Review of Regional Research (2024) 44:119–140
https://doi.org/10.1007/s10037-024-00207-x

ORIGINAL PAPER

Forecasting first-year student mobility using explainable machine learning techniques

Marie-Louise Litmeyer[1] · **Stefan Hennemann**[1]

Accepted: 9 February 2024 / Published online: 21 February 2024

Abstract In the context of regional sciences and migration studies, gravity and radiation models are typically used to estimate human spatial mobility of all kinds. These formal models are incorporated as part of regression models along with co-variates, to better represent regional specific aspects. Often, the correlations between dependent and independent variables are of non-linear type and follow complex spatial interactions and multicollinearity. To address some of the model-related obstacles and to arrive at better predictions, we introduce machine learning algorithm class XGBoost to the estimation of spatial interactions and provide useful statistics and visual representations for the model evaluation and the evaluation and interpretation of the independent variables. The methods suggested are used to study the case of the spatial mobility of high-school graduates to the enrolment in higher education institutions in Germany at the county-level. We show that machine learning techniques can deliver explainable results that compare to traditional regression modeling. In addition to typically high model fits, variable-based indicators such as the Shapley Additive Explanations value (SHAP) provide significant additional information on the differentiated and non-linear effect of the variable values. For instance, we provide evidence that the initial study location choice is not related to the quality of local labor-markets in general, as there are both, strong positive and strong negative effects of the local academic employment rates on the migration decision. When

✉ Marie-Louise Litmeyer
Marie-Louise.Litmeyer@geogr.uni-giessen.de

Stefan Hennemann
Stefan.Hennemann@geogr.uni-giessen.de

[1] Economic Geography, Department of Geography, Justus Liebig University Giessen, Senckenbergstr. 1, 35390 Gießen, Germany

controlling for about 28 co-variates, the attractiveness of the study location itself is the most important single factor of influence, followed by the classical distance-related variables travel time (gravitation) and regional opportunities (radiation). We show that machine learning methods can be transparent, interpretable, and explainable, when employed with adequate domain-knowledge and flanked by additional calculations and visualizations related to the model evaluation.

Keywords Spatial Mobility · High School-to-University Transition · Machine Learning · Gravitation Model · Radiation Model

1 Introduction

Since the early 2000s, the number of students in Germany has increased significantly more than predicted in forecasts (see, Nutz 1991; KMK 2005; Gösta and von Stuckrad 2007; Wissenschaftliche Dienste des Deutschen Bundestages 2006; Multrus et al. 2017). Reasons for this are the politically desired expansion of higher education offerings, the rising high school graduation rate, the introduction of the bachelor's/master's system, the abolition of compulsory military service, and double high school graduation cohorts. While the deviations of the total predictions are often hard to comprehend due to the effect size of such non-predictable political decisions, it is of great importance for decision makers to forecast spatial patterns of student mobility, since the basic funding is strongly related to the enrolment (HMWK 2015).

Gravity models are typically used to forecast student migration (Sá et al. 2004; Alm and Winters 2009; Cooke and Boyle 2011; Faggian and Franklin 2014). However, these models have some weaknesses, for example, empirical data are needed for fitting (Viboud et al. 2006; Balcan et al. 2009; Kaluza et al. 2010; Krings et al. 2009; Simini et al. 2012). For this reason, Simini et al. (2012) developed the classical radiation model as an alternative approach to estimate mobility between two sites. The advantage of this approach is both, the small amount of data and the parameter freedom. In addition to the classical radiation models, other models have taken up the approach in recent years and have developed the original idea into promising strands (Masucci et al. 2013; Yan et al. 2014; Ren et al. 2014; Kang et al. 2015; Lenormand et al. 2012; Liu and Yan 2019; among others). Nonetheless, these models often do not accurately describe mobility flows (Litmeyer et al. 2023) because a variety of regional characteristics, hard and soft location factors, such as employment rates (Cooke and Boyle 2011; Dotti et al. 2013), play a role in the choice of higher education location besides study location attractiveness, proxied by current enrolment, and distance. The incorporation of co-variates into regression equations is usually improving the model performance greatly. However, as there are typically complex non-linear relationships among the variables and among the observations in spatial interactions, it needs a complex approach to arrive at a valid goodness-of-fit with formal models.

Machine learning algorithms are capable of handling this complexity and the non-linearity but are often criticized for their black-box characteristics of the estimation

procedure and, more importantly, for the limited capability to provide details for the effect sizes of individual variables. However, there are significant improvements in providing transparent, interpretable, and explainable machine learning methods, in recent years (Miller 2019; Roscher et al. 2020 for comprehensive reviews on the requirements for explainable machine learning). Recently, Morton et al. (2018) and Spadon et al. (2019) used the machine learning algorithm XGBoost to show for the USA and Brazil that this algorithm is particularly well suited to predict commuting as one case of spatial interaction and conclude that this method may also be a significant improvement for other cases of spatial interactions. Moreover, the combination with additional calculations of coefficients such as the so-called SHAP values opens a consistent way to interpret the influence of individual variables on the estimate (Morton et al. 2018; Spadon et al. 2019). In this article, the XGBoost algorithm is applied to the case of the transition phase from high-school graduation to higher education, which is frequently related to migration. Therefore, we estimate the number of first-year students per German county, based on a comprehensive set of hard and soft factors of college location choice. The following questions are answered:

- How can we employ machine learning algorithms such as the XGBoost algorithm in order to deliver comprehensible results that are transparent, interpretable and explainable, when extended with model specific and variable specific indicators and visualizations?
- How do results compare to those presented in the literature, typically based on formal regression models, and what are additional insights into the knowledge domain of student migration from the machine learning approach, leading to an original contribution of such techniques?

The second chapter introduces the higher education landscape in the case study area of Germany. This is followed on the regional characteristics for migration processes of first-year students derived from the international literature and the state-of-the-art. The methodology used is then presented, and presents the results. This is followed by a discussion and an outlook.

2 Germany as a higher education location

The German higher education system has changed considerably in the last 30 years after German reunification. First, new universities were founded in the 1990s in the territories of the former German Democratic Republic (GDR, 64 in total) (Erhart 2002). The new establishments, since 2000 (91 in total), are often private schools, private universities, satellite campuses, regional offshoots of existing vocational academies, or spin-offs from universities and research institutions (e.g., the Baden-Württemberg Cooperative State University or the Karlsruhe Institute of Technology (HRK 2019; KIT 2018)). In addition, the Bologna Declaration of June 13, 1999, led to the harmonization of study structures with bachelor's and master's degrees and thus to greater compatibility and comparability in the European Higher Education Area (EHEA 2016). Besides the changes to the degrees, other changes, both internal

and external to higher education institutions, were made. Between 2006 and 2007, tuition fees were initially introduced in all western German states (except Bremen, Rhineland-Palatinate and Schleswig-Holstein) and abolished again by 2014 due to political changes and changes in government (Kauder and Potrafke 2013). The switch from a nine-year to an eight-year Abitur (KMK 2012) and the suspension of compulsory military service in 2011 (Deutscher Bundestag 2011) led to significant increases in student numbers. The average annual growth rate in the number of first-semester students (German and foreign first-semester students in all types of higher education institutions) from 2008 to 2016 was 2.82%. In total, enrolment increased from 396,800 to 509,760 during that time. In 2011, the highest number of students in the first semester was measured at 518,748 students (Statistisches Bundesamt 2008–2017). Overall, Germany as a knowledge location is particularly well suited as a study area since universities and universities of applied sciences are (relatively) evenly distributed throughout the country due to historical and political reasons. Moreover, Germany consists of rather homogeneous budgeting situations in the higher education system, when compared to the strong disparities in the Anglo-American higher education system. This means that, unlike in the USA, there are no "educational deserts" in Germany (Hillman 2016) and the lack of tuition fees does not lead to large distorting effects.

3 Motives for student mobility

The most important aspect, for a study location decision, is the spatial proximity between the origin and destination (Alm and Winters 2009; Dotti et al. 2013; Gibbons and Vignoles 2012). In general, student migration intensity decreases exponentially with increasing distance (e.g., Montgomery 2002; Sá et al. 2004; Frenette 2004, 2006; Spiess and Wrohlich 2010; Alm and Winters 2009; Dotti et al. 2013; Gibbons and Vignoles 2012). Along with increasing distance, emotional costs, e.g. giving up social ties, are a barrier to student mobility in addition to higher relocation and transportation costs (Winters 2011; Dotti et al. 2013, 2014). However, e.g., marketing activities of universities (Vrontis et al. 2007) or strong collaboration between universities and schools reduce the negative effects of geographical distance (Raab et al. 2018). Regions with a higher degree of urbanization and higher population density are more attractive and often draw in students (Sá et al. 2004; Cullinan and Duggan 2016; Weisser 2019). The same is true for regions with high employment rates (Cooke and Boyle 2011; Dotti et al. 2013). Furthermore, especially in agglomeration areas, financial aspects such as rent levels (Dotti et al. 2013) and future earning potential in the home and destination regions play a role in study choices (Sá et al. 2004). The direction of impact of tuition fees, on the other hand, is not always clear and depends on the level of fees charged (Spiess and Wrohlich 2010; Dwenger et al. 2012; Dotti et al. 2013). In Italy (Ciriaci 2014), the U.S. (Cooke and Boyle 2011), and Ireland (Walsh et al. 2018), a significant impact of the quality of higher education teaching on student mobility could also be measured, while no impact could be detected by Sá et al. (2004) for the Netherlands. The employment rate of graduates (Sá et al. 2011), faculty and student ratios (Sá et al. 2004), expen-

diture per student (Cullinan and Duggan 2016), research intensity (Adkisson and Peach 2008) or the place in international rankings (Ciriaci 2014) served as measurement indicators for quality. Both the educational background of parents (Lörz 2008) and gender (Belfield and Morris 1999; Ciriaci 2014) influence student mobility. In addition, different studies suggest that potential students often migrate to student-dominated regions or regions with a high share of highly educated people due to similar lifestyles, as well as to regions with strong cultural proximity (Buenstorf et al. 2016; Haussen and Uebelmesser 2018). This multitude of indicators explaining student mobility shows that a very large number of highly individual factors can play a role in the decision-making process for and against a particular university. Finally, location- and weather-related amenities are also important in the choice of study location (Kodrzycki 2001).

The migration patterns of first-year students cannot be discussed completely isolated from the whereabouts of students after graduation, as universities contribute greatly to regional economic activities (e.g., Kodrzycki 2001; Marinelli 2013; Dotti et al. 2013; Krabel and Flöther 2014; Kitagawa et al. 2022). In this context, universities and colleges as centers for research and development as well as teaching and training students occupy a special position in the (regional) innovation system (Geissler and König 2021). On the one hand, they generate knowledge, make it available to other stakeholders and promote the development of the next generation of scientists. It has been observed for years that more and more people are doing their doctorate and working at universities after completing their doctorate (e.g. Briedis et al. 2014; Buenstorf et al. 2023). On the other hand, the training of students is an important aspect for the labor market.. The private sector benefits from the well-trained graduates as well as from the corresponding knowledge of the universities and can thus improve its innovative capacity (Fritsch and Slavtchev 2007).

Another aspect is that cooperation between private-sector companies and universities in the manner of scientific publications, seminars, workshops and informal relationships have a positive influence on the transfer of academic knowledge to industry (Fritsch and Slavtchev 2007). However, academic knowledge is relatively immobile in this context, so geographical proximity and graduate ties play a vital role. This offers the advantage of directly increasing a region's human capital endowment and thus having an impact on its innovation potential in the medium to long term.

Accordingly, the retention of graduates in a region is relevant, even before higher education policy measures such as scholarships. However, the effectiveness of scholarships is controversial (Groen 2004; Busch and Weigert 2010; Geissler and König 2021). For Germany, Busch and Weigert (2010) and Buenstorf et al. (2016) showed that more than half the number of graduates take up employment in the university region and the corresponding state or return to their home region.

Furthermore, it can be observed that graduates and scientists often work near their home university and newly founded innovative companies also actively seek spatial proximity to universities. Whereby basically regional differences exist between urban and non-urban areas as well as the fields of study (Marinelli 2013; Buenstorf et al. 2016; Kitagawa et al. 2022). Krabel and Flöther (2014) and Kitagawa et al. (2022) were able to demonstrate that urban areas or metropolitan regions have a high

retention of university graduates, while rural areas are characterized by a higher mobility requirement of graduates. In non-urban areas, the establishment of a company at the university location seems to increase the retention in the region. It can be seen that the retention rate of graduates in natural sciences is significantly higher in urban regions. One reason for this is that labor markets in agglomeration areas increase the match between STEM graduates and STEM professions (Kitagowa et al. 2022). Krabel and Flöther (2014) and Teichert et al. (2020) were able to show that graduates are more likely to stay in the university region if they gain work and professional experience in the university region during their studies (Teichert et al. 2020).

This wide range of indicators explaining student mobility highlights that a large number of highly individual factors can play a role in the decision for or against a particular university.

4 Methodology

We seek to predict the number of students at any German county that hosts a higher education institution, based on the aggregate of dyadic migration decisions. In order to be able to predict the weighting of each connection more reliably, we employ three XGBoost regressors, each of it, representing the number of first-year students who migrate from their home county i, i.e. the place of high-school graduation, to the university location j.

The XGBoost algorithm is a method that uses the mathematical data representation of decision trees. Decision trees are non-parametric and often used in supervised machine learning. They use loss functions to evaluate the gradual improvements of the predictions during the learning process. Therefore, they belong to the class of ensemble learning problems. The procedure starts with an initial calculation of a simple model (a tree), which is used to predict the training data. The error of these predictions compared to the actual values is then determined by a loss function and another tree is created to minimize these errors (gradient descent optimization). This process is repeated and with each new tree the error of the previous tree is corrected. Since all machine learning methods have a stochastic element, model outputs may not be deterministic, compared to formal regression modeling. Thus, the whole procedure is usually repeated to arrive at ensembles of converged predictions of all trees and are then averaged.

In a basic model 1, we consider the first-year students to m_i and n_j at the home and university locations, and the distance r_{ij} between the locations. This is equivalent to a gravitation model, but using non-linear estimation techniques from machine learning.

A second model is based on the seminal idea of Simini et al. (2012), who introduced a radial "opportunity" component. This led to a significant improvement of the forecast of commuter movements for the U.S. at the municipality level, utilizing a very reduced set of variables (number of inhabitants in the destination and origin region (m_i; n_j) as well as s_{ij} defined as inhabitants from all locations within radius ij

around i, the total number of commuters T_i in the system) and without parameters. Formally, it follows that

$$T_{ij}^{\text{radial}} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} = \frac{\vartheta}{M} \frac{m_i^2 n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}$$

Transferring these considerations to the mobility of first-year students, it follows that m_i is defined as the number of high-school graduates m at place i. For the university location j, n_j is chosen at time t-1. It represents the number of students in the first semester at time t-1. Basically, it is assumed that future students compare university locations considering different conditions. From these considerations, it follows that s_{ij} describes the total number of freshmen at time t-1 within the radius of the distance between the home county and the future university location, around the home county, and thus represent all potential university locations in the vicinity. To further calculate the average freshman migration T_{ij} from location i to j, the average freshman migration rate at time t-1 for the entire country is also calculated. ϑ is the total number of all mobile students (excluding students whose home region corresponds to the university region) and M describes the number of all first-year students in Germany. Thus, this second model is an extension of model one by adding s_{ij} as the representation of all other opportunities within a given distance ij around i. Again, the XGBoost regressor is allowing for non-linear relationships among the variables and observations. The calculation is performed using the R package 'xgboost' (Yuan 2023).

The final model 3 incorporates 28 co-variates to model 2 to control for important aspects in the study location choice of high-school graduates. A student decision is modeled as

$$\boldsymbol{U_k} := \left\{\boldsymbol{u_k^1}, \ldots, \boldsymbol{u_k^{28}}\right\}, \boldsymbol{for\, k} \in \{\boldsymbol{i}, \boldsymbol{j}\} \boldsymbol{and}\, \boldsymbol{u_k^m} \in \mathbb{R} \boldsymbol{for\, m} \in \mathbb{N}$$

To comprehend the most relevant motives, derived from the literature for the home and the university location respectively (cf. Table 1). That is, for each interaction and set:

$$\mathbb{R}^{|\boldsymbol{S}|} \ni \boldsymbol{S_{ij}} := \left\{\boldsymbol{r_{ij}}, \boldsymbol{s_{ij}}, \boldsymbol{U_i}, \boldsymbol{U_j}\right\},$$

the following function is sought (Spadon et al. 2019):

$$\mathit{weight} : \mathbb{R}^{|\boldsymbol{S}|} \longrightarrow \mathbb{N}$$

The co-variates control for infrastructure (e.g. Accessibility of IC/EC/ICE stations), supporting/soft location factors (e.g. Guest overnight stays), and environmental aspects (e.g. Average temperature) and can be defined and statistically described as follows:

Due to the choice of methodology, it is not necessary to normalize the variables accordingly. To be able to calculate the regressor, the data are first divided into a training data set (70%) and a test data set (30%). This is followed by a 5-fold

Table 1 Overview of the used variables

Variable	Definion	*n*	Min	25% quantile	Median	Mean	75% quantile	Max	sd
Migration of first-year students	First-year students moving from home region i to college region j	90,626	0	0	0	5	0	15,003	73
Travel time	Travel time from district town to district town in minutes	90,626	0	230	364	373	503	956	184
m	Number high-school graduates at place i at time t	401	55	306	493	742	870	13,871	993
n	Number of first-year students at time t-1	401	0	0	96	1079	910	26,797	2524
s	Total number of first-year students within the radius between home district and future university location around the home district, representing opportunities in the vicinity	90,626	0	741,751	1,554,702	1,549,084	2,320,843	3,128,331	907,418
Building land prices	Average land prices in € per m^2	344	12	39	83	138	174	2027	188
Average age	Average age of the population in years	401	40	43	44	44	46	50	2
Household income	Average household income in € per inhabitant	401	1293	1622	1775	1786	1919	3028	215
Childcare rate for young children	Proportion of children under 3 years of age in day-care facilities as a percentage of children in the corresponding age group	401	14	23	27	31	34	61	12
Higher education funding (long-term)	Actual expenditures (2020) by the Federal Government in cooperation between the Federal Government and the Länder for research buildings and large-scale equipment in accordance with Art. 91 of the Basic Law, grants under the Excellence Initiative and the Higher Education Pact (areas of lump-sum program funding and the Teaching Quality Pact) (long-term) in € per inhabitant	401	0	0	0	97	19	2700	288

Table 1 (Continued)

Variable	Definion	*n*	Min	25% quantile	Median	Mean	75% quantile	Max	sd
Population density	Inhabitants per km^2	401	36	115	200	531	661	4713	699
Accessibility of motorways	Average travel time to the nearest highways junction in minutes (2021)	401	0	7	10	13	16	54	8
Accessibility of IC/EC/ICE stations	Average travel time to the nearest IC/ICE station in minutes (2021)	401	0	14	22	22	31	75	15
Accessibility of regional centers	Average travel time to the nearest regional centre in minutes (2021)	401	0	14	24	24	33	69	16
Average distance supermarkets	Population-weighted linear distance to the nearest supermarket or discount store (2021)	401	328	672	1180	1179	1575	2987	546
Average distance public transport stops	Population-weighted linear distance to the nearest public transport stop with at least 20 departures per day (2020)	401	144	232	392	580	694	6978	573
Guest overnight stays	Overnight stays in accommodation facilities per inhabitant	397	0	2	3	6	6	45	7
GDP per inhabitant	Gross domestic product in 1000 € per inhabitant	401	16	26	32	36	38	178	16
Broadband supply	Proportion of households with broadband coverage of 100 Mbit/s in % (2017)	401	1	43	59	60	79	99	22

Table 1 (Continued)

Variable	Definion	*n*	Min	25% quantile	Median	Mean	75% quantile	Max	sd
Employment rate	Employees subject to social insurance at place of residence per 100 inhabitants of working age in %	401	43	56	59	59	62	69	4
Employees in IT and scientific service professions	Proportion of employees subject to social insurance contributions in IT and scientific service occupations as a percentage of employees subject to social insurance contributions	401	1	2	2	3	4	20	2
Employees with academic degrees	Proportion of employees subject to social insurance at the place of residence with an academic vocational qualification among employees subject to social insurance in %	401	5	9	11	13	15	42	6
Recreational areas	Recreational area in m^2 per inhabitant	401	14	39	51	68	72	384	55
Avarage temperature	Multi-year mean temperature in °C for the districts 1981–2010	401	7	9	10	10	10	78	4
Precipitation	Multi-year mean precipitation in mm for the districts 1981–2010	401	505	685	778	805	860	1866	198
Global radiation	Multi-year mean global radiation in W/m^2 for the counties 1981–2010	401	111	115	120	121	126	134	6

Source: own presentation based on BBSR (2023); Deutscher Wetterdienst (2018, 2022, 2023); FDZ (2019); INKAR (2023)

cross-validation and the tuning of the hyperparameters. For this purpose, a grid with the hyperparameters (eta, gamma, min_child_weight, max_depth) is formed and all possible variants are tested so that the Sörensen index is maximized. The hyperparameter eta corresponds to the learning rate and stands for the step size that is used during the update to prevent overfitting. In addition, gamma is adjusted. This parameter stands for the minimum loss reduction that occurs when the nodes are split. Basically, the larger gamma is, the more conservative the algorithm becomes. In addition, max_depth is used to specify the maximum depth of the tree with the aim of controlling overfitting. The fourth parameter that is adjusted is min_child_weight and also aims to minimize overfitting. In this case, the larger the value, the more conservative the algorithm becomes.

In addition, 70% percentage of regional features (columns) is used in the construction of each tree to counteract possible endogeneity problems. This means that each tree is only built with 70% of the columns. Then, the tuned regressor is applied to the test dataset and the goodness of fit is evaluated using various indicators (Spadon et al. 2019).

The widely used parameter to assess interaction is the Soerensen index (Soerensen 1948). It is used to measure fluctuation and indicates the correctly reproduced proportion of pendulum flows in simulated networks. The similarity measure can take values between 0 and 1. Provided a value of 0 is assumed, there is no correspondence with the original pendulum flows. For 1, the empirical network fully corresponds to the simulated network. Comparatively, the advantage of the Soerensen index is that it maintains sensitivity in more heterogeneous data sets and is less sensitive to outliers (McCune and Grace 2002). The measurement indicator is calculated as follows where ${T_{ij}}^{\text{empric}}$ represents the empirical and T_{ij}^{model} the calculated commuter flows (Soerensen 1948):

$$SI = \frac{2\sum_{i=1}^{N}\sum_{j=1}^{N} min\left({T_{ij}}^{\text{empric}}, T_{ij}^{\text{model}}\right)}{\sum_{i=1}^{N}{T_{ij}}^{\text{empric}} + \sum_{j=1}^{N}T_{ij}^{\text{model}}}$$

The evaluation is complemented by the Mean Square Error, the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), the Pearson correlation coefficient and the adjusted R^2 value.

For the evaluation of the XGBoost regressor, a reconstruction of the weighted structure of the mobility network is also performed. For this purpose, the influence of the characteristics on the forecast is determined using the SHAP value (Shapley Additive Explanations value) (Lundberg and Lee 2017; Lundberg et al. 2018a, b). As early as 1953 Lloyd Shapley introduced Shapley values in the context of game theory (Shapley 1953). The basic idea is that the prediction of the model is made with and without the feature in question. It should also be noted that the order in which new features are added has an effect on the model, so all permutations of the feature orders must be calculated.

The advantage of this approach is that the effects of the characteristics on the prediction of the individual data become possible, since in the case presented here, the SHAP value measures the contribution or importance of a county to the forecast.

For this, a graph (e.g., Figs. 2 and 3) is created based on the SHAP values using the R package SHAPforxgboost to better interpret the results (Liu et al. 2021). The most important characteristic is placed at the top. The SHAP values can also be graphically displayed, so that the effect of each feature can be immediately recognized. For each feature, a dot representing the predicted association is drawn. Thus, it is possible to determine the distribution of the impact of each feature on each interaction. Points that are in the negative range indicate that this predicted association has a negative impact on the model's prediction performance. Conversely, a point in the positive range is an indication that the prediction is improving. The colors also represent the SHAP value and vary from low (yellow) to high values (purple). In each row the mean value of the amounts of all SHAP values for the respective variable is given. In contrast to standardized beta coefficients in traditional regression analysis, SHAP visual representations can help differentiating non-linear relations between the dependent and the independent variables in an explorative way.

Overall, only a few examples so far use the full range of options of feature extraction and SHAP analysis along with an informative visualization. One of the notable exceptions is Li (2022). However, most authors still use XGBoost as a black box for prediction without addressing the contributions of the features (e.g. Rahman and Chowdhury 2022), thus, somehow violating the requirements for transparent, interpretable and explainable machine learning applications, as discussed in Miller (2019) or Roscher et al. (2020). The approach presented here, comes intuitively close to formal regression analysis and its interpretation. Through the explicit feature extraction, the detailed effect size analysis through the SHAP values and subsequent visualization of the parameter influence, analysts are able to present important drivers of the effects behind the phenomenon under investigation in a comprehensible way.

5 Empirical results

The dataset includes variables for the 401 counties and independent cities in Germany, which serve as the study area for the influence of study motivation on migration behavior in 2016. We omitted all dyads that include counties that do not host a higher education institution, because there is not option for studying and, thus, no migration potential. The data for the regional characteristics (see Table 1) were taken from the database "Indicators and Maps of Spatial and Urban Development" of the Federal Institute for Research on Building, Urban Affairs and Spatial Development (INKAR 2023). In addition, travel time in minutes between all district cities of the counties were obtained from BBSR (2023) and climatic indicators representing mean values of the counties calculated on the basis of raster data of the German Weather Service (Deutscher Wetterdienst 2018, 2022, 2023). The query of migration flows of students from the home county to the university location was made in the research database Frankfurt (FDZ 2019). For data protection reasons migration flows with less than three students are considered with 0 migrations.

All three models can be evaluated, using the proposed model diagnostics. Table 2 shows the Soerensen index with 0.78, the MAE with 1.57 and the adjusted R^2 value with 0.81 for model 3, qualifying this model as best model. Overall, results

Table 2 Results of the evaluation indicators

Indicator	Model 1	Model 2	Model 3
Soerensen index	0.64	0.69	0.75
MAE	4.32	3.81	3.26
RMSE	103.97	93.47	75.43
R^2 *adjusted*	0.73	0.75	0.80
Correlation coeff. (Pearson)	0.52	0.72	0.87

Source: own calculations based on BBSR (2023); Deutscher Wetterdienst (2018, 2022, 2023); FDZ (2019); INKAR (2023)

improve gradually from model 1 to model 3, when incorporating more information, which is in line with standard procedures in classical regression analyses. There is a great improvement from model 2 to 3, which emphasizes the importance of including larger numbers of conceptually important co-variates. Thus, different from the experience with formal regression models, the acknowledgement of complex interactions between the observations and the variables and the non-linear learning procedure led to an increasingly good fitting of the machine learning model. In addition to that, the introduction of co-variates greatly improves the interpretability and transparency.

Figure 1 shows accessibility and average distances to existing infrastructure facilities (e.g. supermarkets, etc.) are particularly strongly negatively correlated with population density, the number of first-year students and the share of employees with academic degrees. There is a strong positive correlation between the number of first-year students (n) and long-term university expenditures and employees with academic degrees. A strong positive correlation can also be seen for travel time and the number of first-year students between the home region and the university region (s).

Looking at the SHAP values for the XGBoost regressor (see Fig. 2), it is clear that the number of first-year students at the university location is the most important characteristic in all three models. Model 3 demonstrates that university locations with a high number of first-year students have a positive effect on predicting student mobility flows. Locations with low numbers of first-year students have a negative effect, thus, the attractiveness of a study location is self-reinforcing and strongly path-dependent. Moreover, very skewed distributions and outliers seem to induce more extreme SHAP values, since all the SHAP values for variable n that are below –5, consists of locations with very low spatial interactions. Another relevant aspect is travel time as especially short travel times have a high impact on the predictive performance of the models, while long travel times decrease the predictive performance. Furthermore, it can be seen that the parameter s, introduced by Simini et al. (2012)—which represents the alternative opportunities for the study location selection with radius of the distance between the home region and the future university location—can also be identified as another important regional characteristic in models 2 and 3 (cf. Figs. 2 and 3). Locations in the vicinity of which there are a large number of first-year students, on the other hand, have a negative influence on the prediction of migration. It becomes clear that the number of first-year students at

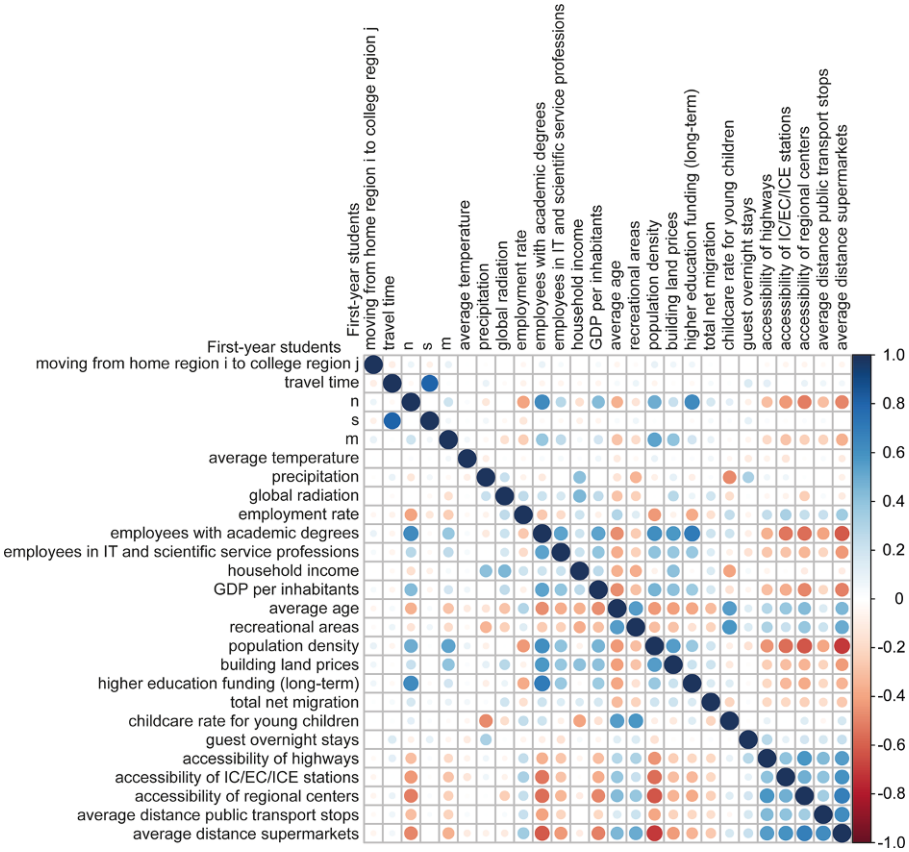

Fig. 1 Correlation of variables used. Source: own representation based on own calculations (BBSR 2023; Deutscher Wetterdienst 2018, 2022, 2023; FDZ 2019; INKAR 2023)

the place of residence plays a role in all three models and that large locations benefit from their local pool of high school graduates.

Looking at model 3 (cf. Fig. 3), it becomes apparent that the regional characteristics in the respective university regions are of high importance. Among other things, long-term university funding measures have a surprisingly large influence on the forecast, although a clear direction of effect is not discernible here. Funding may also be interpreted as proxy for other quality related aspects of the students' decision that are hard to capture otherwise such as quality of teaching and research. Furthermore, it becomes clear that low average distances to the nearest supermarket and high population densities at university locations have a positive effect on the calculation of student flows. Low population densities, on the other hand, have the opposite effect. Also, a high proportion of people employed in the IT sector, high childcare rates for young children and small households improve the prediction. Similar observations can be made for overnight guest stays. Locations with high numbers of overnight stays in the home and university regions positively influence the forecast.

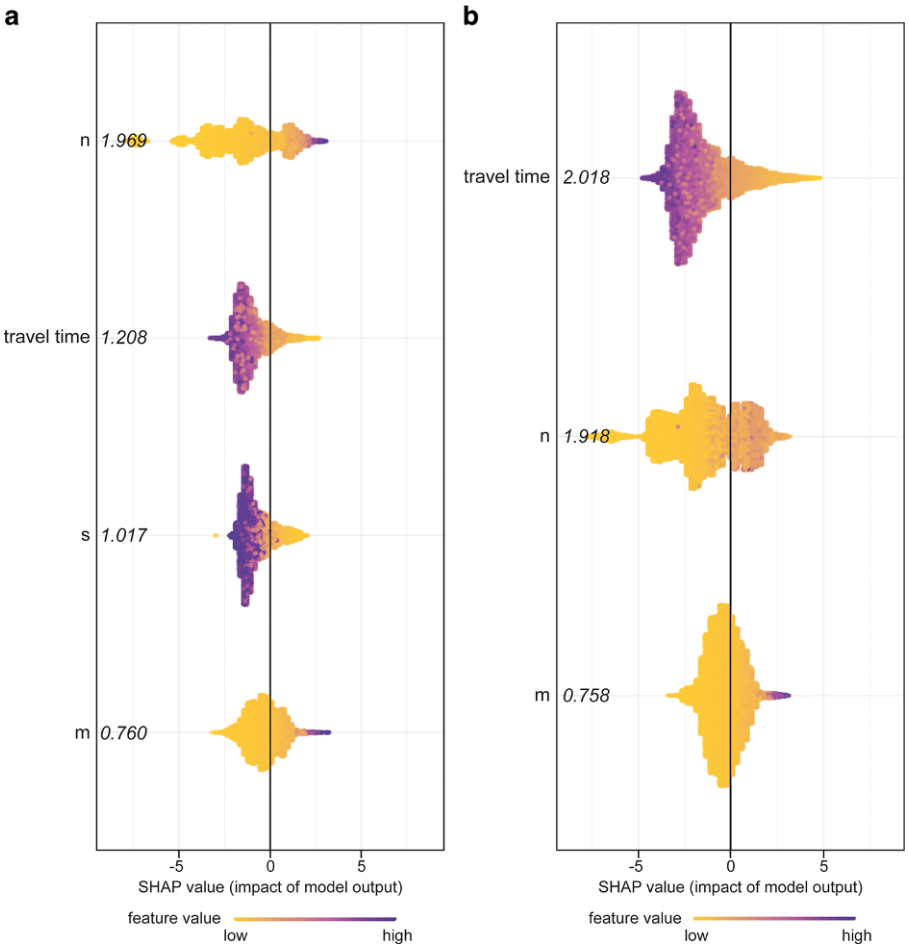

Fig. 2 SHAP values for the XGBoost regressor (model 1 & 2). Source: own representation based on own calculations (BBSR 2023; INKAR 2023)

Counties with a high age structure cause a deterioration of the predicted mobility flows. This could also be determined for the population structure in the home county. In addition, high global radiation in the university region and a low proportion of local recreation areas is an aspect that is also of positive significance for the forecast. Fundamentally high employment rates at the place of residence and work tend to lead to a deterioration of the forecast. A more differentiated look at the proportion of employees with an academic vocational degree at the university location shows a high proportion has both a positive and a negative effect. No clear statements can be made for all other characteristics, such as household income.

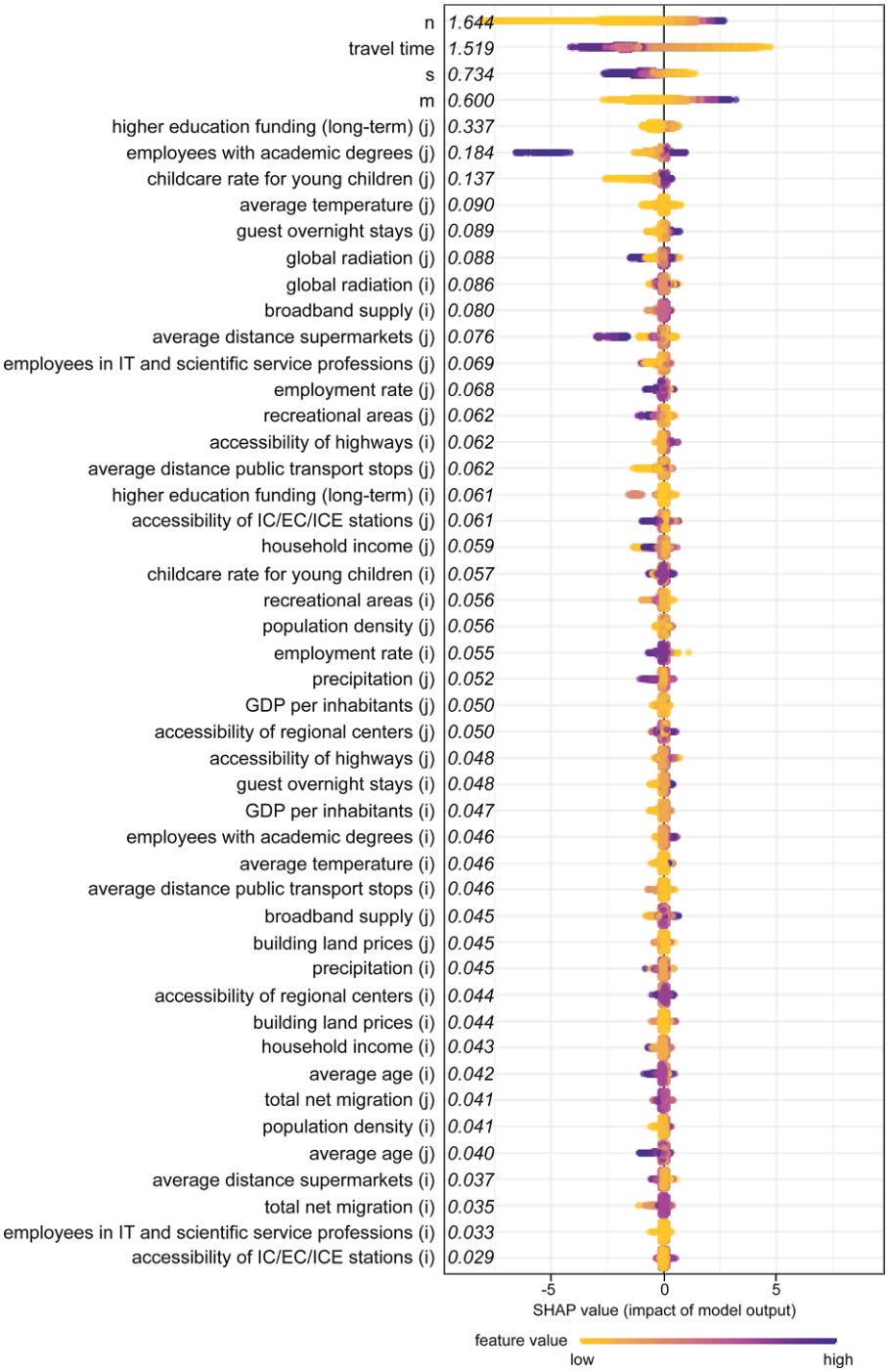

Fig. 3 SHAP values for the XGBoost regressor (model 3). Source: own representation based on own calculations (BBSR 2023; Deutscher Wetterdienst 2018, 2022, 2023; FDZ 2019; INKAR 2023)

6 Discussion and conclusion

This article was exploring, what machine learning methods can offer for spatial interaction modeling. We provided evidence that algorithms such as the XGBoost algorithm can deliver comprehensible results that are transparent, interpretable and explainable. We have suggested a set of model diagnostics and visualizations to support the interpretability of the results. What seems most important from the knowledge domain perspective is the need for a comprehensive acknowledgement of independent variables and co-variates. Machine learning techniques are already providing good model fits to the empirical data with few parameters as presented in models 1 and 2 in this study. However, the predictions would be less explainable without a decent amount of additional conceptually derived variables. It must also be discussed at this point that the prediction using the XGBoost algorithm has an explorative character due to the hyper-parameterization. The prediction of the migrations can be optimized through the targeted control and coordination of the parameters through loss functions which guarantee a gradient descent.

Nevertheless, the approach also offers advantages. In particular, the visualization by means of the SHAP-values offers a deeper insight into the black-box of the algorithm. It contributes to the understanding of the individual positive or negative influence of each region. This also enables to measure the respective influence of counties or municipalities in other areas of interest for regional phenomena. This is something, traditional regression methods cannot offer.

Overall, the results from the XGBoost Regression compare very well to the state-of-the-art, presented in the literature concerning our case study on high-school-graduates' migration to their preferred place of study. The socio-economic structure of the respective university region is of great relevance. The most important aspect, as discussed in the literature is the number of first-year students in the previous year, which can be interpreted as attractiveness of the location for prospective students, and the travel times to the university location. Locations with many students have a positive influence on the forecast.

In addition, interactions with very small migration movements (< 10) have a strong negative influence on the forecast. Migratory movements that are somewhat larger (≥ 20) also have a negative influence on the forecast performance, but this is significantly smaller.

Besides the size of the university location, the proximity to the home region is of particular relevance in the forecast. It becomes evident that the choice of university location is strongly dependent on the number of opportunities in the surrounding of the home location. In regions of origin where there are many first-years and opportunities, the likelihood of choosing a particular college location decreases. Contrastingly, in locations where there are few universities in the immediate vicinity, it can be assumed that these universities will accept many high school graduates.

Among the regional characteristics, aspects related to agglomeration effects are very important. High population densities, a well-developed infrastructure and (e.g. proximity to the nearest supermarket) basic services, a high proportion of employees, large numbers of overnight stays, small household sizes and a high rate of childcare lead to an improvement in the forecast. This is particularly interesting since these

regional characteristics are mainly aspects that also play a role for graduates. In other words, regions with such a structure benefit more than average from the immigration of first-year students. This may also contribute to the recent observation of increasing employment opportunities in academia (Buenstorf et al. 2023).

The present analysis does not consider individual factors of the high-school graduates in the decision-making process, such as gender, educational background, and family ties, due to a lack of data. Likewise, indicators that represent qualitative aspects of a study location such as the quality of teaching or the structure of the study program were also excluded. Thus, the attractiveness of the university location remains obscure and hidden behind the residual variable n in our case study. This being said, there is need to further explore the capacities of machine learning for the purpose of the development of new indicators that grasp such fuzzy concepts like the attractiveness of a region. One promising avenue in this respect is discussed in Kriesch (2023), who suggests using machine learning and large language modeling for the classification of website data and the production of new regionalized variables in empirical studies in the field of economic geography and regional sciences. Reflecting the encouraging results from the analysis presented here, it seems worthwhile to further explore what the dynamic field of machine learning has to offer for our knowledge domain of economic geography and regional sciences.

Acknowledgements We would like to thank Lisett Diehl for her cartographic support. We also thank the two anonymous reviewers for their constructive comments on this paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest M.-L. Litmeyer and S. Hennemann declare that they have no competing interests.

References

Adkisson RV, Peach JT (2008) Non-resident enrolment and non-resident tuition at land grant colleges and universities. Educ Econ 16(1):75–88. https://doi.org/10.1080/09645290701563156

Alm J, Winters JV (2009) Distance and intrastate college student migration. Econ Educ Rev 28(6):728–738. https://doi.org/10.1016/j.econedurev.2009.06.008

Balcan D, Colizza V, Goncalves B, Hud H, Ramasco J, Vespignani A (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci U S A 106:484–489. https://doi.org/10.1073/pnas.0906910106

BBSR (2023) Fahrzeiten in Minuten von Kreisstadt zu Kreisstadt (on request)

Belfield C, Morris Z (1999) Regional migration to and from higher education institutions: scale, determinants and outcomes. High Educ Q 53(3):240–263. https://doi.org/10.1111/1468-2273.00129

Briedis K, Jaksztat S, Preßler N, Schürmann R, Schwarzer A (2014) Berufswunsch Wissenschaft. In: Laufbahnentscheidungen für oder gegen eine wissenschaftliche Karriere. Forum Hochschule

Buenstorf G, Geissler M, Krabel S (2016) Locations of labor market entry by German university graduates: is (regional) beauty in the eye of the beholder? Rev Reg Res 36(1):29–49. https://doi.org/10.1007/s10037-015-0102-z
Buenstorf G, Koenig J, Otto A (2023) Expansion of doctoral training and doctorate recipients' labour market outcomes: evidence from German register data. Stud High Educ: 1–27
Busch O, Weigert B (2010) Where have all the graduates gone? Internal cross-state migration of graduates in Germany 1984–2004. Ann Reg Sci 44(3):559–572
Ciriaci D (2014) Does university quality influence the interregional mobility of students and graduates? The case of Italy. Reg Stud 48(10):1592–1608. https://doi.org/10.1080/00343404.2013.821569
Cooke TJ, Boyle P (2011) The migration of high school graduates to college. Educ Eval Policy Anal 33:202–213. https://doi.org/10.3102/0162373711399092
Cullinan J, Duggan J (2016) A school-level gravity model of student migration flows to higher education institutions. Spat Econ Anal 11(3):294–314. https://doi.org/10.1080/17421772.2016.1177195
Deutscher Bundestag (2011) Aussetzung der allgemeinen Wehrpflicht beschlossen. https://www.bundestag.de/dokumente/textarchiv/2011/33831649_kw12_de_wehrdienst-204958. Accessed 21 Jan 2022
Deutscher Wetterdienst (2018) Vieljähriges Mittel der Raster der Niederschlagshöhe für Deutschland 1981–2010. https://opendata.dwd.de/climate_environment/CDC/grids_germany/multi_annual/precipitation/. Accessed 20 July 2023
Deutscher Wetterdienst (2022) Raster der vieljährigen Mitteltemperatur in °C für Deutschland – HYRAS-DE-TAS. https://opendata.dwd.de/climate_environment/CDC/grids_germany/multi_annual/hyras_de/air_temperature_mean/. Accessed 20 July 2023
Deutscher Wetterdienst (2023) Raster der vieljährigen mittleren Globalstrahlung in W/m^2 für Deutschland – HYRAS-DE-RSDS. https://opendata.dwd.de/climate_environment/CDC/grids_germany/multi_annual/hyras_de/radiation_global/. Accessed 20 July 2023
Dotti NF, Fratesi U, Lenzi C, Percoco M (2013) Local labour markets and the interregional mobility of Italian university students. Spat Econ Anal 8:443–468. https://doi.org/10.1080/17421772.2013.833342
Dotti NF, Fratesi U, Lenzi C, Percoco M (2014) Local labour market conditions and the spatial mobility of science and technology university students: evidence from Italy. Rev Reg Res 34(2):119–137. https://doi.org/10.1007/s10037-014-0088-y
Dwenger N, Storck J, Wrohlich K (2012) Do tuition fees affect the mobility of university applicants? Evidence from a natural experiment. Econ Educ Rev 31(1):155–167. https://doi.org/10.1016/j.econedurev.2011.10.004
EHEA (2016) The Bologna Declaration of 19 June 1999. http://ehea.info/me-dia.ehea.info/file/Ministerial_conferences/02/8/1999_Bologna_Declaration_English_553028.pdf. Accessed 2 May 2019
Erhart M (2002) Gemeinsame Strukturen finden – Der „Masterplan" zu Beginn der 90er Jahre. In: Stifterverband für die Deutsche Wissenschaft e. V. (ed) 10 Jahre danach – Zur Entwicklung der Hochschulen und Forschungseinrichtungen in den neuen Ländern und Berlin. Stifterverband für die Deutsche Wissenschaft, pp 70–73
Faggian A, Franklin RS (2014) Human capital redistribution in the USA: the migration of the college-bound. Spat Econ Anal 9:376–395. https://doi.org/10.1080/17421772.2014.961536
FDZ (2019) Statistik der Studenten. 1992–2017. https://www.forschungsdaten-zentrum.de/de/bildung/studenten. Accessed 6 June 2019
Frenette M (2004) Access to college and university: does distance to school matter? Can Public Policy. https://doi.org/10.2307/3552523
Frenette M (2006) Too far to go on? Distance to school and university participation. Educ Econ 14(1):31–58. https://doi.org/10.1080/09645290500481865
Fritsch M, Slavtchev V (2007) Universities and innovation in space. Ind Innov 14(2):201–218
Geissler M, König J (2021) 'See you soon?!' Mobility, competition and free-riding in decentralized higher education financing. Reg Stud 55(4):665–678
Gibbons S, Vignoles A (2012) Geography, choice and participation in higher education in England. Reg Sci Urban Econ 42(1–2):98–113. https://doi.org/10.1016/j.regsciurbeco.2011.07.004
Gösta G, von Stuckrad T (2007) Die Zukunft vor den Toren. Aktualisierte Berechnungen zur Entwicklung der Studienanfängerzahlen bis 2020. https://www.che.de/download/che_prognose_studienanfaengerzahlen_ap100-pdf/?wpdmdl=11190ind=5d1a0a27b4528. Accessed 21 Jan 2022
Groen JA (2004) The effect of college location on migration of college-educated labor. J Econom 121(1–2):125–142
Haussen T, Uebelmesser S (2018) No place like home? Graduate migration in Germany. Growth Change 49(3):442–472. https://doi.org/10.1111/grow.12249

Hillman NW (2016) Geography of college opportunity: the case of education deserts. Am Educ Res J 53(4):987–1021. https://doi.org/10.3102/0002831216653204
HMWK (2015) Hochschulpakt 2016–2020. https://wissenschaft.hessen.de/si-tes/default/files/media/hmwk/hsp_2016-2020.pdf. Accessed 12 July 2019
HRK (2019) Download von Hochschulliste. https://www.hochschulkom-pass.de/hochschulen/downloads.html. Accessed 21 Jan 2022
Inkar (2023) Indikatoren und Karten zur Raum- und Stadtentwicklung. https://www.inkar.de. Accessed 20 July 2023
Kaluza P, Kölzsch A, Gastner MT, Blasius B (2010) The complex network of global cargo ship movements. J R Soc Interface 7:1093–1103. https://doi.org/10.1098/rsif.2009.0495
Kang C, Liu Y, Guo D, Qin K (2015) A generalized radiation model for human mobility: spatial scale, searching direction and trip constraint. PLoS ONE 10(11):e143500. https://doi.org/10.1371/journal.pone.0143500
Kauder B, Potrafke N (2013) Government ideology and tuition fee policy: evidence from the German states. CESifo. Working Paper: 4205
KIT (2018) Geschichte – Forschungszentrum und Universität: Pioniere in Forschung und Lehre. http://www.kit.edu/kit/geschichte.php. Accessed 21 Janaury 2022
Kitagawa F, Marzocchi C, Sánchez-Barrioluengo M, Uyarra E (2022) Anchoring talent to regions: the role of universities in graduate retention through employment and entrepreneurship. Reg Stud 56(6):1001–1014
KMK (2005) Prognose der Studienanfänger, Studierenden und Hochschulabsolventen bis 2020. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2005/2005_10_01-Studienanfaenger-Absolventen-2020.pdf. Accessed 21 Janaury 2022
KMK (2012) Vorausberechnung der Studienanfängerzahlen 2012–2025. Fortschreibung. Stand, 24, 2012. https://www.kmk.org/fileadmin/Dateien/pdf/Statistik/Vorausberech-nung_der_Studienanfaengerzahlen_2012-2025_01.pdf. Accessed 21 Janaury 2022
Kodrzycki YK (2001) Migration of recent college graduates: evidence from the national longitudinal survey of youth. N Engl Econ Rev 1–2:13–34
Krabel S, Flöther C (2014) Here today, gone tomorrow? Regional labour mobility of German university graduates. Reg Stud 48(10):1609–1627
Kriesch LJ (2023) Web Mining und Natural Language Processing als methodisches Komplement in der Wirtschaftsgeographie. https://jlupub.ub.uni-giessen.de/handle/jlupub/16306. Accessed 4 Aug 2023 (Diss. Univ Giessen)
Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: a model for inter-city telecommunication flows. J Stat Mech Theory Exp. https://doi.org/10.1088/1742-5468/2009/07/L07003
Lenormand M, Huet S, Gargiulo F, Deuant G (2012) A universal model of commuting networks. PLoS ONE 7:45985. https://doi.org/10.1371/journal.pone.0045985
Li Z (2022) Extracting spatial effects from machine learning model using local interpretation method: an example of SHAP and XGboost. Comput Environ Urban Syst 96:101845
Litmeyer M, Gareis P, Hennemann S (2023) Comparing student mobility pattern models. Eur J Geogr 14(1):21–34. https://doi.org/10.48088/ejg.m.lit.14.1.21.34
Liu E, Yan X (2019) New parameter-free mobility model: opportunity priority selection model. Phys A 526:12102. https://doi.org/10.1016/j.physa.2019.04.259
Liu Y, Just A, Mayer M (2021) Package 'SHAPforxgboost
Lörz M (2008) Räumliche Mobilität beim Übergang ins Studium und im Studienverlauf: Herkunftsspezifische Unterschiede in der Wahl und Nachhaltigkeit des Studienortes. Bild Erzieh 61:413–436
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Curran Associates, pp 4768–4777
Lundberg SM, Erion GG, Lee S-I (2018a) Consistent individualized feature attribution for tree ensembles. arXiv preprintarXiv:1802.03888
Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, Liston DE, Low DK-W, Newman S-F, Kim J, Lee S-I (2018b) Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed 2:749. https://doi.org/10.1038/s41551-018-0304-0
Marinelli E (2013) Sub-national graduate mobility and knowledge flows: an exploratory analysis of onward- and return-migrants in Italy. Reg Stud 47(10):1618–1633
Masucci A, Serras J, Johansson A, Batty M (2013) Gravity versus radiation models: on the importance of scale and heterogeneity in commuting flows. Phys Rev E 88:22812. https://doi.org/10.1103/PhysRevE.88.022812

McCune B, Grace J (2002) Analysis of ecological communities. MjM Software Design, Gleneden Beach
Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
Montgomery M (2002) A nested logit model of the choice of a graduate business school. Econ Educ Rev 21(5):471–480. https://doi.org/10.1016/S0272-7757(01)00032-2
Morton A, Piburn J, Nagle N (2018) Need a boost: a comparison of traditional commuting models with the XGboost model for predicting commuting flows (short paper). GIScience. https://doi.org/10.4230/LIPIcs.GISCIENCE.2018.51
Multrus F, Majer S, Bargel T, Schmidt M (2017) Studiensituation und studentische Orientierungen. 13. Studierendensurvey an Universitäten und Fachhochschulen. Bundesministerium fuer Bildung und Forschung (BMBF)
Nutz M (1991) Räumliche Mobilität der Studierenden und Struktur des Hochschulwesens in der Bundesrepublik Deutschland
Raab J, Knobem J, Aufurth L, Kaashoek B (2018) Going the distance: the effects of university—secondary school collaboration on student migration. Pap Reg Sci 97(4):1131–1149. https://doi.org/10.1111/pirs.12288
Rahman MS, Chowdhury AH (2022) A data-driven eXtreme gradient boosting machine learning model to predict COVID-19 transmission with meteorological drivers. PLoS ONE 17(9):e273319
Ren Y, Ercsey-Ravasz M, Wang P, González MC, Toroczkai Z (2014) Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat Commun 5:5347. https://doi.org/10.1038/ncomms6347
Roscher R, Bohn B, Duarte MF, Garcke J (2020) Explainable machine learning for scientific insights and discoveries. IEEE Access 8:42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199
Sá C, Florax RJ, Rietveld P (2004) Determinants of the regional demand for higher education in the Netherlands: a gravity model approach. Reg Stud 38(4):375–392. https://doi.org/10.1080/03434002000213905
Sá C, Tavaresc DA, Justinod E, Amarale A (2011) Higher education (related) choices in Portugal: joint decisions on institution type and leaving home. Stud High Educ 36(6):687–703. https://doi.org/10.1080/03075071003725534
Shapley LS (1953) A value for n-person games. In: Contributions to the theory of games, vol 2, pp 307–317
Simini F, Gonzalez MC, Maritan A, Barabasi AL (2012) A universal model for mobility and migration patterns. Nature 484:96–100. https://doi.org/10.1038/nature10856
Soerensen T (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biol Skr 5:1–34
Spadon G, Carvalho A, Rodrigues-Jr JF, Alves L (2019) Reconstructing commuters network using machine learning and urban indicators. Sci Rep 9:11801. https://doi.org/10.1038/s41598-019-48295-x
Spiess C, Wrohlich K (2010) Does distance determine who attends a university in Germany? Econ Educ Rev 29(3):470–479. https://doi.org/10.1016/j.econedurev.2009.10.009
Statistisches Bundesamt (2017) Fachserie 11 Reihe 4.1. Bildung und Kultur. Studierende an Hochschulen. Wintersemester 2007/2008 – Wintersemester 2016/2017. Statistisches Bundesamt, Wiesbaden
Teichert C, Niebuhr A, Otto A, Rossen A (2020) Work experience and graduate migration: an event history analysis of German data. Reg Stud 54(10):1413–1424
Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT (2006) Synchrony, waves, and spatial hierarchies in the spread of influenza. Science 312:447–451. https://doi.org/10.1126/science.1125237
Vrontis D, Thrassou A, Melanthiou Y (2007) A contemporary higher education student-choice model for developed countries. J Bus Res 60(9):979–989. https://doi.org/10.1016/j.jbusres.2007.01.023
Walsh S, Flannery D, Cullinan J (2018) Analysing the preferences of prospective students for higher education institution attributes. Educ Econ 26(2):161–178. https://doi.org/10.1080/09645292.2017.1335693
Weisser R (2019) How personality shapes study location choices. Res High Educ. https://doi.org/10.1007/s11162-019-09550-2
Winters M (2011) Studium und Studienreform im Vergleich der Bundesländer. In: Pasternack P (ed) Hochschulen nach der Föderalismusreform, pp 215–280
Wissenschaftliche Dienste des Deutschen Bundestages (2006) Der Studentenberg – Kollaps der Universitäten oder Illusion? Ein kritischer Beitrag zur aktuellen Diskussion. https://www.bundestag.de/resource/blob/418880/251afebe1c84c24d81ad39c8bbf34334/WD-8-212-06-pdf-data.pdf. Accessed 21 Janaury 2022

Yan X-Y, Zhao C, Fan Y, Di Z, Wang W-X (2014) Universal predictability of mobility patterns in cities. J R Soc Interface 11:20140834. https://doi.org/10.1098/rsif.2014.0834
Yuan J (2023) Package 'xgboost