trends show a growing population size in three out of every four German districts. As a result, there are currently severe shortages of housing and childcare in regions that were projected to decline but have instead grown in recent years. Other regions, by contrast, continue to lose young people in particular. Most of these differences between regions stem from within-country as well as international migration. An important question for both regional demographic research as well as local policy-makers is thus how well net migration rates in cities and rural districts can be predicted several years into the future. In this study, we develop models that predict migration (both within-country as well as international migration) at the level of municipalities for two demographic groups, namely young people aged 18 to 24 years, and families (people aged 30 to 49 years and underage children). We collect data on economic, demographic and other characteristics such as distances to large cities or universities for around 3,000 German municipalities (*Gemeinden*). The model is trained on a subset of these data from the period 2005-2009 and predicts net migration rates among young people on an unseen test dataset in the future (i.e. for the period 2011-2015). The results show that the model can predict future net migration by young people aged 18 to 24 years reasonably well ($R^2 > 0.5$), although there were quite significant changes during the period under study, for example refugee immigration to Germany. Family migration, on the other hand, cannot be predicted equally well ($R^2 = 0.25$). Some important lessons emerge concerning the predictability of regional and international migration and the usefulness of demographic forecasts for local policy-makers.

Keywords: Migration · Machine learning · Municipalities · Prediction · Demographic change

URL: www.comparativepopulationstudies.de
DOI: 10.12765/CPoS-2020-08en
URN: urn:nbn:de:bib-cpos-2020-08en5

1 Introduction

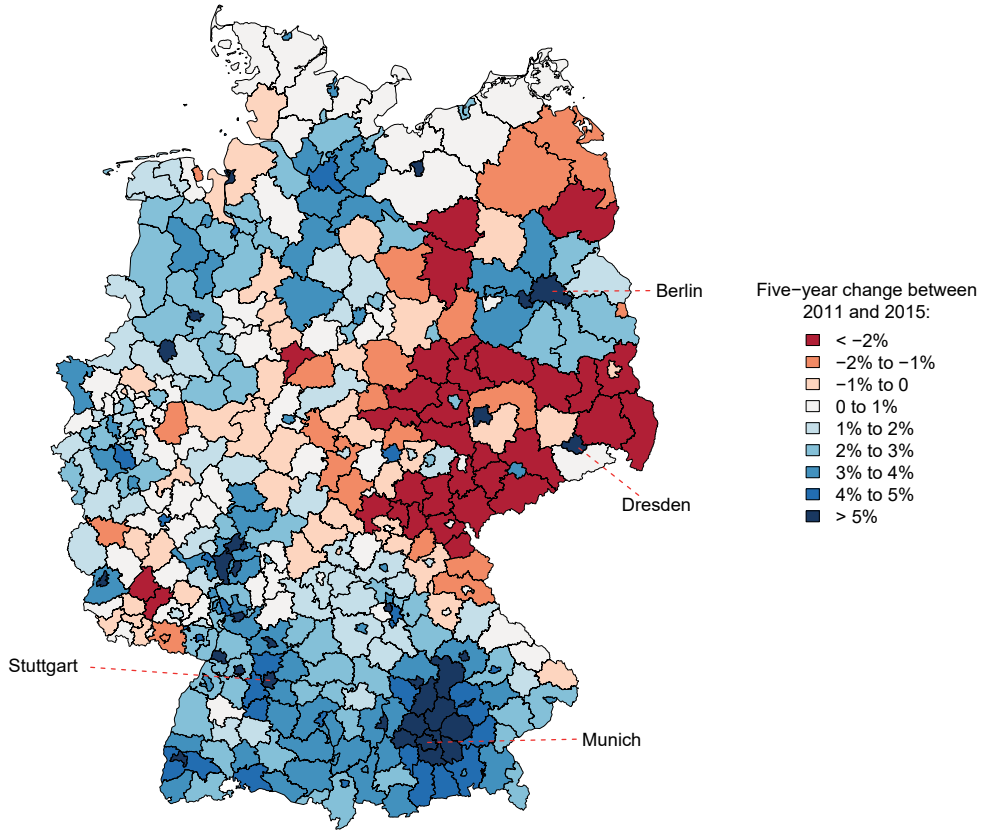
Germany's population will decline drastically throughout the 21st century – this had been the predominant belief among demographers, politicians and the general public in Germany for decades. In its 2009 forecasts, the German Federal Statistical Office was still predicting a substantial drop in the national population from 82 million to between 65 and 70 million by 2060 (*Statistisches Bundesamt* 2009). In recent years, however, the demographic outlook has changed dramatically. As of 2019, more people live in Germany than ever before. And according to the 2019 revision of the official forecasts, this is unlikely to change in the coming decades (*Statistisches Bundesamt* 2019). Instead of the "demographic implosion" that had often been envisioned, Germany's population size is likely to remain above 80 million by mid-century. The Federal Statistical Office's estimates are in line with other recent forecasts (e.g. *Fuchs et al.* 2018; *Vanella/Deschermeier* 2018; *Weber* 2019) which arrive at much higher values for mid-century population size than had long been assumed.

It goes without saying, however, that there are strong regional differences within Germany regarding the demographic outlook. Some metropolitan areas are preparing for population density to increase more and more in the near future, whereas other regions are already experiencing quite dramatic rates of out-migration by mainly young people and, as a consequence, accelerated demographic ageing and decline (*Schlömer* 2006; *Mai/Scharein* 2009). But again, local forecasts have been proven wrong by the reality in many regions. For instance, a decade ago, the German Federal Institute for Research on Building, Urban Affairs, and Spatial Development (BBSR) predicted that the majority of districts and cities (*Kreise und kreisfreie Städte*) in Germany will experience a population decline between 2005 and 2025 (*BBSR* 2008). Even in the federal states of Bavaria and Baden-Württemberg in the south of the country, both of which experienced strong economic and demographic growth in the past, many regions were projected to stagnate or decline (see also *Swiaczny* 2015).

This is in contrast with the actual recent (2011-2015) trend displayed in Figure 1. The general geographic pattern of growth versus decline matches the predictions by the BBSR and others: the populations of rural regions in the eastern states mostly declined, whereas big cities and their surrounding areas increased in population size. However, there are also marked differences between predictions and reality. Between 2011 and 2015, only 23 percent of all German districts actually experienced population decline, instead of the majority that had been expected by the BBSR. This means that around 30 percent of all German districts – many of them smaller cities and rural areas – were projected to decline but instead gained inhabitants.

Of course, it is still uncertain whether this trend will continue over the coming years. But the current trend already has considerable consequences in many areas, including a severe shortage of housing and child care, overcrowded public and private transport in large as well as smaller cities, and a considerable backlash against efforts to protect the environment and the fight against climate change (*Weber/Sciubba* 2019). In the past, many cities got rid of public housing projects or cut educa-

Fig. 1: Population growth in German districts between 2011 and 2015



Note that the subsequent analyses will use data on the smaller-scale municipalities (Gemeinden, N > 10,000), but the visual inspection is easier using the larger aggregation.

Source: own calculations, *Bertelsmann Stiftung* (2018)

tion budgets because it was thought they were not needed in a future characterised by population ageing and decline. Today, soaring housing prices and a shortage of teachers and child care workers are among the main problems frequently cited by local politicians. Thus, demographic forecasts that turn out to be wrong can clearly cause substantial problems in many policy areas.

The main reason for the sudden turnaround in German demography is obvious. Between 2012 and 2016 alone, net immigration to Germany amounted to more than three million people, whereas the *Federal Statistical Office's* (*Statistisches Bundesamt* 2009) forecasts only expected between half a million in the "low" immigration variant and one million immigrants in the "high" variant to settle permanently in the same period. Not only did the surge in refugee migration (especially during 2015) contribute to these figures, but also sharp increases in other types of immigration,

above all intra-EU movements from the new accession states in the East and South-East.

International migration is the most difficult component in demographic projections. Sharp fluctuations in migration are the main reason why demographic forecasts in Europe have not become more accurate over time, despite massive advances in computational power and statistical methods (*Keilman* 2008). As an example, the forecasts by *Alders et al.* (2007) assumed that migration to Germany until 2050 would be moderate, and lower than migration to other EU countries such as Portugal or Spain. The authors justified this by stating that the high flows of refugees were "probably over", labour migration from Central and Eastern Europe was assumed to be more balanced, and high unemployment in Germany would deter immigration (*Alders et al.* 2007: 60). All of these trends have seen a drastic reversal just a few years after publication of the forecasts.

Evidently, it is still unclear whether net migration to Germany will remain higher than the Federal Statistical Office's "high immigration" variant of 200,000 people per year as is currently the case (in 2018, this figure was 386,000), or whether it will, for instance, drop to the previous decade's (2000-2009) average figure of just 96,000. But migration has by far the largest impact on the accuracy of demographic forecasts (e.g. *Weber* 2015a). Fertility and mortality have only changed gradually throughout the past decades, but migration has been very volatile, changing from negative net migration in 2009 to a historical record figure in 2015, all within the space of a few years. Moreover, contrary to previous occasions when vast differences were recorded in fertility or mortality rates between cities and rural areas, today there are only minor differences across German regions as far as these two indicators are concerned (*Bujard/Scheller* 2017). Thus, the uncertainty of demographic forecasts for cities and regions stems mainly from inter-regional and international migration.

How well, then, can regional population figures be forecast into the future? The aim of this study is to develop a model to predict future net migration (including both within-country mobility and international migration) for German municipalities based on various socio-economic, demographic, and geographic characteristics. We focus on relocations into municipalities by two demographic groups, namely young people aged 18 to 24 years, and middle-aged families (people aged 30 to 49 with underage children). Attracting younger demographic groups is often stated as a goal for cities and regions that are facing population ageing and decline. This does not mean that only young people contribute to economic growth; on the contrary, research finds that having many workers in their forties leads to greater levels of productivity in a country as opposed to having a younger workforce (*Feyrer* 2007). In addition, many demographers challenge the notion that low fertility and population ageing are per se problematic, finding that per capita income can actually benefit from below-replacement fertility and a moderately declining population size (*Lee et al.* 2014). There are many aspects, including environmental considerations, that might be taken into account in a normative discussion of population ageing and decline which goes beyond the scope of this study. Nevertheless, younger people

are clearly regarded by local governments as an economic advantage and an asset against demographic ageing.

The question, therefore, is as follows: Can the migration component of regional population change be predicted at all? Or are the possibilities of cities and regions for anticipating future demographic developments limited, given the recent changes in trends?

2 Regional migration in Germany: recent trends

Before turning to predictions, it is first instructive to look at descriptive trends in residential mobility and migration in Germany in recent years. These trends constitute the explanandum and the variation that can be captured by the subsequent models. Figures 2 and 3 show the geographic variation in net migration into districts between 2011 and 2015. Note that the numbers include both international migrants as well as within-country mobility.

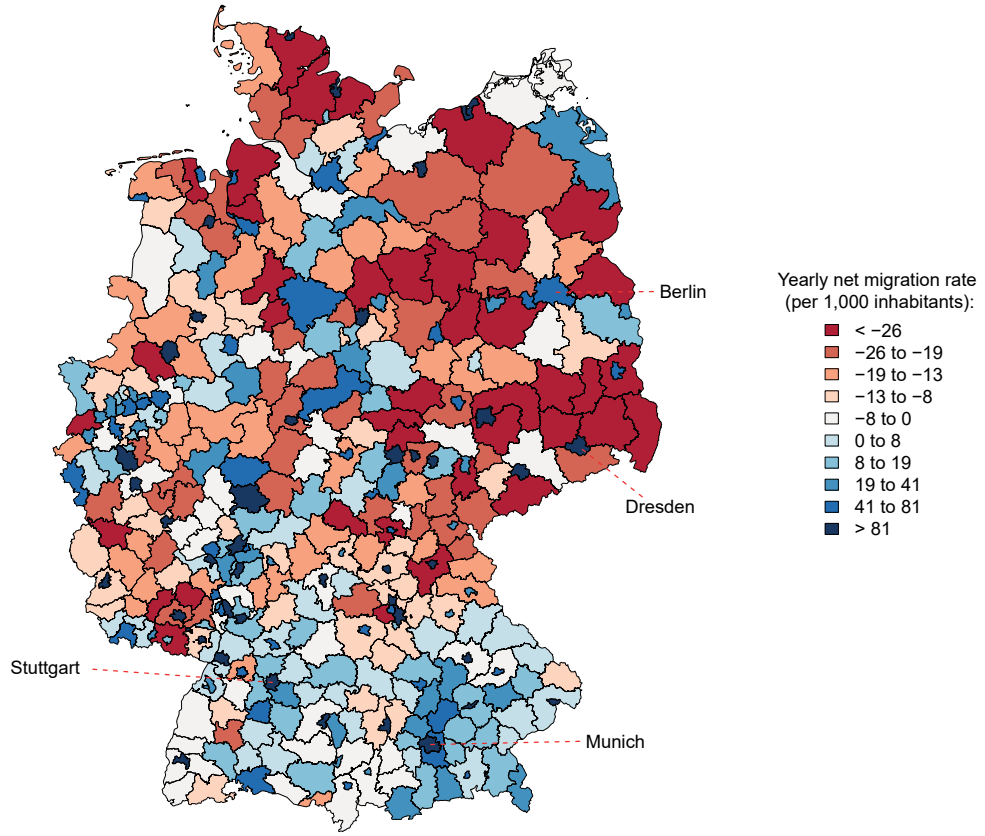
As Figure 2 shows, young people aged 18 to 24 years display a clear pattern of leaving rural areas and moving into big cities. There are a few notable differences across regions, though. The surrounding areas of Berlin or Dresden see many of their young inhabitants leave for the cities. In Munich or Stuttgart, by contrast, the surrounding districts are gaining young inhabitants as well.

The geographic pattern for families with underage children (Fig. 3) stands in marked contrast to the mobility behaviour of young adults. Looking at middle-aged people and children, we see big cities mostly losing these inhabitants while the suburban areas, especially around Berlin and Munich, show a massive increase.

It is also important to look at the temporal stability of these trends. Figure 4 shows changes in migration rates among young adults as well as families in the 13 largest German cities with a population size of at least 500,000. As the graph shows, the changes in the young-age (18 to 24 years) migration rates were generally small between 2005 and 2015. Even as early as the period between 2005 and 2009, all cities were gaining many young adults on balance (see x-axis), and this pattern remained stable up to 2015. Leipzig is an exception. Here, the net migration rate of young adults doubled, making the East German city the most popular destination (per capita) for young people between 2011 and 2015, followed by nearby Dresden. Munich, previously the most popular city among 18 to 24 age group, is the only major city where the migration rate decreased slightly over time, presumably due to housing prices that are unrivalled in all of Germany. Nevertheless, the greater Munich area attracts large numbers of young people who move there from other parts of the country in order to work or study there.

If, by contrast, we look at changes in the vertical dimension in Figure 4, the picture looks drastically different. During the period between 2005 and 2009, all 13 large German cities were characterised by a negative net migration rate among the age groups 30 to 49 and 0 to 17 years. Families had, on average, left these cities for the surrounding districts. Just a few years later, however, this trend had reversed and eight out of 13 cities are now net receivers of families.

Fig. 2: Net migration rate, age group 18 to 24 years, in German districts between 2011 and 2015

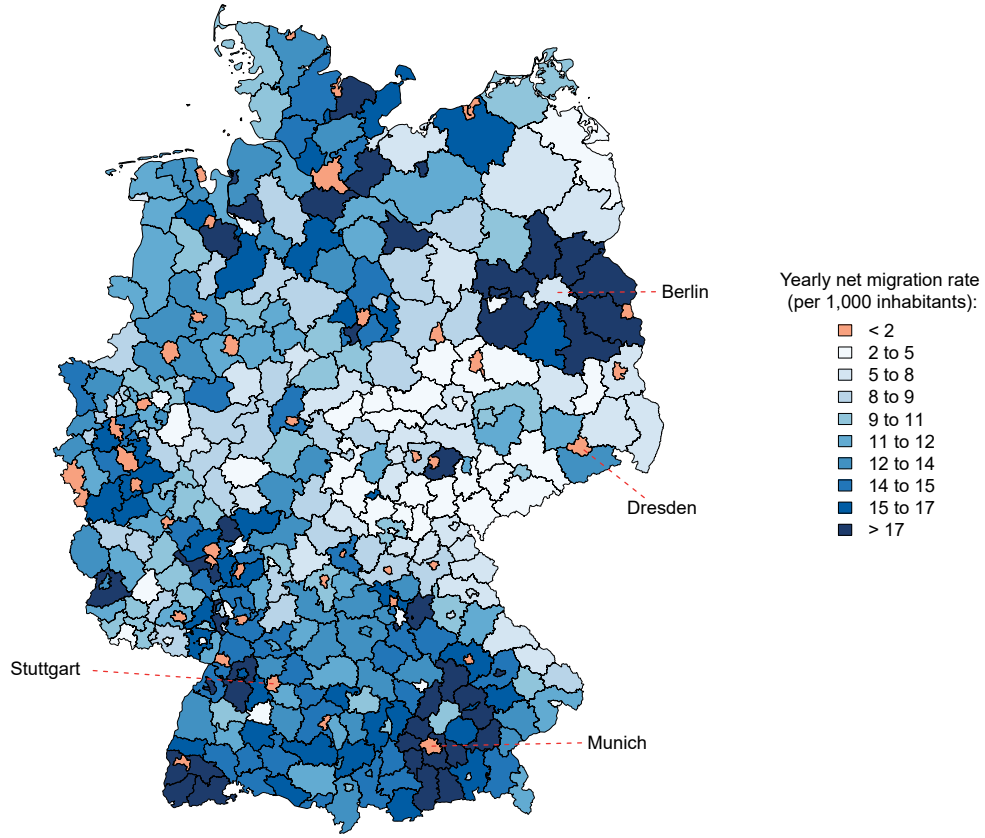


Source: own calculations, *Bertelsmann Stiftung* (2018)

Figure 5 shows the same analysis as Figure 4, but in this case the data are limited to German nationals in the respective age groups. As indicated by the graph, there is a massive decrease in net migration among German families in the largest German cities. All cities had already lost German nationals in the age groups 30 to 49 and 0 to 17 years during the period between 2005 and 2009, but the balance became markedly more negative by 2011-2015. This means that the positive total net migration in this demographic group shown in Figure 4 can be attributed entirely to foreign nationals who moved into the big cities in much larger numbers in recent years.

This is an important insight, showing that an average trend (the increase in family migration rates) can mask two contradicting trends that partially cancel each other out. On the one hand, natives are leaving the big cities in ever greater numbers whereas immigrants more than compensate for this loss. On the other hand, young immigrants and Germans in the 18 to 24-year age group are moving into the cities, and without immigrants the net migration rates would not look drastically different.

Fig. 3: Net migration rate, age group 30 to 49 years and 0 to 17 years, in German districts between 2011 and 2015

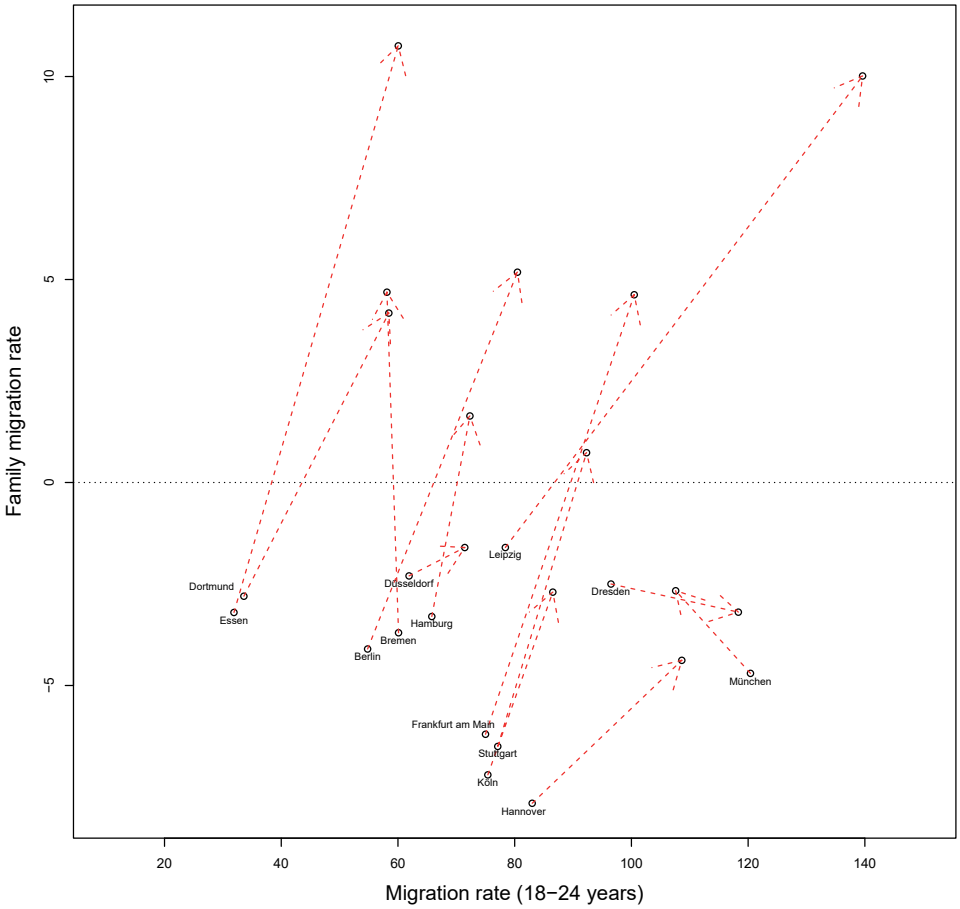


Source: own calculations, *Bertelsmann Stiftung* (2018)

3 Regional migration and residential mobility: theoretical considerations

In general, two strands of literature provide theoretical accounts that seek to explain why people relocate and which places are more attractive to movers than others. Macro-level accounts on the determinants of migration focus on the characteristics of countries or regions that attract or repel people (e.g. *Massey et al.* 1993). The terminology of "push and pull factors" (*Lee* 1966) has long transcended the scientific literature and infiltrated into public discourse. Among the founding documents for macro-level migration research are *Ravenstein's* (1885) classic inquiries into the determinants of movements between cities. His "laws" stated, for instance, that long-distance migrants tend to target large cities with economic opportunities over rural areas. *Zipf* (1946) formulated the well-known hypothesis that the volume of migration between two places is proportional to the product of their population size

Fig. 4: Net migration rate, age group 18 to 24 years (x-axis) and age groups 30 to 49 years and 0 to 17 years (y-axis), in the 13 largest German cities 2005-2009 (start of arrow) vs. 2011-2015 (arrowhead)

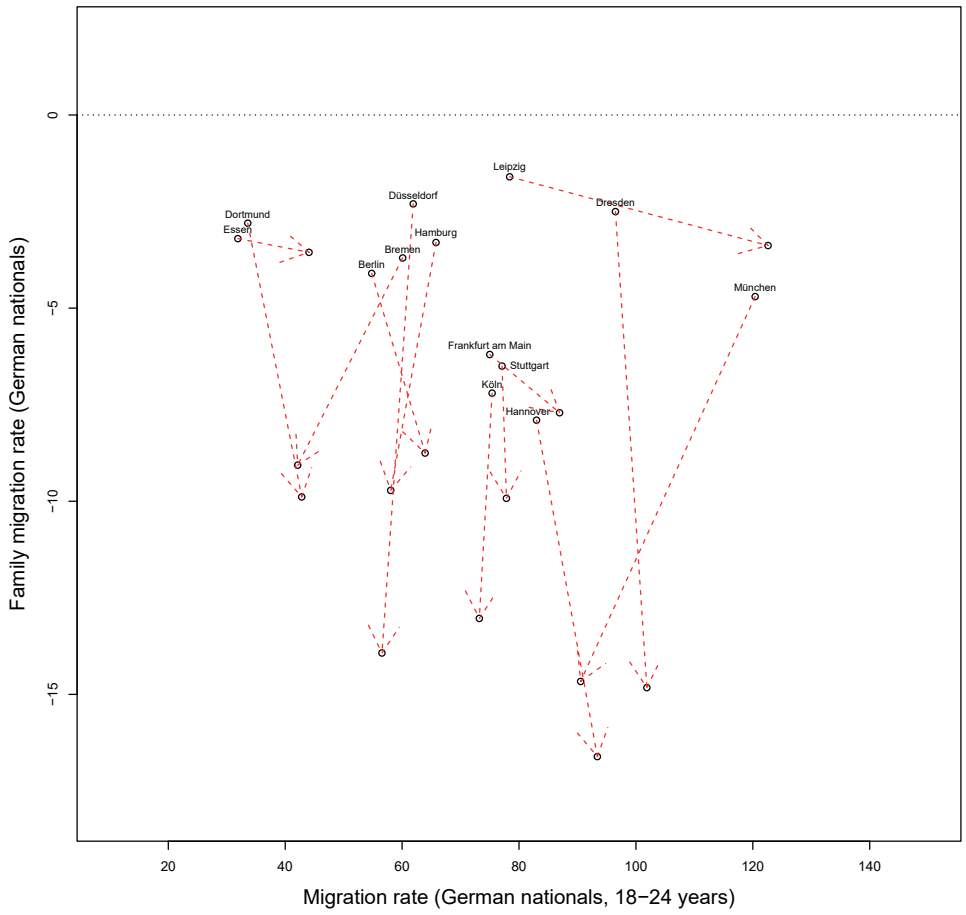


Source: own calculations, *Bertelsmann Stiftung* (2018)

and inversely proportional to the distance between the two places. These simple models have inspired more recent research that relates global migratory movement mostly to geographic and demographic characteristics of origin and destination countries (e.g. *Cohen et al.* 2008; *Kim/Cohen* 2010).

Besides these "gravity" variables (population size and distance), macro-level research into migration most often relates observed migration rates to economic prosperity and growth as well as to the labour market (e.g. *Lowry* 1966). In general, a problem for macro-level migration theory based on human capital or gravity models is the recurrent finding that the number of people that do actually migrate is small, despite huge wage differentials and other supposed push- and pull-factors (*Speare* 1974). On the one hand, the effect of economic variables such as unemployment is not straightforward and sometimes found to be zero or negative (*Hatton/Tani* 2005;

Fig. 5: Net migration rate among German nationals, age group 18 to 24 years (x-axis) and age groups 30 to 49 years and 0 to 17 years (y-axis), in the 13 largest German cities 2005-2009 (start of arrow) vs. 2011-2015 (arrowhead)



Source: own calculations, *Bertelsmann Stiftung* (2018)

Windzio 2004). On the other hand, non-economic variables are also of importance, such as the reluctance to move to culturally different regions even within a country (e.g. into regions with different dialects) (*Bauernschuster et al.* 2014).

Micro-level theories seek to explain migratory behaviour on the basis of individual characteristics of (potential) migrants. The micro-founded neoclassical model assumes that people migrate to maximise their well-being and, accordingly, migration occurs if the expected utility outweighs the costs of moving (*Sjaastad* 1962). Among the individual-level theories, there are also accounts from life course research which explain migratory behaviour through psychological models of the decision-making process (e.g. *Kley* 2011). Research on residential mobility often fo-

cuses on individual characteristics such as age, education, employment status, or psychological factors such as affinity to risk (*Greenwood* 1985).

Of course, the need to combine micro-level and macro-level models has long been voiced (e.g. *Cadwallader* 1989). In a combined model, "objective" characteristics of places are perceived by individuals as subjective variables, which the individuals combine to estimate utility scores of different places (*Cadwallader* 1989). If the subjective attractiveness of a potential destination is greater than the current place of residence, an individual disposition for mobility exists (considering migration) which may translate into actual migratory behaviour, conditional on constraining factors such as costs (*Kalter* 1997). As a general framework, the concept of subjective expected utility (SEU) can serve as a starting point for integrating various middle-range theories (*De Jong et al.* 1983). Individuals subjectively evaluate their origin as well as potential destination places and compare them to their own aspirations. An SEU model of residential mobility in accordance with *Cadwallader* (1989: 497) could be formulated as follows:

$$SEU_i = \sum_{j=1}^{m} w_j U_{ij} - C_i \tag{1}$$

where the subjective expected utility SEU of moving to region i is a function of the subjective utility ratings U of regional attribute j, weighted by an individual-specific importance measure w, and the costs C of moving to region i. The attributes considered by individuals when assigning values to different places might include, among others, employment opportunities, length of commute to existing job, green space, crime rates, or in case of families the perceived quality of schools.

Importantly, we assume that the weights that individuals place on these factors can change over the life course (*Kley* 2011). That is, individuals in different age groups will have different rank orders of places because the subjective importance assigned to various regional characteristics has changed. In particular, the presence or absence of children influences personal preferences regarding residential location choices and commuting behaviour (*Kim et al.* 2005). For instance, research finds that young adults are more likely to plan to leave a city if there are better career opportunities in the potential destination or better opportunities to pursue own interests relative to the current place of residence. In the family phase, however, these effects are absent and instead factors related to family life inspire the plan to move (*Kley* 2011).

Below, we discuss different preferences for residential mobility among young adults and families, focusing on four trends that we deem especially important for predicting future patterns of mobility among these two demographic groups in Germany.

Educational expansion

Between 2000 and 2015, the share of school leavers attending university increased drastically, from around 30 percent to more than 50 percent. This means that while two decades ago, the average school leaver started vocational training as part of the traditional dual system in Germany, today the majority of school leavers begin some form of tertiary education. This has a considerable impact on the residential mobility of young people. Starting around the year 2000, the large inflow of young people aged 18 to 24 has caused Germany's big cities to grow once again, counteracting the previous decades of de-urbanisation (*Gans* 2013). Whereas opportunities for vocational training in crafts and trades are available in equal measure in urban as well as rural regions, tertiary education is mostly confined to cities. An ever increasing share of first-year university students among young people implies that secondary school leavers will leave rural regions situated a long way from universities in even greater numbers.

This process also affects people's mobility decisions in later stages of the life course. Once a university degree has been obtained, the likelihood of moving back to the countryside diminishes since there are fewer employment opportunities for job-seekers with tertiary education. After children have been born, tertiary-educated residents of large metropolitan areas might still express a desire to leave the city, but this is constrained by employment opportunities that are largely restricted to urban areas. As a consequence, young families living in big cities are likely to move into suburban municipalities close to these cities, but less likely to go back to the peripheral rural regions where they may have come from (*Milbert*/*Sturm* 2016).

Socio-economic and ethnic segregation

Another feature that consistently predicts residential mobility of family-age people in particular is the socio-economic or ethnic composition of an area (e.g. *South*/*Crowder* 1997; *van Ham*/*Feijten* 2008). In Germany, there is a close correlation at the neighbourhood level between the share of migrants and lower socio-economic status, but at the more aggregate regional level, the opposite is true (*Weber* 2015b). Thus, prosperous and ethnically diverse metropolitan areas are presumably attracting families, while at the level of municipalities, the concentration of ethnic minorities is likely to lead to greater out-migration.

Residential segregation by ethnic background has historically been rather low in Germany, as compared with other countries such as France, the United Kingdom, or the United States (e.g. *Musterd* 2005). However, there are also signs that ethnic and socio-economic segregation have been on the rise in Germany in recent years. The residential segregation of welfare recipients, for instance, has increased in 80 percent of all German cities over the past few years (*Helbig*/*Jähnen* 2018). Half of all cities now have neighbourhoods in which more than 50 percent of all children are dependent on social security transfers (*Helbig*/*Jähnen* 2018). In addition, it is well known that residential segregation is stronger among recent migrants than among migrants from previous waves who tend to move into more mixed neighbourhoods

over time (e.g. *Friedrichs* 1998). As such, the recent sharp increase in immigration to Germany is likely to lead to an increase in the level of segregation, at least in the short term.

In line with *Schelling's* (1971) well-known theory of segregation, one could assume that municipalities with a growing proportion of migrants or socio-economically deprived people will, at some point, start witnessing an increase in emigration rates among the native population even though there is no historic precedent for this. It is likely that there will be a tipping point at which an accelerated level of outmigration by middle-class members of the municipality is triggered. To a certain degree, this can already be observed in big German cities with a negative correlation between the concentration of ethnic minorities in a neighbourhood and the net migration rate of natives (*Weber* 2015b). A high concentration of immigrants is, for example, often associated with the fear of a decrease in the educational quality of schools due to language problems and cultural misunderstandings (*SVR* 2016). However, ethnically mixed but prosperous areas (such as university quarters) are still popular (*Weber* 2015b).

Tight housing markets

One of the most politically salient issues in recent years has been the sharp increase in housing prices throughout most of Germany. A tight housing market has previously been linked to low residential mobility in Germany (*Clark*/*Drever* 2000). Some people might consider moving, but financial constraints impede actual resettlement. For many young adults wishing to move to the city to enrol at university or start vocational education, living in the core cities might not be affordable and they instead resort to surrounding municipalities with good public transport connections. This is what the descriptive view already suggests for Munich in particular (see Fig. 2 and 5). In the family age groups, the desire to leave the city may be amplified by the inability to find larger apartments within the city limits after the birth of a child (*Busch* 2016).

There is also obviously endogeneity with regard to the effect of housing costs. Since demand for housing as a consequence of local population growth is one of the most important determinants of intranational variation in housing prices – other factors such as the interest rate are constant for all regions – residential mobility affects housing prices at least as much as housing prices affect residential mobility. Whereas the former relationship is positive, however, the latter is negative. Rising housing prices are thus a strong indicator of population growth in the past, but can deter future inflows if the price level surpasses the acceptance limit of potential in-movers.

The housing market also interacts with residential segregation along socio-economic or ethnic lines and with geographic location. On the one hand, high rents may impede people with a low socio-economic status from moving into certain municipalities or prompt them to leave in the case of "gentrification" of previously cheaper areas. On the other hand, depending on the demographic composition of a municipality, potential in-movers may be more or less willing to pay high housing

prices. Municipalities within commuting distance of big, prosperous cities with a low proportion of socio-economically deprived residents or ethnic minorities are probably most attractive to families who want to move into the broader metropolitan area. In this case, housing prices are likely to be higher.

Increase in international migration

While the previous theoretical considerations have mostly focused on residential mobility within Germany, the descriptive view has already suggested that international migration plays a major role in determining the demographic prospects of many regions. Mobility patterns of international migrants are different to those of internal migrants (compare Fig. 6 and 7). There is abundant evidence that international migrants tend to move into big cities where other migrants already live (e.g. *Mayda* 2010). We can thus hypothesise that, in terms of the weighting placed on different attributes of a municipality, equation (1) looks different for migrants than it does for natives. In particular, we assume that migrants are willing to pay a larger share of their income to housing if it allows them to live close to kinship and fellow countrymen in the big cities. For recent refugees in particular, it should also be noted that their location choices (if they do have a choice) are less sensitive to the housing market.

In sum, the following assumptions can be made about the determinants of residential mobility among young people and families:

- Young adults' (18 to 24 years) propensity to move to large cities with universities will increase over the study period due to the continued expansion of tertiary education.
- In the most prosperous cities where housing prices are highest, the surrounding areas with access to good public transport connections will also attract young people, and fewer school leavers originating from these areas will move out of their parental home into the city due to the higher living costs.
- For rural areas farther away from large cities, a crucial aspect in attracting young people or preventing them from leaving will be whether universities (of applied sciences) are within reach. The labour market and opportunities for vocational education still play a vital role, but tertiary education is gaining in importance over time.
- Families who value access to green areas, larger living spaces, and schools with a low share of socio-economically deprived children are more likely to move to suburban or rural areas with low unemployment and high income levels compared with large cities.
- However, since an increasing share of young-to-middle-aged mothers and fathers has undergone tertiary education, these suburban or rural areas need to be in reach of large cities with a diversified economy. Thus, areas that are within short distances of large cities, have access to public transport, and only a small number of residents with a low socio-economic status, will likely lead to high inflows of family-age people.

- Rising housing costs and an increasing concentration in cities of residents with a low socio-economic status will accelerate this trend of families moving towards suburban and rural areas.
- Places with a high share of immigrants can be expected to attract more immigrants in the future – the volume of these flows, however, is highly conditional upon national and international factors such as domestic migration policies, and is hardly predictable on a local basis.

4 Building predictive models of municipal migration

Explanations versus predictions

Explanations and predictions are both fundamental to science. For *Hempel* and *Oppenheim* (1948: 138), "an explanation of a particular event is not fully adequate unless [it] […] could have served as a basis for predicting". An "explanation" in the deductive-nomological tradition is the logical derivation of statements about particular observations from a more general theoretical model – this is also sometimes termed "scientific prediction" (*Dowding/Miller* 2019). A "prediction", by contrast, in our understanding is closer to what *Dowding* and *Miller* (2019) call "pragmatic prediction", namely the assignment of values to previously unseen cases (possibly in the future), i.e. to cases that were not used to train the model that is the basis for prediction. In many disciplines, such as demography and other social sciences, applied research often uses a mixture of these approaches. In these disciplines, there are no universally applicable laws but often, instead, a set of middle-range theories and auxiliary assumptions about the causal mechanisms at work that are translated into a statistical model. There is arguably a continuum between a purely data-driven predictive model and a purely deductive-hypothetical approach, with theory-informed empirical research in the middle.

The big difference between "explanatory" and "predictive" approaches is thus not so much the underlying philosophy of science. Rather, they differ in their main goal of what should be achieved by the empirical endeavour. In an explanatory approach, as is the dominant practice in many applied statistical fields such as population studies and economics, a hypothesis is translated into a formal model and tested with inferential statistical techniques such as linear regression and null-hypothesis significance testing (in the frequentist paradigm) on the basis of empirical data. If the regression coefficient is statistically different from zero, the association found in the data is unlikely given a null hypothesis is assumed to be true. If some other conditions are met, the effect is assumed to be causal. Formally, in a standard linear regression model of the form $\hat{y} = \hat{\alpha} + \hat{\beta}x$, where values of an outcome $\hat{y}$ are estimated as a function of an explanatory variable x with parameter weight $\hat{\beta}$ (and a constant $\hat{\alpha}$), the explanatory approach is most concerned with the estimation of parameter $\hat{\beta}$ (*Mullainathan/Spiess* 2017). "Mere" predictions are often dismissed as not advancing the field by providing reliable evidence based on theory (*Keuschnigg et al.* 2018).

By contrast, a predictive model is more concerned with the outcome $\hat{y}$. The goal is to maximise predictive accuracy as defined, for instance, by the number of correctly predicted events as a proportion of all observed events in case of a dichotomous outcome. Predictive models are much more widely used in the computer science and machine learning communities, and increasingly also in applied fields in business and industry (e.g. *Kuhn/Johnson* 2013). In these fields, there is often less emphasis put on "explanation", and some even dismiss the necessity for theoretical models, causality and explanations per se (*Anderson* 2008). Indeed, in many applications such as the prediction of credit default, customer churn, or accident risks, it might suffice to have a data-driven approach with high predictive accuracy without a deeper understanding of the causal mechanisms for *why* two specific items are often bought together, or why a scanned image of a handwritten signature is more likely to belong to someone named "Smith" rather than "Miller". The coefficient of a specific independent variable $\hat{\beta}$ is less interesting compared with correct forecasts of the outcome. Some model characteristics such as hyperparameters (e.g. number of hidden units in a neural network) are not even substantially informative.

A common critique of most explanatory approaches is the lack of model validations via out-of-sample predictions (e.g. *Schrodt* 2014). Since in most cases all of the data is used to obtain the model parameters, the model is likely to be overfitted (e.g. *Cranmer/Desmarais* 2017). This means that the parameters are influenced by outliers or other idiosyncrasies of the training dataset and the model would probably fail to produce good estimates for cases that were not part of the training set. This critique is arguably most appropriate if the data analysed do not come from a random sample of a population (as is the case in public opinion research for instance, where the frequentist paradigm of null-hypothesis significance testing was developed), but are instead a non-random sample or the total population – as in, for instance: "all German municipalities with a population greater than 5,000". These shortcomings are arguably major causes for today's replication crisis in science.

Predictive models using machine learning algorithms are increasingly applied to social science research, although their use is still comparatively rare. For instance, *Bansak et al.* (2018) used machine learning algorithms to geographically place refugees in a way that their employment prospects improved significantly. In this application, the main question of interest is not whether a specific independent variable (e.g. a refugee's level of education) influences the outcome significantly. Therefore, a research design for causal identification is less well suited, compared with a predictive model. In another example, *Muchlinski et al.* (2016) find that machine learning algorithms such as random forests provide more accurate predictions of civil war onset as opposed to "traditional" regression-based approaches.

But predictive models alone are no panacea for arriving at more reliable evidence (e.g. *Athey* 2017). Correlations that exist in big databases are often spurious (*Calude/Longo* 2017), and most machine learning algorithms do not give consistent estimates of which features are more important to the outcome and which are less so (*Mullainathan/Spiess* 2017). A machine learning algorithm discarding a certain feature does not therefore constitute a falsification of a theory that claimed this feature to be important. This might be less problematic in some applications (e.g.

image recognition), but in other cases it is important to have at least a plausible idea of the underlying mechanisms at work. A combination of proxy variables that are highly correlated with the "real" causally relevant factors often produce predictive accuracies that are just as good as models informed by the true mechanisms. As a consequence, policy-related interpretations of the form: "Since X is given high importance by the algorithm, improving X will result in higher Y", are invalid. This is important to note because practitioners and policy-makers might be tempted to interpret the results in such a way.

Thus, Susan *Athey* (2019) does not expect the increasing use of machine learning techniques in disciplines such as economics to have any substantial impact on the statistical theory of causal identification. The latter will continue to be a main goal for researchers in these fields. Instead, machine learning algorithms will complement the standard econometric approaches, which might put an end to the "statistical monoculture" of linear and logistic regression that *Schrodt* (2014) complained about. In addition, practices that are widely used in machine learning such as cross-validation and out-of-sample predictions can help applied research in economics and social sciences to overcome the problem of overfitted models and purported findings that fail replication (*Cramner/Desmarais* 2017).

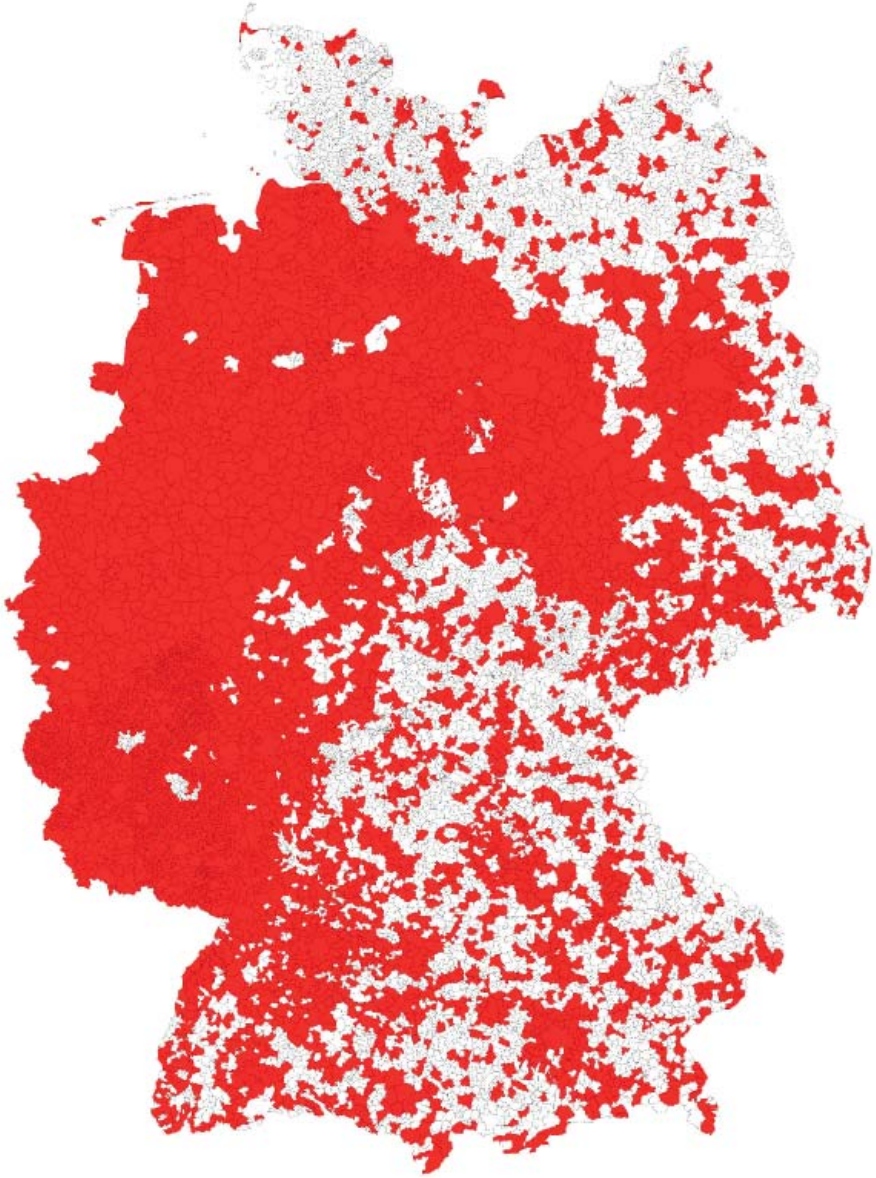
Our application is located somewhere on the continuum between a data-driven predictive model and a deductive-nomological explanatory approach. On the one hand, our goal is to make out-of-sample predictions: How well can future migration rates be predicted for municipalities that were not part of the training dataset? On the other hand, our predictive model is informed by theories and well-established knowledge about mechanisms behind migration and residential mobility. Since there is no grand theory in demography and other social sciences, the various middle-range theories referred to in the previous chapter can guide variable selection and transformation (feature engineering). As such, the evidence produced by the model will hopefully prove to be more stable compared with pragmatic, purely data-driven approaches. However, strictly testing these theories of migratory behaviour or inferring policy recommendations is beyond the scope of this approach.

Data

The main data source in this project is the "Wegweiser Kommune" (*Bertelsmann Stiftung* 2018), a database that contains various demographic and socio-economic indicators for more than 3,000 municipalities (*Kommunen*) in Germany. All German municipalities with a population greater than 5,000 are included in the database, although a small number of municipalities have dropped below this mark since the first reference year from which the data are available (2008). The largest municipality is the city of Berlin with a population of more than 3.3 million, whereas the average population size of a municipality in the dataset is around 24,000.

We thus have a non-random sample of municipalities that also shows systematic geographical patterns (see Fig. 6): Especially in rural areas in the states of Bavaria, Schleswig-Holstein, and Mecklenburg-Western Pomerania, large parts of the map consist of smaller municipalities with populations below 5,000. In these states, we

Fig. 6: Map showing German municipalities that are included (in red) in the dataset (grey areas = not included)



Source: *Bertelsmann Stiftung* 2018

lack data on many rural areas. In the states of North Rhine-Westphalia or Saxony-Anhalt, by contrast, local government reforms resulted in fewer, but larger municipalities. Finally, in Lower Saxony and Rhineland-Palatinate, there are many associations consisting of multiple smaller municipalities (*Samtgemeinden*) which are treated as a single municipality in the analyses. In the latter states, there are thus fewer areas

for which we lack data. All in all, considering that many smaller municipalities are represented as unions of municipalities in the dataset, 5,850 or 51.3 percent of all German municipalities are included in the analysis. However, the total sample size is 3,151 since associations of municipalities are treated as a single observation.

The outcome variable of interest in this study, namely net migration, is reported in the dataset as the average yearly value for the five years leading up to the reporting year. In other words, for the 2015 edition, net migration refers to the average net migration value (per 1,000 inhabitants) during the period from 2011 to 2015. This is our target variable which is reported for a variety of age groups. We focus on the two pre-defined demographic groups that the authors (*Bertelsmann Stiftung* 2018) refer to as "educational migration" (age group 18 to 24 years) and "family migration" (age groups 30 to 49 and 0 to 17).

Among the many indicators (N = 184) in the dataset, we excluded a number of variables that were either irrelevant to our question or too close to the definition of the dependent variable (e.g. total net migration, total population change). In addition, we excluded many indicators that were highly correlated ($r > .95$) with others; for instance, total balance of commuters out of a municipality vs. balance of male commuters vs. balance of female commuters, etc. We also excluded indicators that had more than 50 percent missing values, which is often the case for indicators that are only available at the district level or for larger agglomerations (e.g. share of children with a migrant background in kindergartens). This leaves us with 38 indicators from the source database that are potentially considered for inclusion in the models (see Appendix). For each algorithm, a recursive feature elimination procedure is built into the modelling pipeline in such a way that the model is trained with the optimal subset of features determined via cross-validation. This means that the algorithm may use all of the variables listed in the Appendix or only a subset if this results in greater predictive accuracy in the training data.

Adding to the indicators from the Bertelsmann dataset, we assembled a number of other variables and merged them with the dataset. In particular, these were:

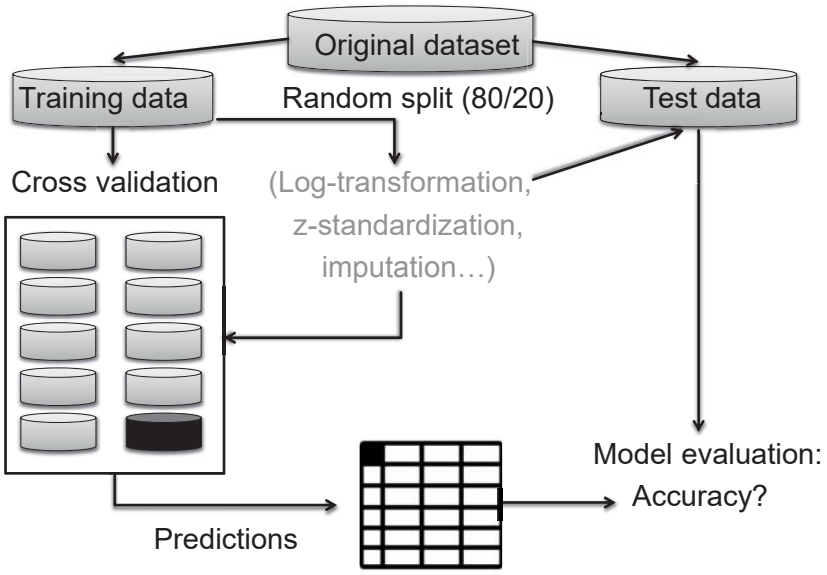
- Information on all long-distance and regional train stops in Germany obtained from the German railway company Deutsche Bahn. For each municipality, we coded whether a train station is present and whether long-distance trains also serve this station. As an additional variable, we calculated the distance to the nearest municipality with a station served by long-distance trains. These indicators are included because theoretical considerations suggested that good public transport connections are vital in attracting young people to municipalities.
- We calculated the distance from each municipality to the nearest city with more than 50,000, more than 100,000, or more than 500,000 inhabitants. The idea is that short distances to cities, especially if combined with other features such as good public transport links, is important for predicting net migration (see theory section).
- We coded whether a municipality had a university, how many students were enrolled at this university, as well as the distance to the nearest municipality with a university. We also counted the total number of university students enrolled within a radius of 20 km around each municipality.

- A few indicators from the original dataset were transformed; e.g. we calculated the change in the share of foreigners over the past two years.

Models of municipal migration

The basic idea is to train models based on past migration rates between the years 2005 and 2009 and to use them to predict migration in the period between 2011 and 2015 in previously unseen municipalities. The dataset consisting of 3,151 municipalities in Germany is divided into a training set (N = 2,543, i.e. 80 percent of the data) and a test set (N = 608, i.e. 20 percent) via a random split. Figure 7 depicts this procedure schematically. Variables with skewed distributions are log-transformed and non-binary indicators are z-normalised, resulting in mean = 0 and standard deviation = 1. This facilitates finding numerical solutions with many machine learning algorithms. Finally, an imputation routine based on a k-nearest-neighbour algorithm is applied to missing values. All of these preprocessing steps are performed with the *preProcess()* function included in R's (*R Core Team* 2018) caret package (*Kuhn* 2018). Importantly, these transformations are applied to the test data using only information (such as mean and standard deviation) from the training data.

Fig. 7: Schematic depiction of the procedure to partition the dataset and arrive at out-of-sample predictions



Source: own design

Next, 10-fold cross validation is performed on the training set. The data are divided into 10 sets where nine serve as training sets and the remaining one as a validation set against which parameters are tuned. The procedure is repeated with

each of the ten sets serving as the validation set, with the result that the entire dataset was used for both training and validation. The model with the best predictive performance on the training set is selected to produce predictions for the test set. As a predictive performance measure, we generally select R^2 since our outcome variable is continuous and it offers the most intuitive interpretation. As robustness tests, all predictions were also judged by their root mean squared error (RMSE), but the overall interpretation of the results did not change.

We consider four popular algorithms: (1) Linear regression as the baseline; (2) a random forest; (3) an extreme gradient boosted tree; and (4) a deep neural network. These are four popular algorithms in applied research, with the latter two in particular often listed as (part of) the winning submissions in machine learning competitions (*Chollet/Allaire* 2018). There are, of course, many other algorithms that are frequently applied to regression problems, but for the sake of simplicity we confine the analysis to these four methods. Tree-based methods are more flexible compared with regression models in terms of handling non-linear relationships and interactions. In the case of multi-collinearity, however, the result of a tree-based algorithm is likely to be unstable since many of the variables could have served as substitutes for others. Random subspace methods used in random forests or gradient-boosted trees can improve this issue. Instead of one tree, many trees are grown using only a random subsample of all features (variables) as well as observations. The final prediction comes from an ensemble of all trees where each tree votes which value to assign to a given observation. Peculiarities of the training dataset such as outliers will have less of an impact on the final model used for predictions.

Gradient boosting machines are also ensemble methods (of trees, in this case), but progress sequentially. After a tree is built, greater weights are assigned to badly predicted cases and the next tree tries to optimise these cases. Ensemble methods usually outperform single learners (such as a decision tree or a linear regression); on the other hand, they cannot easily be interpreted by humans (e.g. *Zhou* 2012). For random forests and (extreme) gradient boosted trees, we use *caret*'s wrapper function *train()* which calls algorithms from the *randomForest* and *xgboost* packages (*Liaw/Wiener* 2002; *Chen et al.* 2019). The model-building pipeline is set up to tune the hyperparameter *mtry*, i.e. the number of variables sampled as potential candidates at each tree node split. A few outliers with z-scores greater than 4 on the dependent variables were dropped since it is likely that these represented either measurement errors or extremely unusual changes, such as the installation of a state-wide reception centre for refugees in a small municipality.

To build neural networks, we use the *keras* package for R with a TensorFlow backend (*Allaire/Chollet* 2018). Deep neural networks are non-linear regression models with stacks of unobserved variables (hidden layers). The units in the hidden layers are a linear combination of the input variables, transformed by a non-linear activation function. We specify a convolutional neural network with two hidden layers with varying numbers of units and dropout layers with different weights for hyperparameter tuning. Deep neural networks have become widely used in recent years due to their ability to learn any kind of function, almost supplanting the need for feature engineering (*Chollet/Allaire* 2018). The fact that functional forms, feature

importance, and interactions between variables remain inside the black box is unproblematic for many applications such as natural language processing or image recognition. In our application, it might be tempting to seek out mechanistic explanations – why is this happening and what can we do against it – but this is not what the method can provide.

5 Results

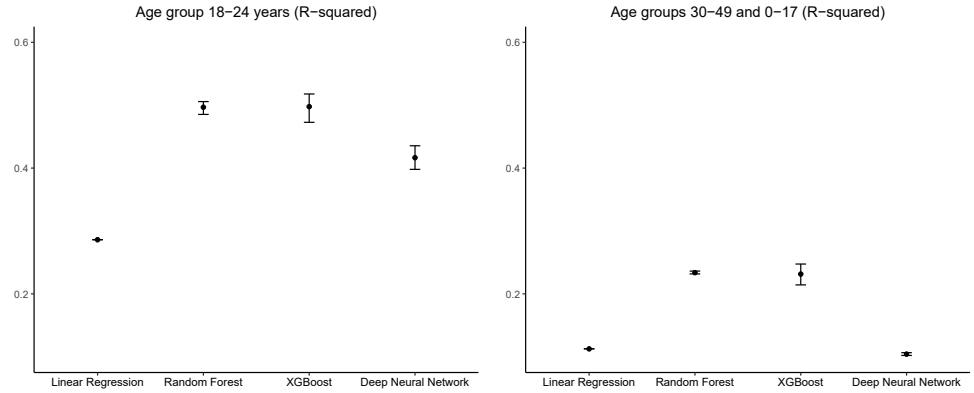
Figure 8 shows the predictive performance of four different algorithms. The models are trained on migration data between 2005 and 2009 and predict migration rates in the period between 2011 and 2015 in 608 previously unseen municipalities. The results show that migration rates among young people (18 to 24 years) can generally be predicted quite well (left-hand panel in Fig. 8). The best performing model (gradient boosted tree) achieves R^2 values of more than 50 percent, equivalent to correlations of $r > 0.7$ between the predictions and the actual observed values. The variation around the average predictive performance (marked with a dot) comes from different combinations of hyperparameters for the models other than linear regression. The well-known bias-variance tradeoff becomes apparent, with linear regression showing no variation in the forecasts but the results are furthest from the actual observations, compared with the other models.

The "family" demographic group, by contrast, is less predictable in its migratory behaviour. $R^2 = .25$ is the best value that could be achieved with an xgboost regressor. This reflects the fact that patterns of family migration appeared to be less stable over time and had changed considerably by 2011-2015 compared with the mid-2000s (see Fig. 4 and 5). Sharp increases in immigration from other countries turned net migration rates in many cities from negative to positive, at the same time as native German families left the cities in ever greater numbers, resulting in higher inflows into suburban and peripheral regions. These trends were "unpredictable" in the sense that, when looking at data from 2005 to2009, an accurate prediction of which cities and rural areas would gain many families on balance and which would lose inhabitants in these age groups would not have been possible.

By contrast, the residential mobility of young people (18 to 24 years) was comparatively stable over time and is less affected by international migration. Yet, R^2 values of around 50 percent mean that there is still variation that is left unaccounted for by the models. Arguably, the trend towards ever higher participation rates in tertiary education that accelerated during the study period is a crucial factor in this regard. This pushes young people to municipalities with universities in even greater numbers than before. At the same time, skyrocketing housing prices in large university cities contribute to a suburbanisation even among this young demographic group, but only in certain regions such as the greater Munich, Stuttgart or Frankfurt areas (see Fig. 2).

Figures 9 and 10 plot the predicted values against the actual observations for the standardised migration rates of young people (Fig. 9) and families (Fig. 10) from the two xgboost algorithms. Again, it is obvious that the predicted values for young-

Fig. 8: Predictive performance (R^2) of four algorithms in predicting migration among people aged 18 to 24 years (left-hand panel) and family-age groups (30 to 49 and 0 to 17 years) (right-hand panel). Variation around coefficients is caused by different hyperparameter values



Source: own calculations

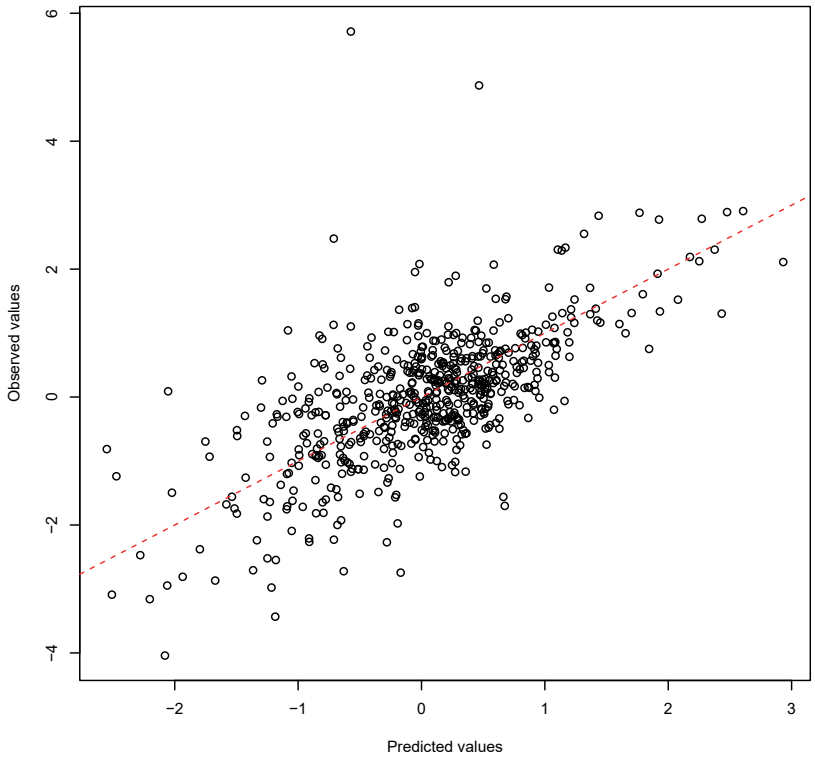
age migration anticipate the actual observations quite well (r = 0.7 between predicted and observed values). With regard to family migration (Fig. 12), this generally proves to be more difficult (r = 0.5), mirroring what we learned from the descriptive analyses.

In order to better understand which features prove useful in predicting migration rates (without making causal claims), Figures 11 and 12 plot decision trees for young-age and family migration, respectively. Note that these are exemplary trees and do not necessarily reflect what the majority of trees in the random forest, for instance, yield as results. But they intuitively show how the algorithm arrives at its results. Here we can also note the trade-off between better interpretability (as provided by the decision tree) and greater predictive accuracy by a "black box" ensemble model (e.g. as a random forest). For the sake of simplicity, tree depth is limited to three. At the leaf nodes, the boxplots show the distribution of the dependent variable in the respective group, the mean value being the prediction made by the model.

For the 18 to 24 years age group, the share of foreigners in a municipality is being used as the first split to separate the data. The second split refers to the share of sealed (urban) land. Municipalities that have a below-average share of foreigners and a very low share of sealed land are characterised by the lowest migration rate, i.e. a large out-migration of young people. The municipalities that attract the largest number of people aged 18 to 24 years are those with an above-average share of foreigners, a very low share of one or two-family homes, and a very high number of university students.

The exemplary regression tree in Figure 11 shows several of the characteristics that make tree-based models in general more flexible compared with linear regression. First, non-linear relationships can easily be accounted for by the model. For instance, while the share of foreigners has a considerable impact when splitting the data near the average value (first split), splitting the data has less of an effect

Fig. 9: Predicted and observed values for the migration rate among people aged 18 to 24 years in 608 municipalities in the unseen test data, 2011–2015 (*z*-standardised)

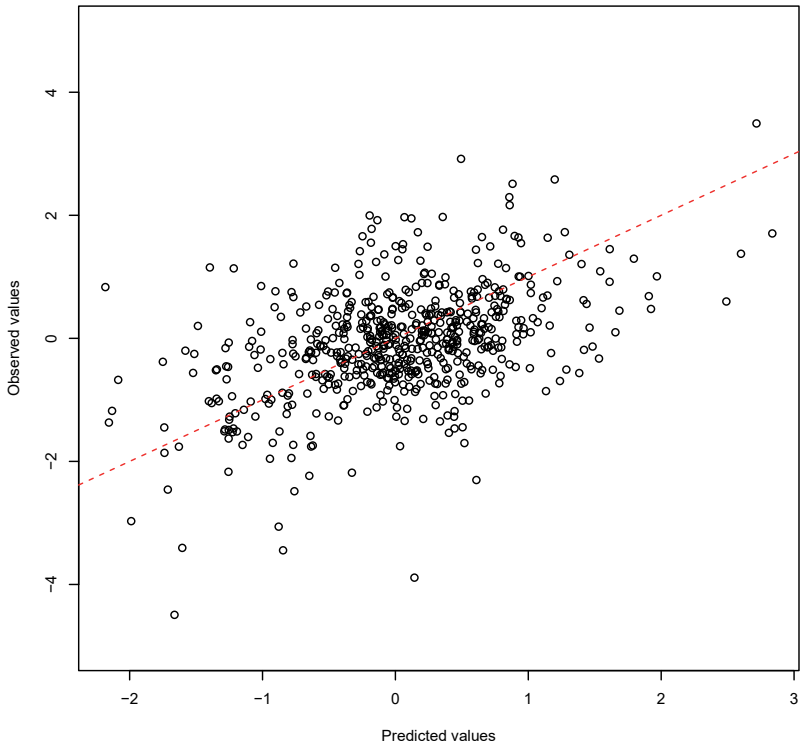


Source: own calculations

at higher values of the predictor (compare nodes 12 and 13). In addition, complex interaction effects are detected by the tree. For instance, as noted above, a greater share of foreigners is generally associated with more in-migration by young people. However, when a high share of foreigners coincides with a rural setting (many one or two-family homes, low share of sealed land, see node 20), these rural municipalities with many foreigners actually attract slightly fewer young people compared with urban and rich municipalities with few foreigners (node 8).

Again, the position of variables and their importance as indicated by the tree cannot be interpreted in a causal way. For instance, the share of foreigners is the most important variable according to the algorithm and is highly correlated with many other variables, in particular those that are characteristics of urban areas. While it is obvious that urban areas with more foreigners show higher inflows of young people, the model provides no evidence for a causal or policy-oriented interpretation of the sort: "Increasing the share of foreigners in a municipality will, everything else being equal, lead to more young people settling there in the future". The only theoretical mechanism that points in this direction is the migrant networks mechanism

Fig. 10: Predicted and observed values for the family-age (30 to 49 and underage children) migration rate in 608 municipalities in the unseen test data, 2011–2015 (z-standardised)

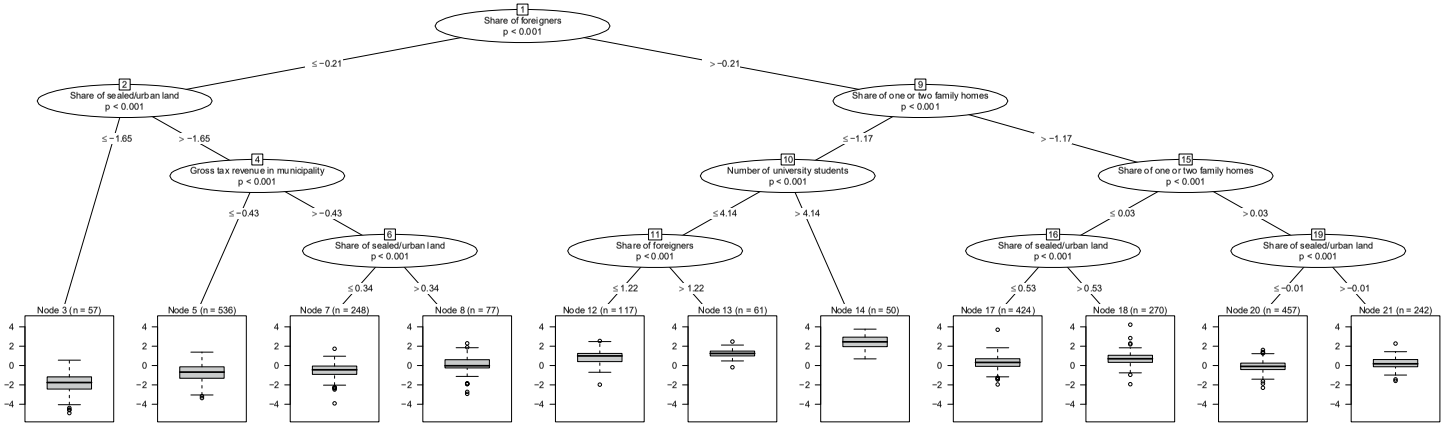


Source: own calculations

which states that international migrants tend to move to places where many of their compatriots already live. However, for larger cities at least (compare Fig. 4 and 5), international migration is not the most decisive factor in this age group.

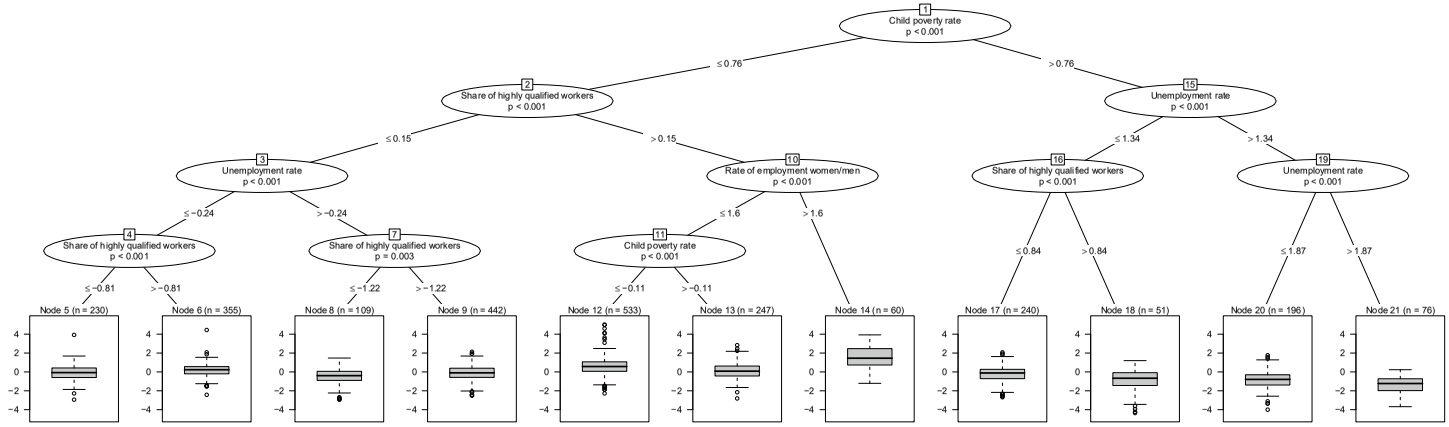
Figure 12 plots a decision tree with depth = 3 for migration rates among people aged 30 to 49 and underage children. The algorithm selects child poverty as the most important feature to divide the dataset. Families leave areas with a high child poverty rate and high unemployment. A low rate of child poverty, many highly qualified workers, and a favourable rate of employment for women in comparison with men are factors associated with a large inflow of families. The lower overall predictive accuracy of the models is reflected by the fact that many of the leaf nodes in Figure 12 do not differ much in their average migration rate.

Fig. 11: Exemplary decision tree for the migration rate among people aged 18 to 24 years



Source: own calculations

Fig. 12: Exemplary decision tree for the family-age (30 to 49 and 0 to 17 years) migration rate



Source: own calculations

6 Conclusions

How useful are regional demographic forecasts? Recent trends in local population growth have disproved many predictions which assumed that most German districts would experience population decline rather than growth. Just a few years ago, given the demographic outlook, public housing projects were abandoned, schools were closed and fewer teachers were hired in many regions. Now, by contrast, as a consequence of considerable population growth throughout most of Germany, there is a severe shortage of housing, too few kindergarten teachers and nurses, and overcrowded public transport, roads, and hospitals. It is therefore obvious that short and medium-term demographic forecasts, if reliable, would be a useful tool for local policy-makers.

How well, then, can municipal migration and residential mobility be predicted five to ten years into the future? In general, this proves to be very difficult, even with sophisticated methods and detailed data at the municipal level. The main reason for this is the volatility of international migration over time. Whereas in the late 2000s, net migration to Germany was close to zero or even negative, an all-time high was recorded just a few years later. Statisticians and qualitative experts failed to predict this trend – and at the sub-national level such predictions become even more difficult. While immigration has decreased in recent years since 2015, it is still at a much higher level than in previous decades. Similarly, while the inflow of refugees is currently much lower than it was in 2015 and 2016, the figure of approximately 180,000 per year in 2018 still represents more than a fivefold increase on rates seen a decade ago. For small municipalities in particular, the uncertainties of how many more migrants will arrive in the next few years, how many will be allowed to stay permanently, and how many will potentially relocate to bigger cities, have a strong influence on the local demographic outlook.

However, not everything is uncertain. The breakdown into two different age groups reveals that some trends are more stable and predictable than others. In particular, the residential mobility of young people aged 18 to 24 years can be predicted quite well five to ten years into the future. The general pattern of these people leaving rural areas and settling in big cities with universities prevails. It is likely that the continued expansion of tertiary education will reinforce the trend of young people leaving remote areas without universities and moving to larger cities. Since more remote areas also provide limited employment opportunities for graduates of disciplines such as cultural studies or media science, these moves will often be permanent. There are also a few mechanisms that have changed, though. The allocation of young refugees to rural areas partly reversed the negative migration rate in some regions. A shortage of housing and skyrocketing rent prices result in young people settling in the suburban surroundings of large industrial areas such as Munich or Stuttgart. As a consequence, Munich is no longer the most sought-after city among young people and is the only large city (together with Bremen) where inflows of young people have decreased somewhat.

What can be done about this? While our statistical models do not allow for causal and policy-oriented interpretations, there are a number of measures that are likely

to affect young-age migration. Given the trend towards an ever greater share of secondary school leavers going on to university, setting up new universities (of applied sciences) in rural areas is likely to prevent some of the young people from leaving the region. These new established universities can specialise in disciplines that are linked to the local labour market. At the more aggregated state and national level, policy-makers might want to reconsider whether the extensive educational expansion of recent years should be continued at the same pace in future. Today, labour shortages are occurring in many jobs that require vocational training, such as kindergarten teachers, nurses, and craftspeople. In rural areas, these jobs dominate over occupations that typically require tertiary education.

A very different set of conclusions can be drawn for residential mobility among families consisting of middle-aged people (30 to 49 years) and underage children. Predicting the migratory behaviour of this group into the future is much more difficult than that for the young age group. In the 2000s, families tended to leave big cities and settle in the surrounding suburbs or in more rural areas. This is still true for native Germans; in fact, this trend has accelerated among this group and the 13 largest German cities witnessed a surge in emigration of German families between 2011 and 2015. However, immigrant families have moved into the large cities in ever greater numbers, reversing the overall trend for residential mobility in the family age group. The majority of the largest German cities are now actually attracting more families on balance than are leaving. In addition, native families leaving the cities contribute to high population growth in the surrounding districts of large cities such as Berlin or Munich, but even municipalities situated further away from the city are now increasingly becoming net receivers of family migration.

There are certainly a few limitations to this study, e.g. regarding the availability of data at the level of municipalities. If we were able to further disaggregate the age groups or provenance (within-country mobility vs. international migration) and analyse directed flows rather than net migration, the predictive accuracy of the models would certainly be higher with regard to some groups (e.g. mobility behaviour of native Germans aged 30 to 39). On the other hand, international migration would still be the hardest factor to predict accurately, so a forecast of total population change in broadly-defined age groups (e.g. underage children in order to forecast the demand for teachers), which is of practical interest to local policy-makers, would still achieve only moderate levels of predictive accuracy.

In sum, it is hard to predict whether the current trends will continue in the near future. Many of the patterns described above are intertwined; for instance, as more immigrants move into the city centres and housing costs rise, natives are more likely to relocate to the suburbs. However, municipalities surrounding Munich or Stuttgart are already highly-priced when it comes to housing and are also densely populated, so if families want lower housing costs, this requires ever further commutes to work in these regions. If immigration from other countries were to decline in the coming years, the populations of big cities would start to decrease on balance, since natives are already leaving these cities in considerable numbers. This might in turn make the core cities more attractive again for young people who are currently moving to the suburbs due to the lack of affordable apartments in the city

centres. Trend reversals of this sort would be hard to anticipate, especially since international migration is the big unknown variable in all demographic forecasts. However, international migration can be regulated by national and European migration policies, so to a certain extent policy-makers have the means to frame the next demographic forecasts.

References

Alders, Maarten; Keilman, Nico; Cruijsen, Harri 2007: Assumptions for long-term stochastic population forecasts in 18 European countries. In: European Journal of Population 23,1: 33-69 [doi: 10.1007/s10680-006-9104-4].

Allaire, Joseph J.; Chollet, François 2018: Keras: R Interface to 'Keras'. R package version 2.2.4 [https://CRAN.R-project.org/package=keras, 31.07.2019].

Anderson, Chris 2008: The end of theory: The data deluge makes the scientific method obsolete. In: Wired Magazine 16,7: 16-07.

Athey, Susan 2017: Beyond prediction: Using big data for policy problems. In: Science 355,6324: 483-485 [doi: 10.1126/science.aal4321].

Athey, Susan 2019: The impact of machine learning on economics. In: *Agrawal, Ajay; Gans, Joshua; Goldfarb, Avi* (Eds.): The Economics of Artificial Intelligence: An Agenda. University of Chicago Press.

Bauernschuster, Stefan et al. 2014: Why are educated and risk-loving persons more mobile across regions? In: Journal of Economic Behavior & Organization 98: 56-69 [doi: 10.1016/j.jebo.2013.12.011].

Bansak, Kirk et al. 2018: Improving refugee integration through data-driven algorithmic assignment. In: Science 359,6373: 325-329 [doi: 10.1126/science.aao4408].

BBSR (Bundesamt für Bauwesen und Raumordnung) 2008. Raumordnungsprognose 2025. BBR-Berichte KOMPAKT 2/2008[https://www.bbsr.bund.de/BBSR/DE/Veroeffentlichungen/BerichteKompakt/2008/DL_2_2008.pdf; 01.08.2019].

Bertelsmann Stiftung 2018: Wegweiser Kommune. Scientific Use File (Jg. 2009-2015) des ZEFIR und der Bertelsmann Stiftung [www.wegweiser-kommune.de, 31.07.2019].

Bujard, Martin; Scheller, Melanie 2017: Impact of Regional Factors on Cohort Fertility. New Estimations at the District Level in Germany. In: Comparative Population Studies 42: 55-88 [doi: 10.12765/CPoS-2017-07en].

Busch, Roland 2016: Inländische Wanderungen in Deutschland – wer gewinnt und wer verliert? In: Zeitschrift für Immobilienökonomie 2,2: 81-101 [doi: 10.1365/s41056-016-0012-3].

Cadwallader, Martin 1989: A conceptual framework for analysing migration behaviour in the developed world. In: Progress in Human Geography 13,4: 494-511 [doi: 10.1177/030913258901300402].

Calude, Cristian S.; Longo, Giuseppe 2017: The Deluge of Spurious Correlations in Big Data. In: Foundations of Science 22,3: 595-612 [doi: 10.1007/s10699-016-9489-4].

Chen, Tianqi et al. 2019: xgboost: Extreme Gradient Boosting. R package version 0.82.1 [https://CRAN.R-project.org/package=xgboost, 31.07.2019].

Chollet, François; Allaire, Joseph J. 2018: Deep Learning with R. Manning.

Clark, William Arthur 1991: Residential preferences and neighborhood racial segregation: A test of the Schelling segregation model. In: Demography 28,1: 1-19 [doi: 10.2307/2061333].

Clark, William Arthur; Drever, Anita I. 2000: Residential mobility in a constrained housing market: implications for ethnic populations in Germany. In: Environment and Planning A 32,5: 833-846 [doi: 10.1068/a3222].

Cohen, Joel E. et al. 2008: International migration beyond gravity: A statistical model for use in population projections. In: Proceedings of the National Academy of Sciences 105,40: 15269-15274 [doi: 10.1073/pnas.0808185105].

Cranmer, Skyler J.; Desmarais, Bruce A. 2017: What can we learn from predictive modeling? In: Political Analysis 25,2: 145-166.

De Jong, Gordon F. et al. 1983: International and internal migration decision making: a value-expectancy based analytical framework of intentions to move from a rural Philippine province. In: International Migration Review 17,3: 470-484 [doi: 10.1177/019791838301700305].

Dowding, Keith; Miller, Charles 2019: On prediction in political science. In: European Journal of Political Research 58,3: 1001-1018 [doi: 10.1111/1475-6765.12319].

Feyrer, James 2007: Demographics and productivity. In: The Review of Economics and Statistics 89,1: 100-109.

Friedrichs, Jürgen 1998: Ethnic Segregation in Cologne, Germany, 1984-94. In: Urban Studies 35,10: 1745-1763 [doi: 10.1080/0042098984132].

Fuchs, Johann et al. 2018: Stochastic Forecasting of Labor Supply and Population: An Integrated Model. In: Population Research and Policy Review 37,1: 33-58 [doi: 10.1007/s11113-017-9451-3.

Gans, Paul 2013: Reurbanisierungstypen in Deutschland: Wissensökonomie und Komponenten der Bevölkerungsentwicklung (2004-2010). In: *Fricke, Axel; Siedentop, Stefan; Zakrzewski, Philipp* (Eds.): Reurbanisierung in baden-württembergischen Stadtregionen. Arbeitsberichte der ARL 14. Hannover: 11-31.

Greenwood, Michael J. 1985: Human migration: Theory, models, and empirical studies. In: Journal of Regional Science 25,4: 521-544 [doi: 10.1111/j.1467-9787.1985.tb00321.x].

Hatton, Timothy J.; Tani, Massimiliano 2005: Immigration and inter-regional mobility in the UK, 1982-2000. In: The Economic Journal 115,507: F342-F358 [doi: 10.1111/j.1468-0297.2005.01039.x].

Helbig, Marcel; Jähnen, Stefanie 2018: Wie brüchig ist die soziale Architektur unserer Städte? Trends und Analysen der Segregation in 74 deutschen Städten. Discussion Paper P 2018-001. Berlin: Wissenschaftszentrum Berlin für Sozialforschung.

Hempel, Carl G.; Oppenheim, Paul 1948: Studies in the Logic of Explanation. In: Philosophy of Science 15,2: 135-175.

Hillmert, Steffen; Hartung, Andreas; Wessling, Katarina 2017: A decomposition of local labour-market conditions and their relevance for inequalities in transitions to vocational training. In: European Sociological Review 33,4: 534-550 [doi: 10.1093/esr/jcx057].

Kalter, Frank 1997: Wohnortwechsel in Deutschland. Wiesbaden: VS Verlag für Sozialwissenschaften [doi: 10.1007/978-3-663-11886-2].

Keilman, Nico 2008: European demographic forecasts have not become more accurate over the past 25 years. In: Population and Development Review 34,1: 137-153 [doi. 10.1111/j.1728-4457.2008.00209.x].

Keuschnigg, Marc; Lovsjö, Niclas; Hedström, Peter 2018: Analytical sociology and computational social science. In: Journal of Computational Social Science 1,1: 3-14 [doi: 10.1007/s42001-017-0006-5].

Kim, Jae Hong; Pagliara, Francesca; Preston, John 2005. The intention to move and residential location choice behaviour. In: Urban Studies 42,9: 1621-1636 [doi: 10.1080/00420980500185611].

Kim, Keuntae; Cohen, Joel E. 2010: Determinants of international migration flows to and from industrialized countries: A panel data approach beyond gravity. In: International Migration Review 44,4: 899-932 [doi: 10.1111/j.1747-7379.2010.00830.x].

Kley, Stefanie 2011: Explaining the stages of migration within a life-course framework. In: European Sociological Review 27,4: 469-486 [doi: 10.1093/esr/jcq020].

Kuhn, Max 2018: Caret: Classification and Regression Training. R package version 6.0-81 [https://CRAN.R-project.org/package=caret, 31.07.2019].

Kuhn, Max; Johnson, Kjell 2013: Applied predictive modeling. New York: Springer [doi: 10.1007/978-1-4614-6849-3].

Lee, Everett S. 1966: A theory of migration. In: Demography 3,1: 47-57 [doi: 10.2307/2060063].

Lee, Ronald; Mason, Andrew, et al. 2014: Is low fertility really a problem? Population ageing, dependency, and consumption. In: Science 346,6206: 229-234 [doi: 10.1126/science.1250542].

Liaw, Andy; Wiener, Matthew 2002: Classification and Regression by randomForest. In: R News 2,3: 18-22.

Lowry, Ira S. 1966: Migration and metropolitan growth: two analytical models. Chandler Pub. Co.

Mai, Ralf; Scharein, Manfred 2009: Effekte der Binnenmigration auf die Bevölkerungsentwicklung und Alterung in den Bundesländern. In: *Cassens, Insa; Luy, Marc; Scholz, Rembrandt* (Eds.): Die Bevölkerung in Ost-und Westdeutschland. Wiesbaden: VS Verlag für Sozialwissenschaften: 75-99 [doi: 10.1007/978-3-531-91832-7_4].

Massey, Durglas S. et al. 1993: Theories of international migration: A review and appraisal. In: Population and Development Review 19,3: 431-466 [doi: 10.2307/2938462].

Mayda, Anna M. 2010: International migration: A panel data analysis of the determinants of bilateral flows. In: Journal of Population Economics 23,4: 1249-1274 [doi: 10.1007/s00148-009-0251-x].

Milbert, Antonia; Sturm, Gabriele 2016: Binnenwanderungen in Deutschland zwischen 1975 und 2013. In: Informationen zur Raumentwicklung 2: 121-144.

Muchlinski, David et al. 2016: Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. In: Political Analysis 24,1: 87-103 [doi: 10.1093/pan/mpv024].

Mullainathan, Sendhil; Spiess, Jann 2017: Machine learning: an applied econometric approach. In: Journal of Economic Perspectives 31,2: 87-106 [doi: 10.1257/jep.31.2.87].

Musterd, Sako 2003: Segregation and integration: a contested relationship. In: Journal of Ethnic and Migration Studies 29,4: 623-641 [doi: 10.1080/1369183032000123422].

Musterd, Sako 2005: Social and ethnic segregation in Europe: Levels, causes, and effects. In: Journal of Urban Affairs 27,3: 331-348 [doi: 10.1111/j.0735-2166.2005.00239.x].

R Core Team 2018: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Austria, 2018, Version 3.4.3.

Ravenstein, Ernest George 1885: The laws of migration. In: Journal of the statistical society of London 48,2: 167-235 [doi: 10.2307/2979181].

Schelling, Thomas C. 1971: Dynamic models of segregation. In: Journal of Mathematical Sociology 1,2: 143-186 [doi: 10.1080/0022250X.1971.9989794].

Schlömer, Claus 2006: Bestimmungsfaktoren der zukünftigen räumlich-demographischen Entwicklung in Deutschland, in: *Gans, Paul; Schmitz-Veltin, Ansgar* (Hrsg.): Räumliche Konsequenzen des demographischen Wandels. Teil 6: Demographische Trends in Deutschland. Folgen für Städte und Regionen (Forschungs-und Sitzungsberichte der ARL, Band 226). Hannover: 4-16.

Schrodt, Philip A. 2014: Seven deadly sins of contemporary quantitative political analysis. In: Journal of Peace Research 51,2: 287-300 [doi: 10.1177/0022343313499597].

Sjaastad, Larry A. 1962: The costs and returns of human migration. In: Journal of Political Economy 70,5 (Part 2): 80-93.

South, Scott J.; Crowder, Kyle D. 1997: Residential mobility between cities and suburbs: Race, suburbanization, and back-to-the-city moves. In: Demography 34,4: 525-538 [doi: 10.2307/3038307].

Speare, Alden 1974: Residential satisfaction as an intervening variable in residential mobility. In: Demography 11,2: 173-188 [doi: 10.2307/2060556].

Statistisches Bundesamt 2009: Bevölkerung Deutschlands bis 2060. 12. Koordinierte Bevölkerungsvorausberechnung. Wiesbaden.

Statistisches Bundesamt 2019: Bevölkerung Deutschlands bis 2060. 14. Koordinierte Bevölkerungsvorausberechnung. Wiesbaden.

SVR (Sachverständigenrat deutscher Stiftungen für Integration und Migration) 2016: Viele Götter, ein Staat: Religiöse Vielfalt und Teilhabe im Einwanderungsland. Jahresgutachten 2016 mit Integrationsbarometer [https://www.svr-migration.de/publikationen/jahresgutachten-2016-mit-integrationsbarometer/, 01.08.2019].

Swiaczny, Frank 2015: Auswirkungen des demographischen Wandels auf die regionale Bevölkerungsdynamik in Deutschland. In: Raumforschung und Raumordnung 73,6: 407-421 [doi: 10.1007/s13147-015-0370-7].

Van Ham, Maarten; Feijten, Peteke 2008: Who wants to leave the neighbourhood? The effect of being different from the neighbourhood population on wishes to move. In: Environment and Planning A 40,5: 1151-1170 [doi: 10.1068/a39179].

Vanella, Patrizio; Deschermeier, Philipp 2018: Stochastic Forecasting Model of International Migration in Germany. In: *Kapella, Olaf; Schneider, Norbert F.; Rost, Harald* (Eds.): Familie – Bildung – Migration: Familienforschung im Spannungsfeld zwischen Wissenschaft, Politik und Praxis. Tagungsband zum 5. Europäischen Fachkongress Familienforschung. Berlin: 261-280 [doi: 10.2307/j.ctvddzpz0.22].

Weber, Enzo; Weigand, Roland 2016: Identifying macroeconomic effects of refugee migration to Germany. IAB-Discussion Paper 20. Institut für Arbeitsmarkt- und Berufsforschung bei der Bundesagentur für Arbeit: Nürnberg.

Weber, Hannes 2015a: Could immigration prevent population decline? The demographic prospects of Germany revisited. In: Comparative Population Studies 40,2: 165-190 [doi: 10.12765/CPoS-2015-05en].

Weber, Hannes 2015b: Mehr Zuwanderer, mehr Fremdenangst? Ein Überblick über den Forschungsstand und ein Erklärungsversuch aktueller Entwicklungen in Deutschland. In: Berliner Journal für Soziologie 25,4: 397-428 [doi: 10.1007/s11609-016-0300-8].

Weber, Hannes 2019: Der demographische Wandel. Mythos – Illusion – Realität. Stuttgart: Kohlhammer.

Weber, Hannes; Sciubba, Jennifer Dabbs 2019: The Effect of Population Growth on the Environment: Evidence from European Regions. In: European Journal of Population 35,2: 379-402 [doi: 10.1007/s10680-018-9486-0].

Windzio, Michael 2004: Kann der regionale Kontext zur „Arbeitslosenfalle "werden? In: Kölner Zeitschrift für Soziologie und Sozialpsychologie 56,2: 257-278 [doi: 10.1007/s11577-004-0034-z].

Zhou, Zhi-Hua 2012: Ensemble methods: Foundations and Algorithms. Boca Raton: CRC Press.

Zipf, George K. 1946: The P 1 P 2/D hypothesis: on the intercity movement of persons. In: American Sociological Review 11,6: 677-686.

Date of submission: 05.08.2019 *Date of acceptance: 10.11.2019*

Dr. Hannes Weber (✉). Universität Mannheim, Mannheimer Zentrum für Europäische Sozialforschung (MZES.). Mannheim, Germany.
E-mail: hannes.weber@mzes.uni-mannheim.de
URL: https://www.mzes.uni-mannheim.de/d7/en/profiles/hannes-weber

Appendix

Tab. A1: Feature labels and descriptive statistics

Feature label	Valid N	Mean	Std. Dev.	Min	Max
Population size	3,148	23,763	93,863	4,215	3,326,002
Population density	3,151	4.027	5.760	0.000	75.586
Share of welfare recipients among workers (per 1,000)	3,125	31.829	4.792	10.000	56.000
Share of welfare recipients among working males	3,099	28.376	5.816	6.000	68.421
Share of employment in primary sector	1,997	1.920	3.000	0.020	29.860
Share of employment in secondary sector	2,944	37.785	15.575	2.870	84.600
Unemployment rate, persons aged 15-25 years	3,114	2.880	2.078	0.300	16.100
Unemployment rate	3,114	4.627	2.756	0.900	19.000
Share of foreigners	2,500	5.491	4.210	0.200	33.200
Share of foreigners older than 65 years	1,769	3.194	2.346	0.100	17.100
Share of employed persons in vocational training	3,102	6.032	1.329	2.200	11.900
Share of welfare recipients among foreigners	1,625	15.181	8.758	0.700	63.200
Part-time workers per 1,000 residents	3,109	95.277	22.903	19.548	220.952
Share of age group 65 to 79 years	3,114	15.647	2.403	7.500	29.100
Birth rate	3,105	7.720	1.073	4.000	12.800
Mortality rate	3,105	10.751	2.436	4.500	26.900
Mean age	3,102	44.331	2.145	37.400	53.300
Share of age group 80 or older	3,114	5.416	1.079	2.200	10.900
Mean apartment size per resident	2,996	46.361	4.109	16.900	69.500
Share of 1-family or 2-family homes	3,099	61.873	18.115	10.500	96.100
Growth in the number of jobs in the past 5 years in %	3,093	8.559	11.094	-46.600	117.900
Share of employment in private sector services	2,961	8.583	6.491	0.200	74.700
Dynamics of employment in private sector services, past 5 years	2,902	46.554	81.142	-96.300	837.500
Share of highly qualified workers	3,043	9.284	4.995	1.500	62.100
Share of employed inhabitants among all inhabitants (18 to 64 years)	3,102	55.399	4.410	21.300	70.900
Share of employed females among all female inhabitants (18 to 64 years)	3,035	50.631	5.307	21.200	71.600
Female employment rate as a share of the male employment rate	3,027	84.553	9.735	56.300	123.100

Tab. A1: Continuation

Feature label	Valid N	Mean	Std. Dev.	Min	Max
Share of highly qualified residents	3,114	11.372	4.965	2.800	37.800
Gross tax revenue in the municipality	2,422	682.145	179.026	226.800	999.900
Rate of commuters into the municipality	2,906	65.092	11.431	7.639	95.568
Rate of commuters out of the municipality	2,915	71.924	15.523	9.301	97.262
Youth unemployment rate	3,115	6.540	3.928	0.800	29.100
Child poverty rate	2,944	11.495	8.000	1.100	51.300
Share of sealed/urban land	2,828	19.021	11.253	2.507	95.000
Long-term unemployment rate	3,087	2.641	2.352	0.136	40.397
Recreational space per inhabitant	2,816	0.005	0.005	0.001	0.060
Employment rate of foreigners relative to total	1,819	66.369	16.056	7.451	133.461
Relative number of fatal accidents in transportation	3,069	4.735	1.854	0.596	21.218
Long-distance train connection	3,151	0.092	0.290	0.000	1.000
Public transport connection	3,151	0.583	0.493	0.000	1.000
Distance to long-distance train connection	3,147	16.937	12.790	0.000	83.882
Distance to city >100,000 inhabitants	3,147	35.388	23.994	0.000	173.557
Distance to city >50,000 inhabitants	3,147	22.633	16.118	0.000	151.670
Distance to city >500,000 inhabitants	3,147	81.283	50.247	0.000	352.297
Number of university students	3,151	1,023	7,481	0.000	185,161
University city (yes/no)	3,151	0.052	0.222	0.000	1.000
Distance to nearest university	3,147	20.743	13.248	0.000	106.145
Number of students in 20 km radius	3,147	13,600	26,780	0.000	210,249
Migration rate, 18 to 24 years (2011-2015 average)	3,122	-8.169	38.149	-128.767	316.429
Migration rate, families (2011-2015 average)	3,132	12.637	9.781	-24.896	184.017
Migration rate, 18 to 24 years (2005-2009 average)	3,085	-21.850	29.031	-130.500	181.600
Migration rate, families (2005-2009 average)	3,082	1.004	6.873	-21.200	46.500
Change in share of foreigners, 2 years	2,367	0.474	0.768	-1.465	1.891

Source: own calculations

Comparative Population Studies

www.comparativepopulationstudies.de

ISSN: 1869-8980 (Print) – 1869-8999 (Internet)

Published by
Prof. Dr. Norbert F. Schneider

Federal Institute for Population Research
D-65180 Wiesbaden / Germany