Contents lists available at ScienceDirect

Applied Geography

journal homepage: www.elsevier.com/locate/apgeog

Learning income levels and inequality from spatial and sociodemographic data in Germany

Oana M. Garbasevschi [a,b,*], Hannes Taubenböck [b,c], Paul Schüle [a,d], Julia Baarck [a,d], Paul Hufe [e,f], Michael Wurm [b], Andreas Peichl [a,d,f,g]

[a] *Ifo Institute — Leibniz Institute for Economic Research at the University of Munich, 81679, Munich, Germany*
[b] *German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), 82234, Weßling, Germany*
[c] *Julius-Maximilians University Würzburg (JMU), Germany*
[d] *Ludwig Maximilian University of Munich (LMU), Germany*
[e] *University of Bristol, Bristol, BS8 1TU, UK*
[f] *IZA, Germany*
[g] *CESifo, Germany*

ARTICLE INFO

Handling Editor: Dr. Y.D. Wei

Keywords:
Regional inequality
Residential segregation
Pre-tax income
Machine learning

ABSTRACT

This study explores the potential of predicting income inequality and income levels from attributes of the built, natural and social environment in Germany. Furthermore, it investigates differences in explanatory variables and estimation accuracy for municipalities with different social and spatial structure profiles. We use income tax data, the 2011 national census, and spatial data from various sources. The explanatory variables capture the spatial variation within the area of interest of characteristics of both the residents and the living environment. Our models explain 54% of the variability in inequality and 73% of the variability in median income levels for a sample of municipalities covering 97% of the country's population. Performance increases for the subsample of municipalities with at least 10,000 inhabitants, attaining 63% for inequality and 80% for income levels. Income inequality and top incomes are better identified in Western, urban, or central locations, while median income is best estimated in Eastern, rural and peripheral locations. The most important predictors are derived from attributes such as nationality, religious affiliation, household composition, residence construction year, as well as the size and density of residences and overall building stock. Our findings further the idea that the joint spatial analysis of population and the built environment can greatly improve our understanding of socioeconomic phenomena—at regional and local levels—beyond conventional data sources.

1. Introduction

In the last decades, income inequality has been increasing in most countries around the globe, particularly due to rising income levels at the top of the distribution: in 2020, the world's top 10% of income earners received more than half of the total global income, while the bottom half of all earners earned only 7% of the total global income (Chancel & Piketty, 2021). In Germany, the top 10% income share increased from little more than 30% in 1980 to more than 40% in 2014 (Bartels, 2019). Attention to the spatial component of inequality is also amplifying. In Europe, income differences between regions in a given country have been growing since 1990 (Rosés & Wolf, 2021). Furthermore, regional differences in Europe are reflected not only in income, but also in terms of urbanization, productivity, innovation potential, employment and education opportunities (Diemer, Iammarino, Rodríguez-Pose, & Storper, 2022; Rodríguez-Pose, Iammarino, & Storper, 2018). These widening gaps foster concerns that regions and people are being left behind.

The regional level of government in EU countries is increasingly acquiring executive capacities in terms of legal, economic, and financial policy-making (Wegschaider, Gross, & Schmid, 2023). Transparent information about local income and inequality indicators reinforces regionally targeted policies by better identifying regions that drive national income inequality, allowing the study of spatial spillovers and the evaluation of policy effectiveness at local levels (De Nicolò, Ferrante, & Pacei, 2022). In Germany, the "central place" concept—the capacity to

[*] Corresponding author. Ifo Institute - Leibniz Institute for Economic Research at the University of Munich, Poschingerstr. 5, München, 81679, Germany.
E-mail address: garbasevschi@ifo.de (O.M. Garbasevschi).

https://doi.org/10.1016/j.apgeog.2023.103058
Received 15 December 2022; Received in revised form 14 July 2023; Accepted 30 July 2023
Available online 12 August 2023

"provide services and infrastructure for the surrounding regions" (Schmidt, Li, Carruthers, & Siedentop, 2021)—is at the core of municipality land development. This means that municipalities receive federal funds and have the power to implement new developments in proportion to their "central place" status (Schmidt et al., 2021). Up-to-date income reporting can enable local leaders to raise awareness of municipal problems and financial needs, either in governmental coalitions or through media coverage, which can have positive social and economic effects locally (Gross & Krauss, 2021).

A large literature shows that people care strongly about their income position relative to their neighbors (Dittmann & Goebel, 2010; Luttmer, 2005; Perez-Truglia, 2020) and municipality-level income distribution is more consequential for income-related well-being than the rank in the national income distribution, as a recent Finnish study shows (Xu et al., 2023). Moreover, high inequality and deprivation at local levels are associated with increased political polarization (Dorn, Fuest, Immel, & Neumeier, 2020) and a destabilizing effect both politically and socially. Furthermore, since people draw assumptions on the income distribution only from limited observations (Marandola & Xu, 2021), perceived and actual income inequality can differ significantly, and have a negative effect on popular support for redistribution policies (Windsteiger, 2022)—policies which could actually benefit people in various income groups. The dissemination of localized data concerning income inequality or income distribution can, in part, attenuate these misperceptions (Windsteiger, 2022) and their negative consequences.

Unfortunately, income data at sub-national levels is often unavailable for many countries, including Germany. Access to microdata from tax records is often restricted and published statistics exist for the national level only (Bartels & Metzing, 2019). While access to survey data is less restricted, the sample size in household surveys like the German Socio-Economic Panel (Goebel et al., 2019) is too small to compute inequality statistics that are representative at the local level. Furthermore, using survey observations to adjust income estimates in small geographic areas requires complex statistical procedures, which are prone to bias (Molina, Corral, & Nguyen, 2022).

In this paper, we provide one potential remedy for such data constraints. We do so by showing how we can learn about municipal income distribution using machine learning methods and open data sources. The selected data sources are in generally easily available for different regions and countries. For all municipalities in Germany, we combine geographically fine-grained spatial data with sociodemographic attributes from the census to predict income levels and income inequality indices such as the Gini coefficient. Conceptually, we establish a relationship between the spatial variation of income, patterns of the natural and built environment, and the distribution of population groups. We integrate this concept into a regional analysis, in which we distinguish regions based on geographical and administrative considerations—East/West Germany and by federal state; based on demography—population size and population density (rural/urban regions); and based on a combination of demographic and socioeconomic factors (peripheral/central regions).

Machine learning has the potential to fill gaps in the estimation of local income levels or inequality. Supervised learning procedures allow the prediction of unknown variables after training on a set of known values and covariates. Continuous refinements over the past years have led to advancements in model complexity and accuracy, an enhanced ability to capture both linear and non-linear relations, and increasingly interpretable models (Aria, Cuccurullo, & Gnasso, 2021; Li, 2022; Wójcik & Andruszek, 2021). A growing body of literature is exploring machine learning techniques that can robustly estimate the effect of diverse factors on inequality (Brunori, Hufe, & Mahler, 2021; Salas-Rojo & Rodríguez, 2022, pp. 27–51).

An important resource for estimating regional income levels and income inequality with machine learning is open spatial data with high spatio-temporal resolution, like remote sensing data and its derivatives. In this sense, two different strands of related literature can be identified: First, some studies aim to provide generalized solutions for extended geographical areas and focus on regions with data scarcity, where official census and survey statistics are difficult to collect (Donaldson & Storeygard, 2016). These studies explore diverse socioeconomic outcomes like household income, wealth, employment opportunities, and GDP, and rely on global remote sensing mapping products and crowd-sourced data: nighttime lights emissions (Ivan, Holobâca, Benedek, & Török, 2019), daytime satellite imagery(Chen et al., 2021; Feldmeyer, Meisch, Sauter, & Birkmann, 2020), land use/land cover or points of interest (Chen et al., 2021; Feldmeyer et al., 2020). The main goal is to obtain a proxy for income, which can then be used for policy measures focused on reducing poverty or increasing the financial resilience of the most vulnerable population categories. While the underlying factors are sometimes obscured by the combination of black box machine learning models and image processing techniques, some determinants of income or wealth can be identified. In studies relying on nighttime lights, the channel appears to be straightforward: areas with higher economic activity show increased use of electricity and thus higher values of measured light emissions. In studies using satellite daytime imagery, income is proxied by land use (Yeh et al., 2020). The density of points of interest—such as transportation nodes, nature or recreation sites—can predict the attractiveness of a municipality, in terms of migration balance and employment opportunities (Feldmeyer et al., 2020).

Second, some studies rely on detailed spatial data, including topographic maps, building models and infrastructure networks (Sapena, Ruiz, & Taubenböck, 2020; Wurm et al., 2019), to explore the complexity of the urban fabric and enable a rich characterization of human settlements' structure. This addresses a broad range of socioeconomic issues, from transport to retail prices to social inequalities. However, these studies often have a narrow geographical focus or deal with only a limited number of locations, such as neighborhoods within a city, large cities within a country or metropolitan areas across continents. Overall, few studies explicitly focus on the prediction of income variables (Casali, Aydin, & Comes, 2022), as we do in this paper.

Combining characteristics of the population and the spatial structure of their surroundings is a growing area of analysis of spatial inequalities (Nijman & Wei, 2020; Patias, Rowe, & Arribas-Bel, 2023). Vulnerable population categories such as migrants, senior citizens or mono-parental households are affected disproportionately by localized factors such as environmental pollution (Rüttenauer & Best, 2021), access to services (Nicoletti, Sirenko, & Verma, 2022), or housing costs (Bartels & Schröder, 2020; Lozano Alcántara & Vogel, 2021, pp. 1–19). Frieden, Peichl, and Schüle (2023) are the first to characterize income inequality at the municipality level in Germany. Helbig and Jähnen (2018) document growing segregation of residential areas along the dimension of social class and age in large German cities, based on administrative information on the composition of schooling districts and the primary residence of social assistance recipients. Also for Germany, Goebel & Hoppe (2015) found effects of both ethnic and social segregation on the persistence of poverty. While social segregation in Europe is relatively low compared to the rest of the world, it is growing (Tammaru et al., 2021), resulting in deeper socioeconomic inequalities. Analyzing the interconnections between individuals and the space they live and work in enables a broader understanding of income inequalities, both at the local and regional level.

Finally, we contribute in three ways to existing research on the estimation of local income levels and inequality. First, we investigate how reliably spatial and sociodemographic variables can predict income levels and inequality at the municipality level. The large sample size enables the evaluation of model generalization, and allows the transfer of acquired insights between different types of municipalities. Second, we differentiate between regions to identify features strongly associated with income levels and inequality and to distinguish patterns in the strength of their relationship with income variables. Third, we discuss these patterns in light of policy implications for regional and local decision makers.

2. Study area and data

Germany is the largest economy in Europe in terms of GDP, the second country by population size and the fifth by area size. The German government is a two-tier federal system, the Federation (Bund) and the Federal States (Länder). The 16 federal states are largely autonomous, have their own constitution and govern the territorial and institutional frame of the local governments. At the lowest level of the local government, federal states are divided into municipalities, currently 10,784. The term municipality denominates four types of settlements: large, medium-sized and small cities, and rural municipalities (Bundesinstitut für Bau- Stadt-und Raumforschun (BBSR), 2017). Based on population density, municipalities are classified as either urban or rural, with the majority of the population (69%) living in urban areas. An alternative type of settlement classification, which includes job availability as a socioeconomic dimension (Bundesinstitut für Bau- Stadt-und Raumforschun (BBSR), 2018), differentiates between central and very central municipalities as areas with high concentration of both population and jobs, and peripheral and very peripheral municipalities as areas defined based on the distance from central municipalities. While the delineation rural/urban or peripheral/central is debatable and open to refinement (Küpper, 2016; Taubenböck et al., 2022), it provides a meaningful classification for the analysis of regional inequalities. Other important regional characteristics are population size, the geographic delineation between the former Eastern and Western Germany, and differentiation by federal state.

2.1. Income data

We used income data from tabulated tax records at the municipality level for the year 2016. These high-quality administrative data provide a reliable source of information of the local income distribution in all German municipalities. As administrative data cannot suffer from survey non-response, the data are highly accurate also for high-income individuals. In contrast, low-income individuals are covered less reliably, as not all adult individuals in Germany file a tax return. For example, a certain share of the 5 million workers with mini-jobs—with a monthly income of up to 450 euros (Drechsel-Grau et al., 2022)—are not included in the data. The data were obtained by filing individual requests to the Statistical Offices of the German federal states. The

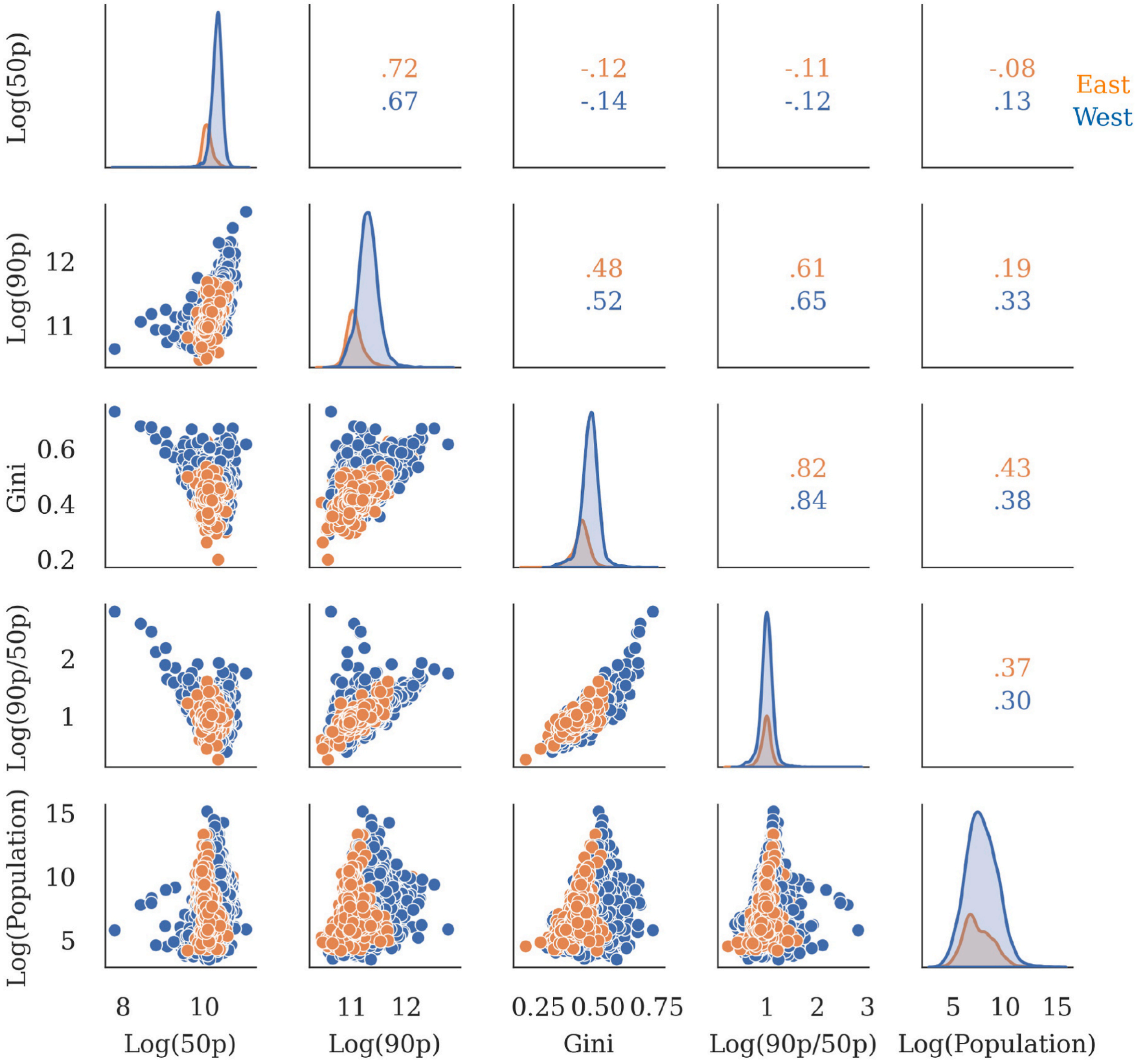

Fig. 1. Pairwise correlations between income variables and population size, for subsamples of Eastern municipalities (2,258) and Western municipalities (8,077). Variables are log-transformed and extreme values are not included, for readability purposes. Panels on the diagonal illustrate the kernel density estimation for the distribution of each variable and subsample. Values in the panels above the diagonal are the Pearson correlation.

tabulated data contain information on the sum of gross pretax income for all taxpayers within two income thresholds. Thus, average incomes per income bracket and per municipality can be determined. For privacy reasons, gross income and number of taxpayers are not reported in some income brackets. We imputed these missing values with procedures described in Appendix A. We then used generalized Pareto interpolation to estimate the income distribution in 100 percentiles (Blanchet, Saez, & Zucman, 2022). Our final sample consists of 10,335 municipalities.

Based on these data, we computed two inequality measures: the Gini coefficient and the ratio of the 90th to the 50th income percentile (90p/50p ratio), which measures inequality in the upper tail of the income distribution. The Gini coefficient is a standard inequality indicator, invariant to scale and ranging from 0 (perfect equality) to 1 (perfect inequality). Further variables of interest include the 50th (median income) and 90th income percentiles. Since the distribution of both median and 90th percentile income are right skewed, we log-transformed all income-derived variables, except for the Gini, prior to model building. The Gini coefficient takes values between 0.20 and 0.73 and the 90p/50p ratio values between 1.25 and 16.75. Gini and the 90p/50p ratio are highly correlated (r = 0.79) at the municipality level, while the Gini and the 90th percentile of income are moderately correlated (r = 0.60). There is no connection between inequality and median income: high levels of inequality are associated with both high and low median incomes, as illustrated in Fig. 1.

The spatial distribution of median income and Gini inequality in Germany is shown in Fig. 2, while corresponding maps of the 90th percentile income and the 90p/50p ratio are included in Appendix B. In terms of income levels, there is a clear divide between the relatively poorer East and more affluent municipalities in West Germany, particularly in Bavaria and Baden-Württemberg. Eastern municipalities however, exhibit lower levels of income inequality. The 5% of municipalities with the lowest Gini values are municipalities with less than 1000 inhabitants, located mainly in the Eastern states of Mecklenburg-Western Pomerania and Thuringia, and the Western state of Schleswig-Holstein. The 5% of municipalities with the highest Gini coefficients (516 municipalities) are municipalities of all sizes, out of which 25% have less than 1000 inhabitants, and 25% have more than 10,000 inhabitants. The municipalities with the highest inequality are spread among different states, but are located almost exclusively in West Germany.

Median and top incomes display a moderately high spatial autocorrelation pattern (global Moran's I statistic of .6, $p - value < 2.2e - 16$). The Gini and the 90p/50p ratio display a moderate spatial autocorrelation (global Moran's I statistic of .38, $p - value < 2.2e - 16$). As such, it appears to be the case that municipalities are more likely to have other municipalities with a similar level of income, rather than inequality, in their vicinity. In Appendix B, we illustrate the spatial autocorrelation relationship for median income and the Gini, highlighting the states with the most distinguishable patterns. Bavaria in particular contains many municipalities that are surrounded by municipalities with higher than average inequality. Instead, in most of Mecklenburg-Western Pomerania instead, a pattern of low-low inequality is observed. Rhineland-Palatinate and Schleswig-Holstein have the most heterogeneous distribution of inequality.

2.2. Spatial and sociodemographic data

Sociodemographic population attributes were extracted from the 2011 German Census (Destatis, 2011). The Census provides information on, among others, age, nationality, family size and composition, size of

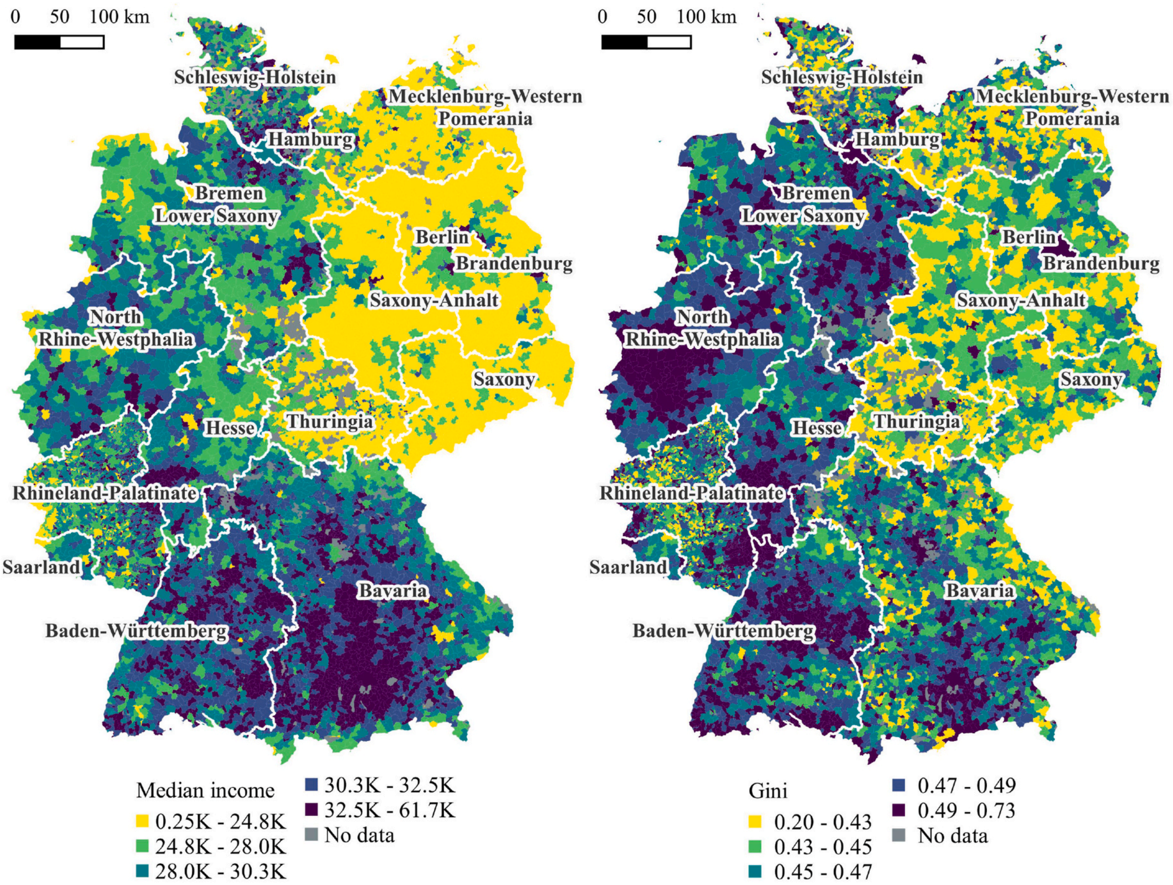

Fig. 2. Gini and median income values for German municipalities. Variables are color-coded based on the distribution quintiles. The labels and white-colored boundaries indicate the 16 federal states.

residence, home-ownership status, and type and construction year of residential buildings. A detailed description of each variable is provided by Destatis (2011). The data are aggregated to the municipality level, and in 1 km × 1 km or $100m \times 100m$ areas defined by the INSPIRE geographical grid, a Pan-European standard for equal-area grid systems (Bundesamt für Kartographie und Geodäsie (BKG), 2019).

The first type of spatial data used in our analysis is land use/land cover information extracted from remote sensing data. We used a fine-grained classification of land cover in Germany, based on the Copernicus LUCAS (Land Use/Cover Area Frame Survey) reference dataset (Weigand, Staab, Wurm, & Taubenböck, 2020). Each $10m \times 10m$ area is labeled with one of seven classes: artificial land—land assigned for urban and economic purposes, occupied by buildings and infrastructure, open soil, water areas, and four different classes of vegetation. We complemented this high-resolution dataset with the CORINE (Coordination of Information on the Environment) European land use map, a 100m resolution classification that comprises 42 classes (Büttner, 2014). The land use classes mapped with CORINE include agricultural areas, forests, three urban classes – continuous, discontinuous, and urban green – as well as areas defining roads, rails, and airports. Additionally, we used nighttime light emission data from the Visible Infrared Imaging Radiometer Suite with the Day and Night Band (VIIRS/DNB). The VIIRS/DNB sensor produces since 2012 images with a daily coverage and a spatial resolution of 740m. Due to high sensitivity to lower light levels, the data are especially useful for mapping urban areas, since urban areas are brighter than the rural surroundings. We used the annual values for 2016 (Earth Observation Group (EOG), 2021), which is a composite aggregated from monthly cloud-free data (Elvidge, Zhizhin, Ghosh, Hsu, & Taneja, 2021).

The spatial resolution of available land use data does not allow for further differentiation of land use within urban areas. For a more refined view of the built environment, we have thus relied on the building dataset for Germany provided by the German Federal Agency for Cartography and Geodesy (Bundesamt für Kartographie und Geodasie (BKG), 2021). It consists of LoD-1 (level-of-detail 1) building data: geometry of the building's ground floor and building height. Building functions are recorded as codes with associated text labels (Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 2020). We grouped the 980 individual labels into 20 functional classes including: residential, residential annexes, retail and services, health care, schools, agriculture, office buildings, factory, and industrial. In the absence of a universally accepted ontology of building functions, we followed the classifications referenced by the TABULA residential energy project (Loga, Diefenbach, Stein, & Born, 2012). The process of recording building function in the official data is unfortunately not consistent in all federal states, which led to missing or potentially inaccurate labels. Residential buildings are identified in 89% of all municipalities, while identification varies for mixed non-residential buildings: 62% of municipalities have buildings of the type of school, 55% have retail and services buildings, and 30% have health care buildings.

3. Method

The within-location variation of municipality attributes were defined as follows: variables were computed for individual areas of $100m \times 100m$ and 1 km × 1 km, and then aggregated to the municipality level. Additional variable creation methods consist of segregation indices, combinations of social groups and spatial attributes, and localized spatial attributes, as summarized in Table 1. A snapshot of the types of spatial data used, and their spatial resolution, is illustrated in Fig. 3. Income variables were estimated from the resulting features using regression models. We differentiated between models that incorporate only spatial features (created using all data except the census), and models which include all features. By this, we aimed to evaluate the potential of standalone spatial data sources to estimate income variables. The following two subsections describe the variable creation methods, while the final subsection introduces the regression methods used and the general setup.

Table 1
Feature descriptions and types. For feature notation, * represents: (1) a segregation index; (2) minimum (MIN), maximum (MAX), standard deviation (SD), average (AVG), range (RNG), interquartile range (IQR), coefficient of variation (CV); (3) selected neighborhood statistics.

Feature type	Year	Spatial scale	Description	Notation
Income, inequality	2016	Municipality	–	50p, 90p, Gini, 90p/50p
Sociodemographic	2011	$100m \times 100m$ 1 km × 1 km municipality	Segregation indices, Area statistics	Index $*_{\text{spatial scale}}$(Variable)
Land use	2015 2016	$100m \times 100m$ 1 km × 1 km municipality	Gini inequality, Area statistics	
Nighttime lights	2016	1 km × 1 km municipality	Gini inequality, Area statistics	
Buildings	2015	$100m \times 100m$ 1 km × 1 km municipality	Form and function statistics, Area statistics	

3.1. Segregation indices

The census data is published at a spatial resolution of $100m \times 100m$ and 1 km × 1 km. This enabled us to investigate residential segregation at multiple spatial scales, with the aim of relating disparities in income to disparities in the distribution of social groups. Residential segregation expresses how spatially separated two social groups are within census tracts (Massey & Denton, 1988). Segregation indices are multi-faceted constructs that can either measure the degree of separation of a single population subgroup, with respect to the rest of the population, or the degree of separation between multiple population subgroups. Furthermore, single group indices can be subcategorized into aspatial and spatial indices, where aspatial indices rely solely on the division of the population in census tracts, and spatial indices also include information on the spatial relationships between tracts.

The segregation indices investigated cover all dimensions of segregation defined by Massey and Denton (1988): evenness, exposure, concentration, centralization, and clustering. Using the segregation module of the *PySal* Python library (Cortes, Rey, Knaap, & Wolf, 2020), we computed an extensive number of indices, out of which the following were identified as important variables in the relationship with income and inequality: multi-diversity, multi-divergence, distance decay isolation and interaction, Simpsons' concentration and interaction indices. Multi-diversity is an expression of group proportions, measured by the Theil's Entropy index, and ranging from 0, i.e. all individuals are members of the same group, to 1, i.e. an even distribution of individuals across groups (Reardon & Firebaugh, 2002). Multi-divergence is the difference between the overall proportion of groups in the entire area and the proportion of a group in local areas (Roberto, 2015). Higher differences between local and overall proportions signify greater segregation. Distance decay indices compute closeness between social groups in space. Distance decay isolation is the probability that the next person a group member meets anywhere in space is from the same group, while distance decay interaction is the probability of meeting members of the other groups (Morgan, 1983, pp. 211–217). Concentration is represented by the relative amount of physical space occupied by a minority group in an area. Interaction is the opposite concept and is measured as the probability that two individuals chosen at random and independently from the population will not belong to the same group (Reardon & Firebaugh, 2002). Higher interaction values correspond to lower segregation levels.

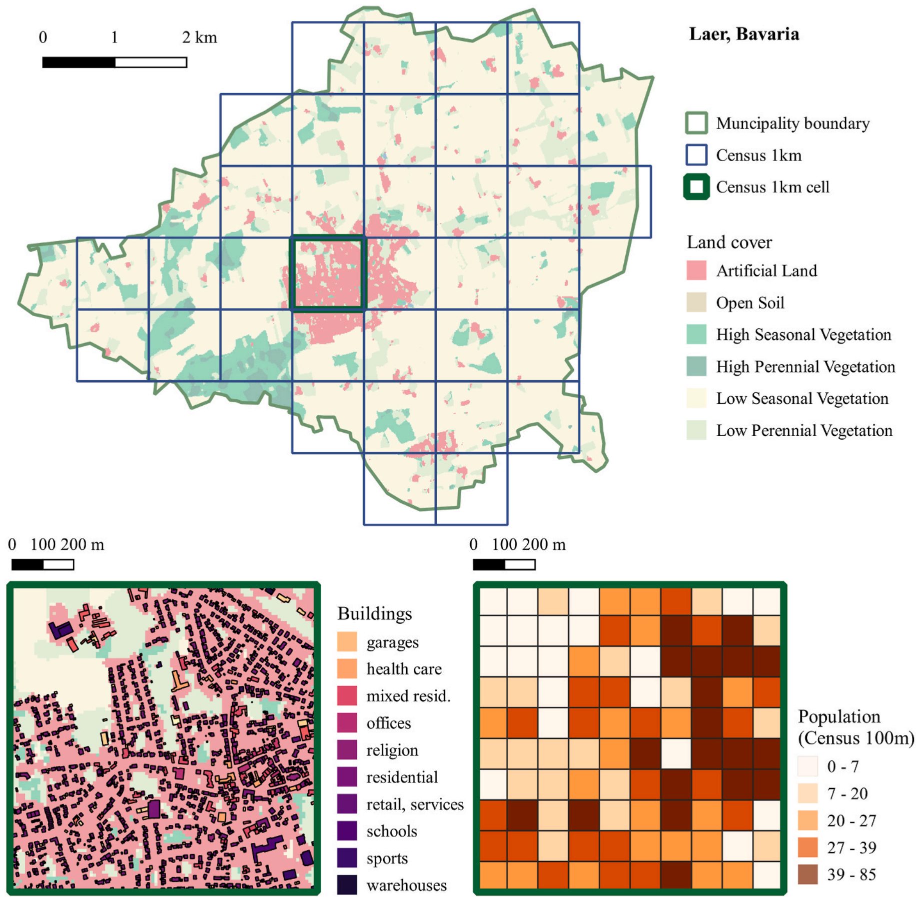

Fig. 3. Overview of different data used and their spatial resolutions, illustrated for the municipality of Laer, Bavaria (top) with a focus on an inner urban area (bottom). Data includes building footprints (Bundesamt für Kartographie und Geodasie (BKG), 2021) and functions (Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 2020), Census data (Bundesamt für Kartographie und Geodäsie (BKG), 2019; Destatis, 2011) and land cover data (Weigand et al., 2020).

The concept of residential segregation is commonly applied to population subgroups defined by demographic or socioeconomic attributes. Following this principle, we constructed indices for sociodemographic attributes expected to be predictive of household income: age, family composition, and households with or without seniors. Nationality and affiliation to the two largest religious denominations, Roman Catholic and Protestant, were also considered. We extended this segregation analysis to include the urban structure, under the premise that the spatial separation of different types of residential forms can be related to the diversity in residents' income levels. At the level of residence units, we considered size and ownership status, and at the level of the individual building, construction year and building type, e.g. single-family houses versus apartment blocks. In Appendix C, we list the groups of sociodemographic variables used to compute segregation indices.

3.2. *Spatial variation indices*

For both types of land cover classification, we computed total areas, and area statistics per capita, or per extent of built-up areas. The method was applied for the entire municipality, and for 100*m* × 100*m* and 1 km × 1 km unit areas. All per capita statistics were computed for the subset of unit areas where population counts are non-zero. The variation of values computed for unit areas within each municipality's boundary was defined by summary statistics, such as minimum, maximum, mean, standard deviation, range, interquartile range, and coefficient of variation. The distribution of nighttime lights intensity was measured similarly. Furthermore, we computed Gini inequality indices for available vegetation per capita, built-up land per capita and nighttime light intensity per capita, in areas of 100*m* × 100*m* or 1 km × 1 km.

Building functions have a threefold purpose in this study: first, they enable the identification of residential buildings; second, a variety of functions can be associated with differences in occupational status and consequently the distribution of skills in the area; last, it allows the identification of accessibility to special services, like schools, healthcare, retail, or worshipping sites. Accessibility was measured as the shortest distance by road to the closest available building providing a specific service. From the GRIP road data, the primary, secondary, and tertiary road types were used (Meijer, Huijbregts, Schotten, & Schipper, 2018). Concerning urban form, we estimated variability in height, and ground and total floor area densities for different building functions, at 100*m* × 100*m*, 1 km × 1 km and municipality levels. We also computed these statistics for areas defined as non-residential, i.e. if the area of residential buildings in a neighborhood is less than or equal to 30% of the total built-up area. Here, the goal was to determine whether municipalities have a distinguishable profile in terms of industry, business, or services.

Furthermore, we computed built-up environment and sociodemographic statistics for neighborhoods with a majority population group. In Germany, individuals with a migration background and the elderly population are at high risk of income poverty after housing costs (Lozano Alcántara & Vogel, 2021, pp. 1–19). In each municipality, we therefore selected as "neighborhoods of interest"the set of 100*m* × 100*m* areas where: (1) more than 50% of inhabitants are 65 years old or older, (2) for more than 25% of inhabitants the country of birth is not Germany, (3) more than 75% of inhabitants define their religion as different from the Catholic denomination, and (4) more than 75% of inhabitants define their religion as different from the Protestant denomination. The exact thresholds for the majority population groups in a 100*m* × 100*m* area are derived based on the top decile values in the distribution of these statistics in all 100*m* × 100*m* areas over the entire country. This means that for each of the four "neighborhoods of interest" types, 10% of all 100*m* × 100*m* areas have been selected.

3.3. Random Forest and multivariate analysis

Based on the constructed spatial and sociodemographic features, the income variables of interest were predicted using Random Forest models. Random Forest is a powerful and flexible ensemble-based machine learning model (Breiman, 2001), with a widespread use in urban analytics, remote sensing, and increasingly, in social sciences (Credit, 2022; Wurm et al., 2019). Its main features include robustness to noise, computational efficiency, and the ability to handle high data dimensionality and multicollinearity of features well (Belgiu & Drăguţ, 2016). In a Random Forest variable interactions are partially and implicitly dealt with in a non-linear manner when considering different variables for node splits in the same decision tree branch (Breiman, 2001; Inglis, Parnell, & Hurley, 2022). This simplifies the model setup by circumventing the requirement for manually defined interactions, in models with large numbers of features, and insufficient prior knowledge about salient interactions. The inbuilt feature selection algorithms of the Random Forest generally show a good performance, as documented in the literature (Degenhardt, Seifert, & Szymczak, 2019; Speiser, Miller, Tooze, & Ip, 2019). We chose the method of permutation importance for feature selection, a procedure where the values of a feature are randomly permuted and subsequent decreases in model performance signal feature relevance (Breiman, 2001). The coefficient of determination R^2 was used as the main model performance indicator.

The data exploration process resulted in an initial pool of explanatory variables consisting of more than 10,000 features: 1158 from land use, 208 from nighttime lights, 50 from distance to amenities, 2295 from buildings, 2804 from sociodemographic data, 1318 from segregation and 2192 from sociodemographic and building data for selected neighborhoods. Hence, the process of feature selection to reduce the dimensionality of the data was a crucial part of model building. Our goal was to identify a diverse as possible set of features related to the target outcome. For this aim, we used an iterative forward-selection procedure, consisting of two steps: First, we built models with features of the same type, i.e. land use, building-derived and sociodemographic features, and we filtered the most relevant features. Second, we merged all best performing features and selected a final set of 20 features. We empirically selected the optimal number of features by prioritizing a minimal set of features for which the addition of new features did not significantly improve model performance. In both steps, we controlled for multicollinearity, using a Pearson's correlation coefficient equal to 0.85 as threshold. Relaxing or strengthening the multicollinearity constraint, using thresholds of 0.8 and 0.9, did not produce significantly different effects. Each feature in the final set was replaced by one of the features strongly correlated with it, filtered out at earlier stages, and for each feature permutation the model performance was tracked. No improvement in model performance was observed.

We repeated the feature selection and modelling process—starting from the same initial pool of variables—for different municipality subsamples: per federal state, per population size, East/West, rural/urban, as well as peripheral/central municipalities. Finally, for a small number of subsamples, the 20 features selected were inspected for significance, and further filtered using an OLS regression model with controls and interaction terms. We introduced variable interactions for pairs of variables that represent the same sociodemographic variable, defined at different spatial scales. An example of such an interaction term would be the share of senior citizens in a municipality taken together with the degree of spatial isolation of senior citizens, or together with the coefficient of variation of the share of senior citizens in 1 km × 1 km areas in the municipality. Only interaction terms which are statistically significant are reported in the results section and Appendix D.

4. Results

4.1. Prediction accuracy and generalization capacity

The likelihood of success in estimating unknown variables through supervised learning can be influenced by (1) choosing the right subgroups of examples from which the machine learning model can learn, and by (2) applying the estimation procedure to compatible groups of observations. For predicting income levels and inequality, we compared regression models built with all municipalities with models built with subsamples defined by a single municipality attribute: population size, East/West, rural/urban, or peripheral/central. It was expected that municipality features would exhibit less variation within-subsamples, which in turn would increase model accuracy.

Data availability, a large population concentration, and diversity of the built environment make cities prime candidates for spatial studies. The variation in population sizes in our municipality sample allowed us to investigate if prediction models can be transferred from large cities also to smaller municipalities. For this aim, we tested different population thresholds to monitor the most significant changes in model accuracy between the two resulting subsamples, as illustrated in Table 2. Increasing the population size threshold results in higher model accuracy. The result is unsurprising: higher thresholds result in increasingly homogeneous samples of municipalities, since large areas are more similar as a group—on multiple dimensions—than when compared with medium- or small-sized areas. At the lower end of population size thresholds, we found that model accuracy drops significantly at the 1000 inhabitants limit. Hence, our model is well suited to the prediction of income levels and also income inequality for all municipalities with at least 1000 inhabitants. Since this subsample represents 67% of all municipalities and 97% of the country's population, the model can therefore be applied at a national scale. Note that the selection of municipalities by population size results in higher prediction accuracies than a selection based on any of the other three criteria, as illustrated in Table 3. Models built for municipalities which are either Western, urban, or central, perform better than models for municipalities which are Eastern, rural, or peripheral.

In terms of explanatory variables, the combination of spatial and sociodemographic features produced the best results. Spatial features alone estimate income levels well, and lead to consistent results for both medium (between 1000 and 10,000 inhabitants) and big (more than 10,000 inhabitants) municipalities. In small municipalities (less than 1000 inhabitants), spatial features alone already produce similar prediction accuracies of inequality as spatial and sociodemographic features combined.

4.2. Global models and features

After showing that it is possible to select a large subsample of municipalities for which income and inequality can be accurately predicted—municipalities with a population of at least 1000 inhabitants—this section documents the most important global features. Global features are variables that can explain a large share of the variance in income levels or inequality in all types of municipalities, i.e. irrespective of their size or geographic location. To identify such global features, we built a baseline OLS model including the natural logarithm of population size and three binary control variables: east, rural, and peripheral. Two OLS models extend the baseline with statistically significant variables highlighted by the Random Forest model: first with area-wide sociodemographic variables, and second with spatial and segregation variables. In Table 4, we report model coefficients and standard errors for the Gini coefficient, while the other results are reported in Appendix D.

We find that the dichotomies between East/West, rural/urban and peripheral/central, alongside population size explain 44% of the variance in median income and 35% (15%) of the variance in the Gini

Table 2
Out-of-sample model performance (R^2) for different samples of municipalities, defined by population size. Performance evaluated with a repeated (n = 500) validation with 50% of the observations as training data. Model (1) variables include only spatial features; model (2) variables include spatial and sociodemographic features.

Dependent variable	Population size									
	$N > 0$		$N < 1,000$		$N \geq 1,000$		$N \geq 5,000$		$N \geq 10,000$	
(model)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Log(50p)	.52	.61	.33	.43	.6	.73	.63	.77	.65	.8
Log(90p)	.55	.65	.29	.35	.61	.74	.6	.78	.59	.77
Gini	.41	.45	.24	.27	.45	.54	.5	.6	.49	.63
Log(90p/50p)	.26	.31	.11	.14	.32	.43	.37	.53	.43	.59

Table 3
Out-of-sample model performance (R^2) for different samples of municipalities, defined by population size. Performance evaluated with a repeated (n = 500) validation with 50% of the observations as training data. Model (1) variables include only spatial features; model (2) variables include spatial and sociodemographic features.

Dependent variable	Sample selection											
	East		West		Rural		Urban		Peripheral		Central	
(model)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Log(50p)	.41	.6	.31	.46	.52	.59	.5	.62	.51	.60	.38	.5
Log(90p)	.37	.55	.46	.57	.52	.58	.52	.63	.47	.55	.47	.61
Gini	.30	.33	.34	.38	.35	.37	.48	.53	.32	.33	.39	.46
Log(90p/50p)	.27	.3	.27	.34	.19	.23	.35	.4	.2	.23	.32	.38

coefficient (the 90p/50p ratio). In Western Germany, the median and the 90*th* percentile of income are 23% higher than in Eastern Germany. Inequality levels, as measured by the Gini coefficient, are fairly similar on average in Eastern and Western Germany, with marginally lower values in Eastern Germany. For rural versus urban areas, the differences are higher for the 90*th* percentile than for median income.

For both income and inequality estimation, five types of important sociodemographic features were identified: age, nationality (citizenship), religious affiliation, family composition and size/type of residence. Incorporating the spatial variation of sociodemographic features improves model performance for all income variables. High median and 90*th* percentile income are associated with large housing units and single-family detached houses. Municipalities with higher top incomes have higher population shares of households with at least one child under 18-years old, and lower shares of households in which one spouse is deceased. Municipalities with higher inequality levels have high shares of citizens born in the EU27 and lower shares of citizens with German nationality. In terms of segregation, the diversity in age and religious affiliation is negatively correlated with top incomes and the 90p/50p ratio. The religious affiliation variable should be interpreted with caution. Declared non-affiliation to either the Catholic or the Protestant denominations in East Germany is associated with a higher degree of atheist beliefs than in West Germany. Consequently, the category “other“ in the religious affiliation variable cannot be associated with an ethnic group or another major religious community.

Spatial features correlate moderately with income inequality but highly with median and top income levels. Individuals with higher incomes live in municipalities with fewer buildings per capita, a lower spatial variation between neighborhoods in total and residential built-up area, and a lower spatial variation in the size of garages and other residential annexes. In these municipalities, buildings are higher and the ratio between the height of residential buildings and the rest of the buildings in a neighborhood is more uniform. High-income municipalities furthermore contain less office and administration space and fewer religious sites. In municipalities with a high degree of income inequality, there is a higher amount of artificial land per capita and a greater variation in building height. These municipalities have overall fewer green spaces, coupled with an unequal spatial distribution of seasonal vegetation per capita. High-inequality municipalities furthermore have smaller garages and residential annexes.

Table 4
Results of the regression analysis of the Gini as dependent variable, for the sample of municipalities with more than 1000 inhabitants (6932 observations). Model variables include: (1) controls, (2) sociodemographic features, (3) spatial features and sociodemographic features with a spatial component. Significance codes: ****p* < .001, ***p* < .01, **p* < .05.

Dependent variable	OLS Model		
Gini	(1)	(2)	(3)
East	−.0298***	−.052***	−.0291***
	(.0008)	(.0017)	(.0018)
Rural	−.0066***	−.0027***	−.0018**
	(.0007)	(.0007)	(.0007)
Peripheral	−.0119***	−.008***	−.0035***
	(.0008)	(.0007)	(.0007)
Log(Population)	.0069***	.0055***	.0051***
	(.0003)	(.0004)	(.0004)
CitizenshipDE		−.2223***	−.1485***
		(.011)	(.0152)
Citizenship1		.1802***	.1557***
		(.0119)	(.0137)
Age20-29		−.1621***	−.2159***
		(.0172)	(.0171)
ReligionOther		.0451***	.0239***
		(.003)	(.0031)
LivingSpace60m-100m		.0076***	−.0625
		(.0042)	(.0048)
IQR$_{100m}$(CitizenshipDE)			−.8415***
			(.0924)
AVG$_{100m}$(2Rooms)			.3219***
			(.0227)
SimpsonsConcentration$_{1km}$(Religion)			.0503***
			(.0087)
MultiDiversity$_{1km}$(Religion)			.0583***
			(.006)
AVG$_{100m}$(ArtificialLandCapita)			.6623***
			(.1084)
AVG$_{100m}$[IQR(BuildingHeight)]			.0023***
			(.0005)
CitizenshipDE x IQR$_{100m}$(CitizenshipDE)			1.053***
			(.1052)
(Intercept)	.425***	.475***	.3619***
	(.0027)	(.0084)	(.0124)
Adj.R2	.3521	.4291	.4889
R2	.3524	.4298	.4901

4.3. Regional models and features

Regional features are variables that are predictive of income level or inequality for specific regions. We estimate variable importance by constructing—starting from the same initial pool of variables—independent models for different types of municipality subsamples: municipalities with less than 1000 inhabitants; East/West, urban/rural, and peripheral/central municipalities, of all population sizes; municipalities per federal state.

As noted previously, inequality is difficult to predict for the subsample of municipalities with less than 1000 inhabitants, which includes predominantly rural and peripheral municipalities with both the highest and lowest Gini values. This holds true for models based on either spatial or sociodemographic features. The relationships between EU27 and German citizenship and inequality observed for larger municipalities still hold for small municipalities. In addition, higher shares of EU27 population are associated with lower median income, whereas residence size and more recent construction years are associated with higher incomes. Furthermore, small municipalities with lower inequality encompass more available green space per capita, less water bodies, and more agricultural sites.

Concerning the East/West, rural/urban and peripheral/central regions, we observed that differences in the correlations of important variables with median income and income inequality are most pronounced between municipalities in East and West Germany, as illustrated in Fig. 4. Residence size is strongly positively associated with median income levels in Eastern municipalities. Concerning residential buildings, Eastern municipalities with higher incomes have more new constructions, higher spatial spread of buildings built before 1978, and higher spatial concentration of owner-occupied residences. The differences in the relationship between spatial features and income is equally of interest, as illustrated in Appendix E. The size and variation in size of residential units and garages is positively correlated with income levels in the East, while being negatively correlated in the West. The peripheral/central division also allows for a good differentiation between municipalities. In contrast, the between-region differences in variable importance are less marked for the rural and urban municipalities.

Concerning individual federal states, two Eastern and two Western states are distinguished in terms of prediction accuracy. Results for all states are summarized in Fig. 5 and detailed in Appendix F. In Saxony-Anhalt, the Gini coefficient can be estimated with an above-average accuracy. Saxony-Anhalt is an Eastern state formed predominantly by municipalities with a population between 1000 and 10,000 inhabitants, and with rather low Gini coefficients. The Gini is correlated with a high concentration of buildings comprised by many individual dwellings (more than 13, as defined in the census), and with higher numbers of buildings classified as sport facilities or retail and services. Additionally, higher inequality is associated with a higher number of divorced couples and households where the mother is the single parent, as well as with lower numbers of households with children. In the Eastern state of Brandenburg, both median and top income can be estimated with significantly higher accuracy than the national average. Brandenburg is the second most sparsely populated state, with highly heterogeneous and dispersed settlements. Higher income levels are registered in areas with higher numbers of residential buildings which were constructed after 1996, and which are spatially clustered. Furthermore, high median income is associated with a high ratio of total area of residential buildings with respect to other types of buildings in the municipality.

The Western states Hessen and North Rhine-Westphalia show high estimation accuracies for both inequality measures. In Hessen, higher values of the 90p/50p ratio are associated with a higher spatial variation in the share of the EU27 population. In addition, higher inequality is associated with higher shares of population of non-EU27 origin living in the neighborhoods with the lowest share of people born in Germany. While median income is estimated with an above-average accuracy in Hesse, the reverse holds true for North Rhine-Westphalia. This can be an indicator that there are other potentially important unidentified factors related to median income levels in this state.

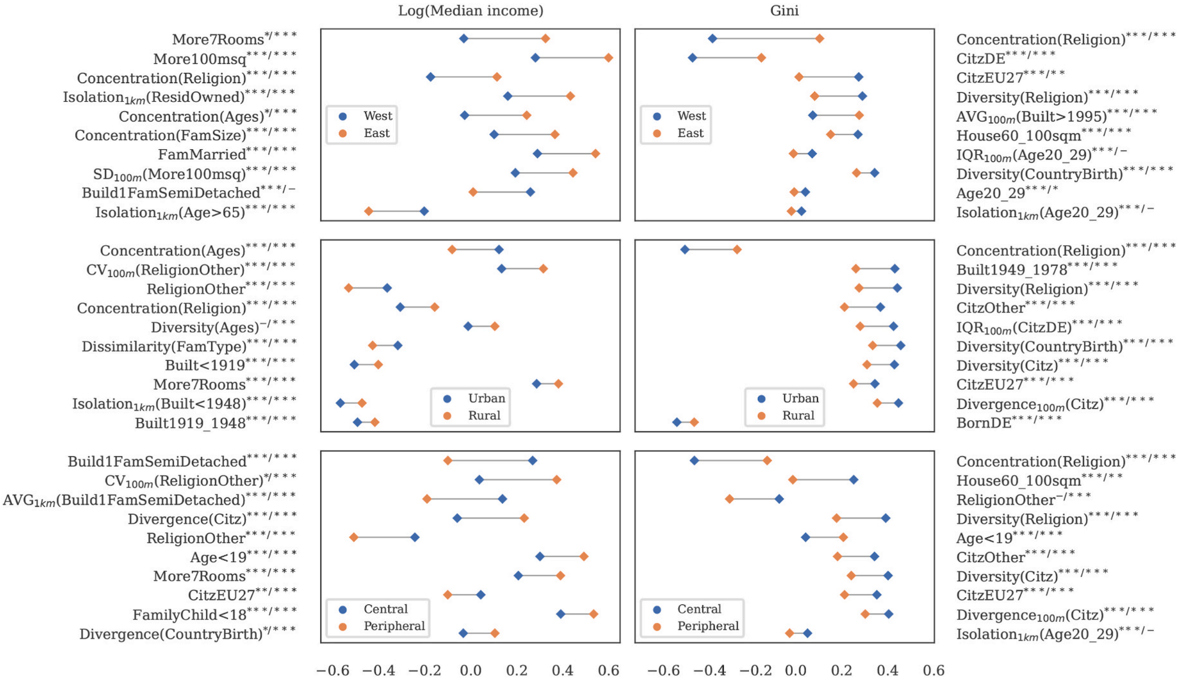

Fig. 4. Pearson correlation between median income (left) and Gini (right), and the most important sociodemographic predictors. Correlations are computed on the East/West (top), rural/urban (middle), and peripheral/central subsamples (bottom), for municipalities of all population sizes. Variables are listed in decreasing order of the absolute difference computed between correlation coefficients for each binary sample split. For all variable correlations computed, the significance level is indicated as superscript, next to the variable name, where the first value corresponds to correlation computed on the West/urban/central subsample, and the second value corresponds to the correlation computed on the East/rural/peripheral subsample. Significance codes: ***$p < .001$, **$p < .01$, *$p < .05$,-not significant.

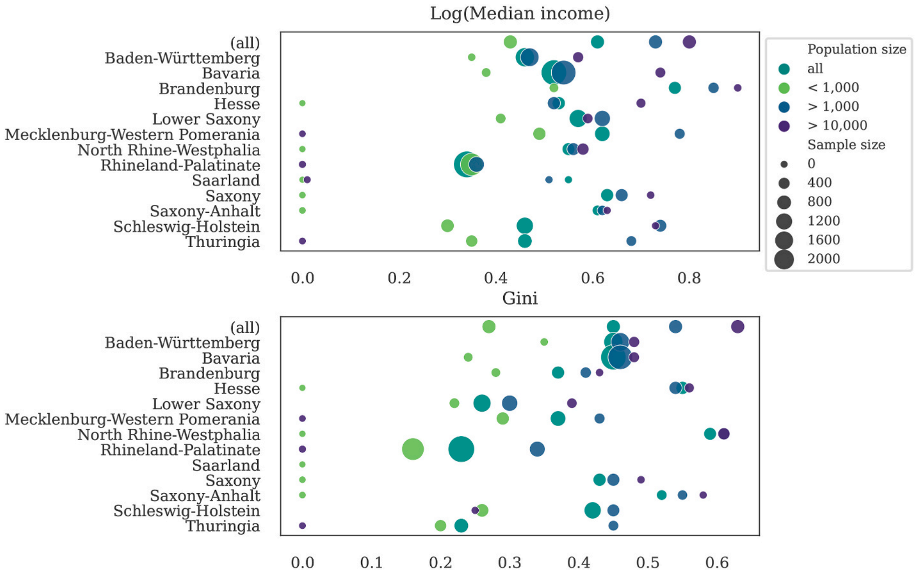

Fig. 5. For median income (top) and Gini (bottom), out-of-sample model performance (R^2) on samples of municipalities defined by federal state and population size. Performance evaluated with a repeated (n = 500) validation with 50% of the observations as training data. Models include spatial and sociodemographic predictors.

5. Discussion

With this study, we have (1) built and estimated a well-performing model to predict regional income levels and inequality mostly based on openly available data, and (2) contributed to a broader understanding of how different regions differ in their spatial and sociodemographic characteristics in relationship with income levels and inequality.

We first show that income levels and inequality can be successfully predicted through a process of supervised learning, with an accuracy of up to .80 for median income, and up to .63 for income inequality as measured by the Gini coefficient. Accuracy is highest in municipalities with more than 10,000 inhabitants, representing 74% of the German population. For municipalities with more than 1000 inhabitants, which represent 97% of the population, accuracies remain good for income (0.73) and moderately good for inequality (0.54). The success of machine learning methods in predicting inequality hence depends on the area under study, which should be considered when applying the method in a different empirical context. Our results are coherent with other studies that have predicted income inequality in comparable areas under study, and that explained between 43% and 68% of the variance in income levels (Khachiyan et al., 2022; Rolf et al., 2021; Sapena et al., 2020; Sapena, Wurm, Taubenböck, Tuia, & Ruiz, 2021). Few studies have tried to predict income inequality with the type of covariates present in our analysis (Sapena et al., 2021). The comparative advantage of our study resides in the large sample size, since except for the US studies of Khachiyan et al. (2022) and Rolf et al. (2021), none of the other studies present a nationwide analysis. Nevertheless, the different perspectives in the related literature show the significant potential of extending the variable pool with complex land or urban form features, consequently improving income and inequality estimation. Concerning regional differences in model accuracy, we observe only small differences between East and West Germany, for all but the median income. Differences are greater between rural and urban, or peripheral and central areas, which produce sub-regions that are similar in their relationship with income. Except for median income, the accuracy in estimating income variables is higher for the urban, central or Western municipalities. A promising avenue of research would be to understand more precisely why some regional characteristics are predictive for specific income variables. Furthermore, it would be interesting to apply our method at an even finer spatial level, the neighborhood.

Our second important contribution is to highlight the role of population size as the most important criterion in differentiating regions in the accurate estimation of income levels and inequality. Studies applied to urban areas with a large population have documented a positive relationship between population size and inequality (Castells-Quintana, Royuela, & Veneri, 2020; Lee, Sissons, & Jones, 2016). Our definition of “large“ and “small“ municipalities is not comparable to such studies. We show with our nationwide exploration of municipality-level pre-tax income data that income inequality—as measured by both the Gini coefficient and the 90p/50p ratio—is also found in medium-sized and small municipalities. Inequality within large cities therefore depicts only a part of the whole picture of inequality in Germany. These results are also reflected in similar studies where sub-national, albeit incomplete, samples of different-sized areas were under investigation, and where no clear relationship between income inequality and population size could be inferred (Martín-Legendre, Castellanos-García, & Sánchez-Santos, 2021). Furthermore, low model accuracy for income estimation in small municipalities points to the existence of important unobserved factors. Small municipalities thus constitute an interesting area of future quantitative research. An analysis of smaller areas is still missing in many research fields: voting behavior (Wegschaider et al., 2023), government coalition agreements (Gross & Krauss, 2021), spatial planning (Eichhorn & Pehlke, 2022), or income indicator estimation (Würz, Schmid, & Tzavidis, 2022). Our study is aligned with the recent literature advocating for open access to small-scale statistical data, and increased recognition of small areas as a distinctive typology of settlements (Academy for Spatial Research and Planning (ARL), 2019). These areas should be investigated with specific research instruments (Academy for Spatial Research and Planning (ARL), 2019), for the exploration of place-appropriate policy measures that will help fulfill the economic potential of all types of regions (Diemer et al., 2022).

Furthermore, our analysis identifies regional factors associated with income and inequality. While the analysis is primarily exploratory, and does not rely on structural models and causal inferences, we highlight in the following paragraphs possible implications for regional and local policy makers, based on a critical assessment of our findings and the relevant literature. Compared to their counterparts, rural, peripheral and Eastern municipalities with higher median incomes are more likely to feature larger residences and a higher share of married couples. Furthermore, we find here a greater spatial segregation by household size and type, in line with the general knowledge that housing preferences differ between households with and without children (Heider, 2019; Cortinovis, Geneletti, & Haase, 2022). Regional growth and employment opportunities could translate into the need to attract a diverse workforce, which implies that appropriate transit and housing options should be provided in smaller municipalities, beyond single family homes and private transportation infrastructure (Gans, 2018, pp. 375–396).

Urban, central or Western municipalities with higher inequality are associated more strongly with lower population shares of German nationals and higher segregation in nationalities and religious affiliations. This relates to previous findings that in Europe, foreign workers are attracted to large metropolitan areas (Benassi, Bonifazi, Heins, Lipizzi, & Strozza, 2020). In Germany, large Western cities have a more diverse population structure, whereas small urban areas show high degrees of ethnic segregation (Buch, Meister, & Niebuhr, 2021). The origin of foreign nationals is also important: There are higher shares of EU27 citizens in medium and large municipalities with high inequality, but also higher shares in small municipalities with low income. Immigrants that newly arrive in a host country tend to cluster in areas with a larger presence of co-ethnics (Chakraborty & Schüller, 2022). The literature on migration identifies both positive and negative effects in terms of labor market success. On the one hand, immigrants have access to informal social networks that provide education and job opportunities, and they benefit especially in situations where the local community includes well-educated individuals, with high employment rates and high wages (Chakraborty & Schüller, 2022). On the other hand, the initial boost in employment can be accompanied by future unemployment and a decreased investment in further human capital development (Battisti, Peri, & Romiti, 2022). Local governments can finance initiatives that amplify the positive effects of co-ethnic associations. For an optimal income mix, housing policies should also be put in place, since individuals with migration background first and foremost search for rental apartments, which can be an insufficient housing stock in small and medium-sized municipalities (Gans, 2018, pp. 375–396).

Our results show that segregation analysis should not be restricted to the residential location of different population groups, but can also be applied to residences themselves. In Eastern or rural municipalities with high incomes, there is a greater spatial separation between owner-occupied residences and the rest than in low-income municipalities. In contrast, there is a more uniform spread of buildings before 1978. In Eastern municipalities with high income and inequality, more buildings were built after 1995, which are also clustered in neighborhoods. This finding can be related to the urban growth that took place after the reunification in economically strong regions, as opposed to the urban shrinkage in many Eastern regions (Heider, 2019). Recent studies in Germany show that strict regional planning regulations reduce construction activities, but at the same time do not have the negative effect on building land prices and rents that is observed in UK or US (Eichhorn & Pehlke, 2022). Moreover, further research is needed to investigate how municipalities interact and adapt to regional regulations (Eichhorn & Pehlke, 2022). Making transparent and explicit the association between planning, growth and income inequality can support the agenda of local policy makers and planners in advocating for further leeway in the implementation of economically viable growth strategies.

We consistently observe built-up density as one of the main spatial indicators of income and inequality. Density is characterized in numerous ways: built-up residential area, the extent of residential annexes, and building height all serve as indicators of building structure and population density (Schug, Frantz, van der Linden, & Hostert, 2021). Population density in residential areas is increasing in many European, and especially German cities (Cortinovis et al., 2022). Densification can have environmentally positive effects, resulting from reduced urban sprawl (Cortinovis et al., 2022; Jehling, Schorcht, & Hartmann, 2020). Whether densification is fairly distributed along all socioeconomic categories of the population is, however, an important question. Our results show that higher variation in neighborhood built-up density is associated with higher income inequality. This finding is consistent with other studies showing that densification affects disproportionately already densely developed areas and populations with lower incomes (Bibby, Henneberry, & Halleux, 2021; Jehling et al., 2020), and that per capita disposable income and per capita built-up area are negatively correlated in many European cities (Masini et al., 2019). Spatial planning and policies on land use are essential in regulating urban form in Germany, and changes in income and transportation, and other market forces, are slower to propagate and impact urban planning, as compared with the US, for example (Schmidt et al., 2021). Timely tracking of local developments in income levels and income inequality can improve the responsiveness of spatial planning to changes in the structure of the local workforce, and more generally, of local demographics. Considerations of fair and appropriate use of the public and private spaces could also become more readily implementable.

The environment where people live and work is essential for a multitude of life outcomes, and the area of residence in particular “determines one’s present and future income“ in terms of employment, education and other opportunities (Bibby et al., 2021; Martín-Legendre et al., 2021). Bigger residential units and greener surroundings were found to correlate positively with higher incomes and higher inequality. Green spaces in urban environments are associated with greater health and psychological well-being (Brindley, Jorgensen, & Maheswaran, 2018; Engemann et al., 2019) and also generate a substantial monetary value, as for example reflected in house prices (Mears, Brindley, Jorgensen, & Maheswaran, 2020). As the quantitatively most important asset class for most households is wealth in real estate (Albers, Bartels, & Schularick, 2022; Wind, Lersch, & Dewilde, 2017), we postulate from our observations that inequality in income is often associated with inequality in wealth. Recent surveys in Germany show that people significantly underestimate the extent of wealth inequality (Bellani, Bledow, Busemeyer, & Schwerdt, 2021). The methodology proposed in this study holds considerable potential for the estimation of residential real estate wealth, which is an important source of inequality, and constituted, in 2018, more than half of total gross wealth in Germany (Albers et al., 2022, pp. 1895–2018).

In Germany, tax, social and labor policies are determined at the federal level, with individual states independently managing the school systems, and also implementing media and cultural policies and governing the police force (Gross & Krauss, 2021). There is however “considerable leeway“ in the application of federal policies at state level, which explains the differences in policy implementation between states (Gross & Krauss, 2021). At the sub-state level, local governments are responsible for executing state and federal laws and have autonomous decision making—among other duties—over the provision of childcare and public transit services, and the collection of local taxes and fees, from businesses and other local agents (Wegschaider et al., 2023). We readily note that the spatial and sociodemographic data available for this study explains to a limited extent differences at state or local levels in income determinants. However, our model-based regional analysis indicates already the pertinence of exploring multiple geographical scales for identifying sub-national variation trends, and supports future lines of research.

Spatial data sources open up innovative opportunities for research, while also bringing about specific limitations. We find that standalone models based on spatial features explain up to 65% of variability in median income and up to 49% of variability in inequality. Land cover/land use data is especially useful for describing small municipalities, while individual building data presents an opportunity for complex analysis for all types of settlements. This potential can be further enhanced by overcoming current limitations. For land use data, higher spatial resolutions would permit a refined mapping of the blue and green environment in cities, or multi-temporal monitoring of built-up changes. For building data, accurate labelling of functions would provide a better view of public amenities and services. Our analysis found only small effects of the mix of building functions or the accessibility to specific services on income levels and inequality. Further analysis of the accessibility to services would be beneficial, especially since most of the missing function labels occur in the data for eastern states with lower income levels.

Another limitation of our study is that not all of the socioeconomic variables used are measured at the exact same point in time. Whereas income is measured in 2016, the census dates back to 2011. The demographic structure of many German municipalities has inevitably changed during this time. The migration influx that occurred in Germany in 2015—largely due to several international refugee movements—is a prime example and cause. However, the impact on the labor market of migrant population largely depends on geographical origins, skills and education levels (Maffei-Faccioli & Vella, 2021; Vanella & Deschermeier, 2020), and in cases like 2015, the absorption in the workforce of migrant population from refugee groups takes place slowly (Brücker, Hauptmann, & Sirries, 2017) and is unlikely to meaningfully impact the income tax declarations in the years immediately following relocation. Concerning the overall population, there are clear trends of spatial disparities in demographic changes, with higher de-growth in Eastern Germany, but also in rural areas in Western Germany (Gans, 2018, pp. 375–396). Also, the most important predicted demographic changes—increasing number of single-family households, of senior citizens, and of citizens with migration backgrounds (Gans, 2018, pp. 375–396)—will very likely have an impact on labor market and earnings. We estimate, however, that no major changes—other than the migration influx—took place in the time horizon 2011–2016, and the interval is short enough to allow for the use of dependent and independent variables at different time points. These time lags may induce measurements errors which cannot be currently exactly estimated, but we speculate that the general trends discovered will still hold true.

The third set of important limitations for our analysis stems from the income tax data. Inequality indices computed from tax income tend to over-represent top earners (Bartels & Metzing, 2019). As a result, our analysis does not fully capture the bottom half of the income distribution. In addition, overall gross income inequality, as measured by the Gini coefficient, is driven mostly by inequality in the top part of the income distribution (Drechsel-Grau et al., 2022). We presume that this is one reason why some of the spatial attributes highlighted in our analysis are also indicators of real estate wealth. Another potential limitation stems from the use of pre-tax income, as opposed to real income, or disposable income. We presume that the relationship between many of the predictors and disposable income would be even more pronounced. Furthermore, due to significant differences in housing costs between the top and the bottom of the income distribution (Bartels & Schröder, 2020; Lozano Alcántara & Vogel, 2021, pp. 1–19), an analysis of disposable income after deducing housing costs could shed more light on regional income inequalities. Finally, recent studies showed that changes in income inequality are reflected in changes in residential segregation, with a time lag of approximately ten years (Tammaru et al., 2021). The future German census data—estimated to be available in 2024 —will allow us to extend the discussion to changes in inequality and changes in population characteristics.

Last but not least, comparable spatial and sociodemographic variables are increasingly available. At the same time, income data reporting is still lacking for a large part of European (and world) sub-national areas. Applying the proposed method to other countries could constitute a fruitful avenue of research. The different institutional contexts, local drivers of inequality and policy considerations should—naturally—be carefully considered when transferring the method. Nevertheless, our study identifies general phenomena associated with income levels and inequality which are relevant for numerous regions. Densification and land use constraints are a pan-European issue (Cortinovis et al., 2022); the integration of people with migration backgrounds to ensure positive social and economical effects locally is an essential policy action area in EU (Benassi et al., 2020); residential segregation and income inequality are interconnected through mechanisms related to the provision of public goods, services and opportunities based on area of residence, an issue both European (Tammaru et al., 2021), and global (Nicoletti et al., 2022).

6. Conclusion

In this paper, we show that it is possible to use machine learning techniques to predict municipality income inequality based on combined social and spatial data. This not only allows us to learn about the distribution of income at a geographically granular level where official income statistics are not easily available, but also delivers insights about the interconnections between the built and natural environment and socioeconomic status. We show that there exists a uniform set of attributes correlated with income, for a heterogeneous set of settlements. Many characteristics that are predictive of inequality in large urban areas, such as patterns of segregation, population diversity, green space availability and building density, are also informative for medium-sized municipalities, which are studied significantly less often in the literature. We therefore believe that our findings support the increasing need of refined and focused research at regional and local levels, and offer an additional stimulus to the growing debate towards increasing statistical data openness.

Financial disclosure/acknowledgements

This work was funded by Volkswagen Stiftung and the Compagnia di San Paolo Foundation, through the international cooperation project INEQUALITREES, grant number A130902.

CRediT author statement

Oana M. Garbasevschi: Conceptualization, Methodology, Formal analysis, Writing - Original Draft.

Hannes Taubenböck: Resources, Writing - Review Editing, Supervision, Funding acquisition.

Paul Schüle: Conceptualization, Data Curation, Writing - Original Draft, Project administration.

Julia Baarck: Data Curation, Writing - Review Editing.

Paul Hufe: Conceptualization, Writing - Review Editing, Funding acquisition.

Michael Wurm: Writing - Review Editing, Supervision.

Andreas Peichl: Resources, Writing - Review Editing, Supervision, Funding acquisition.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this publication.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apgeog.2023.103058.

References

Academy for Spatial Research and Planning. (2019). *Small town research in Germany – status quo and recommendations*. URL: http://nbn-resolving.de/urn:nbn:de:0156-01149.

Albers, T., Bartels, C., & Schularick, M. (2022). *Wealth and its distribution in Germany*.

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV). (2020). Data format description of official building polygons of Germany (HU-DE). URL: https://www.adv-online.de/AdV-Produkte/Standards-und-Produktblaetter/ZSHH/.

Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for random forests 6, 100094. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666827021000475.

Bartels, C. (2019). Top incomes in Germany, 1871-2014. *The Journal of Economic History, 79*. https://doi.org/10.1017/S0022050719000378. URL: https://www.cambridge.org/core/journals/journal-of-economic-history/article/abs/top-incomes-in-germany-18712014/C74EAC8F800F7A17E4691419F143757D.

Bartels, C., & Metzing, M. (2019). *An integrated approach for a top-corrected income distribution* (Vol. 17, pp. 125–143). https://doi.org/10.1007/s10888-018-9394-x. URL: http://link.springer.com/10.1007/s10888-018-9394-x.

Bartels, C., & Schröder, C. (2020). *The role of rental income, real estate and rents for inequality in Germany*. URL: https://newforum.org/wp-content/uploads/2022/01/FNE-WP07-2020.pdf.

Battisti, M., Peri, G., & Romiti, A. (2022). *Dynamic effects of co-ethnic networks on immigrants' economic success* (Vol. 132, pp. 58–88). URL: http://www.nber.org/papers/w22389.

Belgiu, M., & Drăguţ, L. (2016). *Random forest in remote sensing: A review of applications and future directions* (Vol. 114, pp. 24–31). https://doi.org/10.1016/j.isprsjprs.2016.01.011. URL: https://linkinghub.elsevier.com/retrieve/pii/S0924271616000265.

Bellani, L., Bledow, N., Busemeyer, M. R., & Schwerdt, G. (2021). *Perception of inequality and social mobility in Germany: Evidence from the inequality barometer*. URL: http://hdl.handle.net/10419/234612.

Benassi, F., Bonifazi, C., Heins, F., Lipizzi, F., & Strozza, S. (2020). *Comparing residential segregation of migrant populations in selected European urban and metropolitan areas* (Vol. 8, pp. 269–290). https://doi.org/10.1007/s40980-020-00064-5. URL: https://link.springer.com/10.1007/s40980-020-00064-5.

Bibby, P., Henneberry, J., & Halleux, J. M. (2021). *Incremental residential densification and urban spatial justice: The case of England between 2001 and 2011* (Vol. 58, pp. 2117–2138). https://doi.org/10.1177/0042098020936967. URL: http://journals.sagepub.com/doi/10.1177/0042098020936967.

Blanchet, T., Saez, E., & Zucman, G. (2022). Real-time inequality. URL: http://www.nber.org/papers/w30229.pdf.

Breiman, L. (2001). *Random forests* (Vol. 45, pp. 5–32). https://doi.org/10.1023/A:1010933404324. URL: http://link.springer.com/10.1023/A:1010933404324.

Brindley, P., Jorgensen, A., & Maheswaran, R. (2018). *Domestic gardens and self-reported health: A national population study* (Vol. 17, p. 31). https://doi.org/10.1186/s12942-018-0148-6. URL: https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-018-0148-6.

Brücker, H., Hauptmann, A., & Sirries, S. (2017). *Arbeitsmarktintegration von Geflüchteten in Deutschland: Der stand zum jahresbeginn 2017* [Labor market integration of refugees in Germany: The state at the beginning of the year 2017.] 4. URL: https://doku.iab.de/aktuell/2017/aktueller_bericht_1704.pdf.

Brunori, P., Hufe, P., & Mahler, D. (2021). *The roots of inequality: Estimating inequality of opportunity from regression trees and forests*.

Buch, T., Meister, M., & Niebuhr, A. (2021). *Ethnic diversity and segregation in German cities* (Vol. 115), Article 103221 https://doi.org/10.1016/j.cities.2021.103221. URL: https://linkinghub.elsevier.com/retrieve/pii/S0264275121001190.

Bundesamt für Kartographie und Geodäsie (BKG). (2019). Geographische Gitter für Deutschland in Lambert-Projektion (GeoGitter Inspire). URL: https://gdz.bkg.bund.de/index.php/default/open-data/geographische-gitter-fur-deutschland-in-lambert-projektion-geogitter-inspire.html.

Bundesamt für Kartographie und Geodasie (BKG). (2021). 3D-Gebäudemodelle LoD1 Deutschland (LoD1-DE). URL: https://gdz.bkg.bund.de/index.php/default/3d-gebaudemodelle-lod1-deutschland-lod1-de.html.

Bundesinstitut für Bau- Stadt- und Raumforschun (BBSR). (2017). Laufende stadtbeobachtung - raumabgrenzungen. URL: https://www.bbsr.bund.de/BBSR/DE/forschung/raumbeobachtung/Raumabgrenzungen/deutschland/gemeinden/StadtGemeindetyp/StadtGemeindetyp.html.

Bundesinstitut für Bau- Stadt- und Raumforschun (BBSR). (2018). *Raumordnungsbericht* (Vol. 2017).

Büttner, G. (2014). CORINE land cover and land cover change products. In *Land use and land cover mapping in Europe* (pp. 55–74). Springer.

Casali, Y., Aydin, N. Y., & Comes, T. (2022). *Machine learning for spatial analyses in urban areas: A scoping review* (Vol. 85), Article 104050. https://doi.org/10.1016/j.scs.2022.104050. URL: https://linkinghub.elsevier.com/retrieve/pii/S2210670722003687.

Castells-Quintana, D., Royuela, V., & Veneri, P. (2020). *Inequality and city size: An analysis for OECD functional urban areas* (Vol. 99, pp. 1045–1064). https://doi.org/10.1111/pirs.12520. URL: https://onlinelibrary.wiley.com/doi/10.1111/pirs.12520.

Chakraborty, T., & Schüller, S. (2022). *Ethnic enclaves and immigrant economic integration* https://doi.org/10.15185/izawol.287.v2. URL: https://iza.lokomotiv.cloud/articles/ethnic-enclaves-and-immigrant-economic-integration/long.

Chancel, L., & Piketty, T. (2021). *Global income inequality, 1820–2020: The persistence and mutation of extreme inequality* (Vol. 19, pp. 3025–3062). https://doi.org/10.1093/jeea/jvab047. URL: https://academic.oup.com/jeea/article/19/6/3025/6408467.

Chen, Q., Ye, T., Zhao, N., Ding, M., Ouyang, Z., Jia, P., et al. (2021). *Mapping China's regional economic activity by integrating points-of-interest and remote sensing data with random forest* (Vol. 48, pp. 1876–1894). https://doi.org/10.1177/2399808320951580. URL: http://journals.sagepub.com/doi/10.1177/2399808320951580.

Cortes, R. X., Rey, S., Knaap, E., & Wolf, L. J. (2020). *An open-source framework for non-spatial and spatial segregation measures: The PySAL segregation module* (Vol. 3, pp. 135–166). https://doi.org/10.1007/s42001-019-00059-3. URL: http://link.springer.com/10.1007/s42001-019-00059-3.

Cortinovis, C., Geneletti, D., & Haase, D. (2022). *Higher immigration and lower land take rates are driving a new densification wave in European cities* (Vol. 2, p. 19). https://doi.org/10.1038/s42949-022-00062-0. URL: https://www.nature.com/articles/s42949-022-00062-0.

Credit, K. (2022). *Spatial models or random forest? Evaluating the use of spatially explicit machine learning methods to predict employment density around new transit stations in los angeles* (Vol. 54, pp. 58–83). https://doi.org/10.1111/gean.12273. URL: https://onlinelibrary.wiley.com/doi/10.1111/gean.12273.

De Nicolò, S., Ferrante, M. R., & Pacei, S. (2022). *Small area estimation of inequality measures using mixtures of betas* (Vol. 28) https://doi.org/10.6092/unibo/amsacta/7073. URL http://arxiv.org/abs/2209.01985. arXiv:2209.01985 [stat].

Degenhardt, F., Seifert, S., & Szymczak, S. (2019). *Evaluation of variable selection methods for random forests and omics data sets* (Vol. 20, pp. 492–503). https://doi.org/10.1093/bib/bbx124. URL: https://academic.oup.com/bib/article/20/2/492/4554516.

Destatis. (2011). *Ergebnisse des Zensus 2011 zum Download - erweitert*. URL: https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrundyearn.html.

Diemer, A., Iammarino, S., Rodríguez-Pose, A., & Storper, M. (2022). *The regional development trap in Europe* (Vol. 98, pp. 487–509). https://doi.org/10.1080/00130095.2022.2080655. URL: https://www.tandfonline.com/doi/full/10.1080/00130095.2022.2080655.

Dittmann, J., & Goebel, J. (2010). *Your house, your car, your education: The socioeconomic situation of the neighborhood and its impact on life satisfaction in Germany* (Vol. 96, pp. 497–513). https://doi.org/10.1007/s11205-009-9489-7. URL: http://link.springer.com/10.1007/s11205-009-9489-7.

Donaldson, D., & Storeygard, A. (2016). *The view from above: Applications of satellite data in economics* (Vol. 30, pp. 171–198). https://doi.org/10.1257/jep.30.4.171. URL: https://pubs.aeaweb.org/doi/10.1257/jep.30.4.171.

Dorn, F., Fuest, C., Immel, L., & Neumeier, F. (2020). *Economic deprivation and radical voting: Evidence from Germany*. ifo Working Paper 336.

Drechsel-Grau, M., Peichl, A., Schmid, K. D., Schmieder, J. F., Walz, H., & Wolter, S. (2022). *Inequality and income dynamics in Germany* (Vol. 13, pp. 1593–1635). https://doi.org/10.3982/QE1912. URL: https://www.econometricsociety.org/doi/10.3982/QE1912.

Earth Observation Group (EOG). (2021). *Annual VIIRS nighttime lights (VNL)* (Vol. 2). URL: https://eogdata.mines.edu/products/vnl/#annual_v2.

Eichhorn, S., & Pehlke, D. (2022). *Unintended effects of regional planning in Germany* (Vol. 53, pp. 933–950). https://doi.org/10.1111/grow.12615. URL: https://onlinelibrary.wiley.com/doi/10.1111/grow.12615.

Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F. C., & Taneja, J. (2021). *Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019* (Vol. 13, p. 922). https://doi.org/10.3390/rs13050922. URL: https://www.mdpi.com/2072-4292/13/5/922.

Engemann, K., Pedersen, C. B., Arge, L., Tsirogiannis, C., Mortensen, P. B., & Svenning, J. C. (2019). *Residential green space in childhood is associated with lower risk of psychiatric disorders from adolescence into adulthood* (Vol. 116, pp. 5188–5193). https://doi.org/10.1073/pnas.1807504116. URL: https://pnas.org/doi/full/10.1073/pnas.1807504116.

Feldmeyer, D., Meisch, C., Sauter, H., & Birkmann, J. (2020). *Using OpenStreetMap data and machine learning to generate socio-economic indicators* (Vol. 9, p. 498). https://doi.org/10.3390/ijgi9090498. URL https://www.mdpi.com/2220-9964/9/9/498.

Frieden, I., Peichl, A., & Schüle, P. (2023). Regional income inequality in Germany. *CESifo Forum, 24*, 3–8.

Gans, P. (2018). *Demografischer wandel*. URL https://nbn-resolving.org/urn:nbn:de:0156-5599346.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., et al. (2019). *The German socio-economic panel (SOEP)* (Vol. 239, pp. 345–360). https://doi.org/10.1515/jbnst-2018-0022. URL: https://www.degruyter.com/document/doi/10.1515/jbnst-2018-0022/html.

Goebel, J., & Hoppe, L. (2015). *Ausmaß und Trends sozialräumlicher Segregation in Deutschland: Abschlussbericht. Technical Report*. DIW Berlin / SOEP. Berlin: Bundesministerium für Arbeit und Soziales.

Gross, M., & Krauss, S. (2021). *Topic coverage of coalition agreements in multi-level settings: The case of Germany* (Vol. 30, pp. 227–248). https://doi.org/10.1080/09644008.2019.1658077. URL: https://www.tandfonline.com/doi/full/10.1080/09644008.2019.1658077.

Heider, B. (2019). *What drives urban population growth and shrinkage in postsocialist East Germany?* (Vol. 50, pp. 1460–1486). https://doi.org/10.1111/grow.12337. URL: https://onlinelibrary.wiley.com/doi/10.1111/grow.12337.

Helbig, M., & Jähnen, S. (2018). *Wie brüchig ist die soziale architektur unserer städte? Trends und analysen der Segregation in 74 deutschen städten*. WZB Discussion Paper 2018-001.

Inglis, A., Parnell, A., & Hurley, C. B. (2022). *Visualizing variable importance and variable interaction effects in machine learning models* (Vol. 31, pp. 766–778). https://doi.org/10.1080/10618600.2021.2007935. URL: https://www.tandfonline.com/doi/full/10.1080/10618600.2021.2007935.

Ivan, K., Holobâcă, I. H., Benedek, J., & Török, I. (2019). *Potential of night-time lights to measure regional inequality* (Vol. 12, p. 33). https://doi.org/10.3390/rs12010033. URL: https://www.mdpi.com/2072-4292/12/1/33.

Jehling, M., Schorcht, M., & Hartmann, T. (2020). *Densification in suburban Germany: Approaching policy and space through concepts of justice* (Vol. 91, pp. 217–237). https://doi.org/10.3828/tpr.2020.13. URL: https://online.liverpooluniversitypress.co.uk/doi/10.3828/tpr.2020.13.

Khachiyan, A., Thomas, A., Zhou, H., Hanson, G., Cloninger, A., Rosing, T., et al. (2022). *Using neural networks to predict microspatial economic growth* (Vol. 4, pp. 491–506). https://doi.org/10.1257/aeri.20210422. URL: https://pubs.aeaweb.org/doi/10.1257/aeri.20210422.

Küpper, P. (2016). *Abgrenzung und Typisierung ländlicher Räume*.

Lee, N., Sissons, P., & Jones, K. (2016). *The geography of wage inequality in British cities* (Vol. 50, pp. 1714–1727). https://doi.org/10.1080/00343404.2015.1053859. URL: https://www.tandfonline.com/doi/full/10.1080/00343404.2015.1053859.

Li, Z. (2022). *Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost* (Vol. 96), Article 101845 https://doi.org/10.1016/j.compenvurbsys.2022.101845. URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971522000898.

Loga, T., Diefenbach, N., Stein, B., & Born, R. (2012). Tabula: Further development of the German residential building typology. URL: https://www.building-typology.eu/downloads/public/docs/scientific/DE_TABULA_ScientificReport_IWU.pdf.

Lozano Alcántara, A., & Vogel, C. (2021). *Rising housing costs and income poverty among the elderly in Germany* https://doi.org/10.1080/02673037.2021.1935759. URL https://www.tandfonline.com/doi/full/10.1080/02673037.2021.1935759.

Luttmer, E. F. P. (2005). Neighbors as negatives: Relative earnings and well-being, 10.1093/qje/120.3.963 *Quarterly Journal of Economics, 120*, 963–1002. https://doi.org/10.1093/qje/120.3.963. arXiv:https://academic.oup.com/qje/article-pdf/120/3/963/5211069/120-3-963.pdf.

Maffei-Faccioli, N., & Vella, E. (2021). *Does immigration grow the pie? Asymmetric evidence from Germany* (Vol. 138), Article 103846. https://doi.org/10.1016/j.euroecorev.2021.103846. URL: https://linkinghub.elsevier.com/retrieve/pii/S0014292121001793.

Marandola, G., & Xu, Y. (2021). *Mis-)perception of inequality: Measures, determinants, and consequences* https://doi.org/10.2139/ssrn.3898673. URL: https://www.ssrn.com/abstract=3898673.

Martín-Legendre, J. I., Castellanos-García, P., & Sánchez-Santos, J. M. (2021). *Neighborhood inequality and spatial segregation: An analysis with tax data for 40 Spanish cities* (Vol. 118), Article 103354 https://doi.org/10.1016/j.cities.2021.103354. URL: https://linkinghub.elsevier.com/retrieve/pii/S0264275121002547.

Masini, E., Tomao, A., Barbati, A., Corona, P., Serra, P., & Salvati, L. (2019). *Urban growth, land-use efficiency and local socioeconomic context: A comparative analysis of 417 metropolitan regions in Europe* (Vol. 63, pp. 322–337). https://doi.org/10.1007/s00267-018-1119-1. URL: http://link.springer.com/10.1007/s00267-018-1119-1.

Massey, D. S., & Denton, N. A. (1988). *The dimensions of residential segregation* (Vol. 67, pp. 281–315). URL: http://www.jstor.org/stable/2579183.

Mears, M., Brindley, P., Jorgensen, A., & Maheswaran, R. (2020). *Population-level linkages between urban greenspace and health inequality: The case for using multiple indicators of neighbourhood greenspace* (Vol. 62), Article 102284 https://doi.org/10.1016/j.healthplace.2020.102284. URL: https://linkinghub.elsevier.com/retrieve/pii/S1353829219307506.

Meijer, J. R., Huijbregts, M. A. J., Schotten, K. C. G. J., & Schipper, A. M. (2018). *Global patterns of current and future road infrastructure* (Vol. 13), Article 064006 https://doi.org/10.1088/1748-9326/aabd42. URL: https://iopscience.iop.org/article/10.1088/1748-9326/aabd42.

Molina, I., Corral, P., & Nguyen, M. (2022). *Estimation of poverty and inequality in small areas: Review and discussion* https://doi.org/10.1007/s11749-022-00822-1. URL: https://link.springer.com/10.1007/s11749-022-00822-1.

Morgan, B. S. (1983). *A distance-decay based interaction index to measure residential segregation*.

Nicoletti, L., Sirenko, M., & Verma, T. (2022). *Disadvantaged communities have lower access to urban infrastructure* https://doi.org/10.1177/23998083221131044. URL: http://journals.sagepub.com/doi/10.1177/23998083221131044.

Nijman, J., & Wei, Y. D. (2020). *Urban inequalities in the 21st century economy* (Vol. 117), Article 102188 https://doi.org/10.1016/j.apgeog.2020.102188. URL: https://linkinghub.elsevier.com/retrieve/pii/S0143622820301910.

Patias, N., Rowe, F., & Arribas-Bel, D. (2023). *Local urban attributes defining ethnically segregated areas across English cities: A multilevel approach* (Vol. 132), Article 103967 https://doi.org/10.1016/j.cities.2022.103967. URL: https://linkinghub.elsevier.com/retrieve/pii/S0264275122004061.

Perez-Truglia, R. (2020). The effects of income transparency on well-being: Evidence from a natural experiment. *The American Economic Review, 110*, 1019–1054. https://doi.org/10.1257/aer.20160256. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20160256.

Reardon, S. F., & Firebaugh, G. (2002). *Measures of multigroup segregation* (Vol. 32, pp. 33–67). https://doi.org/10.1111/1467-9531.00110. URL: http://journals.sagepub.com/doi/10.1111/1467-9531.00110.

Roberto, E. (2015). *The divergence index: A decomposable measure of segregation and inequality*. arXiv:1508.01167.

Rodríguez-Pose, A., Iammarino, S., & Storper, M. (2018). *Regional inequality in Europe: Evidence, theory and policy implications*. URL: https://cepr.org/publications/dp12841.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., et al. (2021). *A generalizable and accessible approach to machine learning with global satellite imagery* (Vol. 12, p. 4392). https://doi.org/10.1038/s41467-021-24638-z. URL: http://www.nature.com/articles/s41467-021-24638-z.

Rosés, J. R., & Wolf, N. (2021). *Regional growth and inequality in the long-run: Europe, 1900–2015* (Vol. 37, pp. 17–48). https://doi.org/10.1093/oxrep/graa062. URL: https://academic.oup.com/oxrep/article/37/1/17/6211736.

Rüttenauer, T., & Best, H. (2021). *Environmental inequality and residential sorting in Germany: A spatial time-series analysis of the demographic consequences of industrial sites* (Vol. 58, pp. 2243–2263). https://doi.org/10.1215/00703370-9563077. URL: https://read.dukeupress.edu/demography/article/58/6/2243/257657/Environmental-Inequality-and-Residential-Sorting.

Salas-Rojo, P., & Rodríguez, J. G. (2022). *Inheritances and wealth inequality: A machine learning approach 20* https://doi.org/10.1007/s10888-022-09528-8. URL: https://link.springer.com/10.1007/s10888-022-09528-8.

Sapena, M., Ruiz, L., & Taubenböck, H. (2020). *Analyzing links between spatio-temporal metrics of built-up areas and socio-economic indicators on a semi-global scale* (Vol. 9, p. 436). https://doi.org/10.3390/ijgi9070436. URL: https://www.mdpi.com/2220-9964/9/7/436.

Sapena, M., Wurm, M., Taubenböck, H., Tuia, D., & Ruiz, L. A. (2021). *Estimating quality of life dimensions from urban spatial pattern metrics* (Vol. 85), Article 101549 https://doi.org/10.1016/j.compenvurbsys.2020.101549. URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971520302829.

Schmidt, S., Li, W., Carruthers, J., & Siedentop, S. (2021). *Planning institutions and urban spatial patterns: Evidence from a cross-national analysis* https://doi.org/10.1177/0739456X211044203. URL http://journals.sagepub.com/doi/10.1177/0739456X211044203.

Schug, F., Frantz, D., van der Linden, S., & Hostert, P. (2021). *Gridded population mapping for Germany based on building density, height and type from earth observation data using census disaggregation and bottom-up estimates* (Vol. 16), Article e0249044. https://doi.org/10.1371/journal.pone.0249044. URL: https://dx.plos.org/10.1371/journal.pone.0249044.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). *A comparison of random forest variable selection methods for classification prediction modeling* (Vol. 134, pp. 93–101). https://doi.org/10.1016/j.eswa.2019.05.028. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417419303574.

Tammaru, T., Sinitsyna, A., Akhavizadegan, A., van Ham, M., Marcińczak, S., & Musterd, S. (2021). Income inequality and residential segregation in European cities. In G. Pryce, Y. P. Wang, Y. Chen, J. Shan, & H. Wei (Eds.), *Urban inequality and segregation in Europe and China* (pp. 39–54). Springer International Publishing. https://doi.org/10.1007/978-3-030-74544-8_3. URL: https://link.springer.com/10.1007/978-3-030-74544-8_3.

Taubenböck, H., Droin, A., Standfuß, I., Dosch, F., Sander, N., Milbert, A., et al. (2022). *To be, or not to be 'urban'? A multi-modal method for the differentiated measurement of the degree of urbanization* (Vol. 95), Article 101830 https://doi.org/10.1016/j.compenvurbsys.2022.101830. URL: https://linkinghub.elsevier.com/retrieve/pii/S0198971522000746.

Vanella, P., & Deschermeier, P. (2020). *A probabilistic cohort-component model for population forecasting – the case of Germany* (Vol. 13, pp. 513–545). https://doi.org/10.1007/s12062-019-09258-2. URL: http://link.springer.com/10.1007/s12062-019-09258-2.

Wegschaider, K., Gross, M., & Schmid, S. (2023). *Studying politics at the local level in Germany: A tale of missing data* (Vol. 16, pp. 753–768). https://doi.org/10.1007/s12286-022-00551-7. URL: https://link.springer.com/10.1007/s12286-022-00551-7.

Weigand, M., Staab, J., Wurm, M., & Taubenböck, H. (2020). *Spatial and semantic effects of LUCAS samples on fully automated land use/land cover classification in high-resolution Sentinel-2 data 88*, Article 102065 https://doi.org/10.1016/j.jag.2020.102065. URL: https://linkinghub.elsevier.com/retrieve/pii/S0303243419307317.

Wind, B., Lersch, P., & Dewilde, C. (2017). *The distribution of housing wealth in 16 European countries: Accounting for institutional differences* (Vol. 32, pp. 625–647). https://doi.org/10.1007/s10901-016-9540-3. URL: http://link.springer.com/10.1007/s10901-016-9540-3.

Windsteiger, L. (2022). The redistributive consequences of segregation and misperceptions. *European Economic Review, 144*, Article 104073.

Wójcik, P., & Andruszek, K. (2021). Predicting intra-urban well-being from space with nonlinear machine learning. rsp3.12478URL https://onlinelibrary.wiley.com/doi/10.1111/rsp3.12478.

Wurm, M., Weigand, M., Stark, T., Goebel, J., Wagner, G. G., & Taubenböck, H. (2019). Modelling the impact of the urban spatial structure on the choice of residential location using 'big earth data' and machine learning. In *Proceedings of the joint urban remote sensing event (JURSE)*.

Würz, N., Schmid, T., & Tzavidis, N. (2022). *Estimating regional income indicators under transformations and access to limited population auxiliary information* (Vol. 185, pp. 1679–1706). https://doi.org/10.1111/rssa.12913. URL: https://onlinelibrary.wiley.com/doi/10.1111/rssa.12913.

Xu, X., Metsälampi, S., Kichler, M., Kotakorpi, K., Matthews, P. H., & Miettinen, T. (2023). *Which income comparisons matter to people, and how? Evidence from a large field experiment*. FIT Working Paper 10.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., et al. (2020). *Using publicly available satellite imagery and deep learning to understand economic well-being in Africa* (Vol. 11, p. 2583). https://doi.org/10.1038/s41467-020-16185-w. URL: http://www.nature.com/articles/s41467-020-16185-w.