Masías H. et al. *EPJ Data Science* (2024) 13:21
https://doi.org/10.1140/epjds/s13688-024-00455-4

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

Evolving demographics: a dynamic clustering approach to analyze residential segregation in Berlin

Víctor H. Masías H.[1,4*], Julia Stier[1], Pilar Navarro R.[2], Mauricio A. Valle[3], Sigifredo Laengle[4], Augusto A. Vargas[5] and Fernando A. Crespo R.[6]

*Correspondence:
vmasias@fen.uchile.cl
[1] WZB Berlin Social Science Center, Reichpietschufer 50, 10785, Berlin, Germany
[4] Faculty of Economics and Business, University of Chile, Diagonal Paraguay 257, 8320000, Santiago, Chile
Full list of author information is available at the end of the article

Abstract

This paper examines the phenomenon of residential segregation in Berlin over time using a dynamic clustering analysis approach. Previous research has examined the phenomenon of residential segregation in Berlin at a high spatial and temporal aggregation and statically, i.e. not over time. We propose a methodology to investigate the existence of clusters of residential areas according to migration background, age group, gender, and socio-economic dimension over time. To this end, we have developed a sequential mixed methods approach that includes a multivariate kernel density estimation technique to estimate the density of subpopulations and a dynamic cluster analysis to discover spatial patterns of residential segregation over time (2009-2020). The dynamic analysis shows the emergence of clusters on the dimensions of migration background, age group, gender and socio-economic variables. We also identified a structural change in 2015, resulting in a new cluster in Berlin that reflects the changing distribution of subpopulations with a particular migratory background. Finally, we discuss the findings of this study with previous research and suggest possibilities for policy applications and future research using a dynamic clustering approach for analyzing changes in residential segregation at the city level.

Keywords: Berlin; Data Science; Dynamic Fuzzy C–Means; Residential Segregation; Data Visualization

1 Introduction

This manuscript examines the phenomenon of residential segregation in Berlin from a dynamic perspective, using data science to identify patterns in its human geography.

Previous research on residential segregation in Berlin has analyzed its different dimensions. For example, research has focused on residential segregation driven by ethnicity, [1–29]; residential segregation driven by age-group [1, 6, 7, 12, 20, 21, 23, 24, 29]; residential segregation driven by gender [4, 8, 15]; segregation driven by socio-economic factors [2, 3, 10, 12–15, 17, 18, 20, 23, 27, 29]; as well as residential segregation driven by digital segregation [24]. Demography, economics, sociology, geography, and ethnographic studies have explored all these dimensions, among other disciplines. They all support the notion that there is an uneven, clustered, or patchy spatial distribution of subpopulations in residential areas of Berlin.

We focus on the case of Berlin, a city in which historical events have changed both the city and its society. First, we are motivated by the Berlin case because a large body of research examines changes in residential segregation before and during the fall of the Berlin Wall. In fact, before the fall of the Wall, studies were based exclusively on data from West Berlin, as no unified public statistics were available. Then, after the fall of the Wall, data from both parts of the city became available for the first time, and research focused on understanding the structural changes that occurred as a result of the city's reunification. Finally, Berlin has served as a reference point for other comparative studies of residential segregation in other German cities in the context of immigration policy. Further details on the historical and research development of the phenomenon of residential segregation in Berlin can be found in [3, 29, 30].

Second, from a conceptual point of view, residential segregation occurs over time and space. However, to the best of our knowledge, there is no research in the case of Berlin that has carried out an analysis that includes these two components. For example, in Helbig's research on residential segregation [23], the author conducts a time series analysis but does not include the geographical dimension of residential areas. Another example is the work of Marcińczak and Bernt [28], in which they use hierarchical clustering on temporal data, which does not allow the identification of the emergence of new clusters or the disappearance of clusters over time. New methodological approaches are therefore needed to study the dynamics of residential segregation.

Third, we are motivated to explore the possible impact of the 2015 European migrant crisis on residential segregation in Berlin. As the city that received the highest number of refugees in Germany during the crisis, we are interested in examining whether the migration process led to the creation or elimination of clusters that shape the demographic composition of residential areas. The type of changes may include, for example, changes in the number of clusters over time (i.e. macrodynamics) and internal changes in subpopulations over time and space (i.e. microdynamics). Past studies looking at this period have not reported the existence of structural and internal changes within clusters, and one possibility for this situation may be the use of aggregate data or static clustering methods.

To reveal both structural and internal changes within clusters, we propose an exploratory analysis based on data science using a dynamic clustering algorithm. The objectives of this paper are:

- Estimate population density according to dimensions such as migration background, age group, gender, and socio-economics.
- Dynamically determine the number of clusters over time according to dimensions such as migration background, age group, gender, and those reflecting the socio-economic conditions.
- Identify structural as well as intra-cluster changes over time.
- Determine the variables that are over- or under-represented for each cluster in a given year.

As a result, our approach has led to discoveries about residential segregation. First, at a macro or structural level, a new cluster emerged in 2015, which we interpret as a result of the so-called European migrant crisis. At a micro or intra-cluster level, we report which subpopulations are under or over-represented in each cluster over time, revealing a rich dynamic of change in the city. By applying data science principles, it is possible to explore the phenomenon of residential segregation in an unsupervised and dynamic way. The contribution of this research is to present a dynamic analysis of the existing clusters in Berlin for the first time. In other words, the contribution of this paper is that it allows us to dynamically determine the number of clusters and the attributes that are more important over time in this analytical context. In this sense, an approach based on data science offers a huge field of application for the identification of changes in the city.

The remainder of the article is organized as follows. Section 2 provides an overview of previous research. Section 3 presents the methodological approach used to estimate population densities in residential areas and perform dynamic clustering based on calculating different subpopulations in Berlin. Section 4 presents the results, the clusters found, the criteria used to validate them, and the interpretation of the results. Finally, Sect. 5 presents a discussion and conclusions based on the research objectives.

2 Literature review

Data science is used to address complex problems related to sociological, economic and demographic factors. In particular, it is used to study residential segregation using unsupervised approaches. Multivariate and unsupervised methods are often preferred because there is no single view or way of quantifying residential segregation, and there is no baseline or ground truth for conducting supervised analyses.

A range of methodological approaches from the field of data science have been employed to study residential segregation. Spatial concentration patterns have been studied for a long time using a factorial approach [31, 32], which is now better known as factorial ecology [33]. Modern urban data science approaches also use this method. For instance, Benassi et al. [34] developed a composite index using multiple principal component analyses, which has been a revival of this approach. Recently, non-supervised machine learning methods have been employed to recognize patterns of residential segregation. For example, Olteanu-Raimond et al. [35] used traditional self-organizing maps, a type of neural network, to identify emerging patterns. Other researchers (see for example [36]) have used data science to improve the visualisation of changes in segregation and diversity in 61 major US cities between 1990 and 2020. Finally, Masías et al. [29] used unsupervised algorithms commonly used in image processing and remote sensing to generate visualizations and human-understandable information, based on concepts of cognitive psychology.

Approaching the study of residential segregation from a data science perspective, taking into account its spatial and temporal dimensions is a multidimensional problem. In the case of Berlin, for example, several dimensions of residential segregation have been studied. These include the ethnic dimension, which has been studied in Germany under the concept of migration background [1–29]; age or age-group segregation [1, 6, 7, 12, 20, 21, 23, 24, 29], which corresponds to the fact that different age groups are clustered in different parts of the city; gender segregation [4, 8, 15], which is a phenomenon associated with unbalanced gender ratios across space; social or socio-economic segregation [2, 3, 10, 12–15, 17, 18, 20, 23, 27, 29], where people are grouped with others with similar socio-economic characteristics, reflecting their economic opportunities; and the digital segregation dimension [24], which attempts to map access to social media and other digital technologies. The different emphases of some previous publications are summarised below (see Table 1).

Table 1 Previous research on the dimensions of residential segregation in Berlin

Work	Migration background	Age-group	Gender	Socio-economic	Digital
[1]	✓	✗	✗	✗	✗
[3]	✓	✗	✗	✗	✗
[4]	✗	✓	✗	✗	✗
[5]	✓	✗	✗	✗	✗
[6]	✓	✓	✗	✗	✗
[7]	✓	✓	✗	✗	✗
[8]	✗	✓	✓	✗	✗
[9]	✓	✗	✗	✗	✗
[2]	✓	✗	✗	✓	✗
[10]	✓	✗	✗	✓	✗
[11]	✓	✗	✗	✗	✗
[12]	✓	✓	✗	✓	✗
[13]	✓	✗	✗	✓	✗
[14]	✓	✗	✗	✓	✗
[15]	✗	✓	✓	✓	✗
[16]	✓	✗	✗	✗	✗
[17]	✓	✗	✗	✓	✗
[18]	✓	✗	✗	✓	✗
[19]	✓	✗	✗	✗	✗
[20]	✓	✓	✓	✓	✗
[21]	✓	✓	✗	✗	✗
[22]	✓	✗	✗	✗	✗
[23]	✓	✓	✓	✓	✗
[24]	✓	✓	✗	✓	✓
[25]	✓	✗	✗	✗	✗
[26]	✓	✗	✗	✗	✗
[27]	✓	✗	✗	✓	✗
[28]	✓	✗	✗	✗	✗
[29]	✓	✓	✓	✓	✗

As seen in Table 1, most previous studies focus on the ethnic aspect of residential segregation. The study of segregation by age group is the second most common. The third most studied aspect is social segregation. Finally, the least researched aspects are gender segregation and digital segregation. However, as can be seen, previous research has also been a study of more than one dimension at a time. Among the works cited, we would like to highlight the following ones:

- Kemper [6, 7], Arin [3], and Yamamoto [5, 37] contribute to a conceptual, empirical and historical understanding of the emergence of ethnic residential segregation in Berlin from a geographical and economic perspective.
- Nakagawa [4, 8] and Kröhnert [15] understand gender segregation as a consequence of ongoing migration processes within Germany from a socio-demographic perspective.
- Although Helbig's [23] contribution does not consider the demographic spatial dimension of residential areas (i.e. estimates of population density across residential areas), his work emphasizes the temporal dimension of residential segregation.
- The contribution of Marcińczak [28], who conducted a cluster analysis of residential segregation in Berlin, used hierarchical cluster analysis to examine several years of demographic data. However, no dynamic analysis of cluster formation in Berlin was carried out, due to the nature of the clustering method used, which is static. Furthermore, this work is guided by a predefined interpretation of the clusters.
- Kurtenbach's key study [24], which explored digital segregation in Berlin using data from a social media service designed to organise community life in neighbourhoods.

- Finally, innovative techniques for estimating population densities in residential areas in Berlin have been developed. For example, Groß [21] has developed methods for estimating anonymized spatial densities at a higher resolution. Building on this work, Masías et al. [29] have used non-negative matrix factorization to study different facets of residential segregation.

Previous studies on the city of Berlin have adopted a non-dynamic approach. The lack of dynamic clustering methods has led other researchers to use, for example, hierarchical cluster analysis, which does not allow the temporal aspect to be taken into account.

In this context, we aim to perform a data analysis which has the advantage of not being a black box. This allows for a more direct interpretation of the clusters and takes into account the existing dynamics, being more accessible to interpret, which is not the case with factor analysis methods or those using black-box machine learning methods.

3 Methods

3.1 Methodological approach

The proposed methodological approach is three-fold: first, estimating the spatial density of diverse subpopulations over Berlin, with the exclusion of non-residential areas, employing a Multivariate Kernel Density Estimation; second, the spatial densities estimated in the first step are analyzed via a Dynamic Fuzzy C-Means clustering method; finally, human-readable information about the composition of the clusters and their interpretation is generated. The flow chart in Fig. 1 summarizes the methodological approach followed throughout this work.

3.1.1 Data source

The register of residents (Einwohnerregister) from 2009 to 2020, available at the Statistical Office of Brandenburg (see www.statistik-berlin-brandenburg.de), was used as the input for this step. We only used the information regarding the migration background, age group, gender, and socio-economic demographics for each LOR spatial planning area (i.e., Die *lebensweltlich orientierten Räume*).

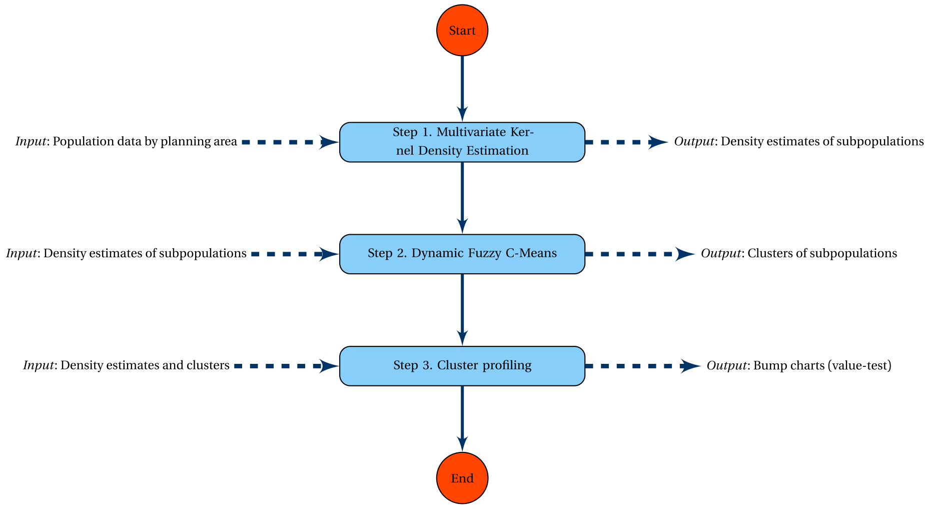

Start
Input: Population data by planning area
Step 1. Multivariate Kernel Density Estimation
Output: Density estimates of subpopulations
Input: Density estimates of subpopulations
Step 2. Dynamic Fuzzy C-Means
Output: Clusters of subpopulations
Input: Density estimates and clusters
Step 3. Cluster profiling
Output: Bump charts (value-test)
End

Figure 1 Methodological steps for the dynamic analysis of residential segregation phenomena

As an indirect measure of the ethnic dimension of residential segregation, the category of migration background has often been used in German sociology. It was first defined in 2005, when it was used in the microcensuses. The official definition used in 2005 is as follows: an individual with a migrant background is defined as "all migrants who entered the current territory of the Federal Republic of Germany after 1949, and all foreigners born in Germany and all those born in Germany as Germans with at least one parent who immigrated to Germany or who was born as a foreigner in Germany" [38, p.6]. In this context, the migrant background is instead referred to as a statistical category based on citizenship and an indirect record of the place of birth of the individual's parents.

Information on the demographic distribution by age group and gender in each LOR area for each year was also used. To estimate the density of males and females in the city of Berlin, we used data on the sex of individuals and information on the number of people in a given age group living in a given LOR. As sex ratios vary from country to country, and international or internal migration processes may have skewed age groups that differ from the destination population, and as this phenomenon has previously been reported as occurring in Berlin (see, [15]), we explore the possibility of residential segregation by gender.

Finally, people experiencing economic hardship in Germany are entitled to receive social benefits as defined in the Second and Third Book of the Social Code (SGB II and SGB III). In principle, any EU or non-EU citizen with a valid residence permit is entitled to SGB II and SGB III benefits after working in Germany for at least one year.

The SGB II (Sozialgesetzbuch Zweites Buch) and the SGB III (Sozialgesetzbuch Drittes Buch) are the two most fundamental laws of the German social security system. SGB II, or "Hartz IV", deals with social benefits for unemployed or low-income persons. SGB II also regulates the payment of unemployment benefits and social assistance. SGB III deals with employment promotion, vocational training, and education. It is aimed at helping people to find and keep a job and to improve their vocational skills. SGB III provides various measures to support job seekers, such as job counselling, placement services, and vocational training programs. Funding provisions and support for companies to create jobs and train their employees are also included.

In summary, while SGB II focuses on providing financial assistance to those in need, SGB III aims to promote employment and vocational training. For this paper, the number of people who obtained benefits under SGB II and SGB III in a given year and city location is considered a proxy measure of social or socio-economic segregation.

3.1.2 Multivariate kernel density estimation in the presence of measurement error

The spatial density of inhabitants is estimated following the work of [39], where a method is proposed to estimate the population density over areas with arbitrary shapes. That method is, in turn, based on a previous publication of the same author in which demographic estimates from rectangular spatial grids of different sizes are computed while introducing measurement error and data anonymization [21]. Other worth mentioning areas of application of the present method are the estimation of ethnic minority settlement areas [21], regional childcare demand estimates [40], regional election analyses [41], or estimates of the incidence of Coronavirus infections over time and space [42].

In the present study, Berlin is divided into spatial units (Planungsräume) whose centroids contain the spatial coordinates (measured in degrees). The technique of [39] is then applied using the LORs (Lebensweltlich-Orientierte Räume) areas on the aggregated number of inhabitants with distinct migratory origins, age, gender, and socio-economic conditions living in each of those spatial units. To obtain corrected density estimates, the non-residential areas were discounted in the analysis (see [43]).

The model used in this work to estimate the corrected spatial density from heaped data (i.e., the arbitrary aggregation of data in a spatial area) in polygons of an arbitrary shape is based on a non-parametric estimation method: the *Multivariate Kernel density estimation* technique. This approach estimates a finite sample's joint probability density function of two or more continuous random variables. In simpler words, it is used to estimate the distribution or spread of the data across more than one dimension when only a finite number of data points are available.

Let $X = \{X_1, X_2, \dots, X_n\}$ be a sample from a multivariate a random variable with probability distribution described by the unknown density function $f(x)$ to be estimated. Each random variable is two-dimensional in our case, i.e., $X_i = (X_{i1}, X_{i2})$, $i = 1, \dots, n$, being X_{i1} and X_{i2} the longitude and latitude coordinates, respectively, and X is the set containing all the available spatial coordinates. Then, the multivariate kernel density estimate at the two-dimensional point x is defined to be:

$$\hat{f}_H(x) = \frac{1}{n|H|^{\frac{1}{2}}} \sum_{i=1}^{n} K\left(H^{-\frac{1}{2}}(x - X_i)\right), \tag{1}$$

where:

- $|\cdot|$ denotes the determinant.
- $K(\cdot)$ is the kernel, a symmetric multivariate density function. This function assigns weights to the observed data points based on their distance from the point where we want to estimate the density. We use the standard *multivariate normal kernel*, i.e., $K(x) = (2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}x^T H^{-1} x}$.
- H is the bandwidth $d \times d$ [1] matrix, characterized by being symmetric and positive definite. It controls the window size in each dimension over which the kernel function operates. A small bandwidth will result in a density estimate that is very sensitive to the data (potentially too sensitive, resulting in over-fitting). In contrast, a large bandwidth may smooth out important features of the data (under-fitting). Therefore, the choice of H is critically important for the accuracy of the kernel density estimations. There exists a lot of discussion in the literature about the selection of the bandwidth matrix. Here, we use the approach of Wand and Jones, as it is done in [21].

In short, a function that returns high values for points close to the data point and low values for points far away is created at each data point, the *multivariate kernel*. The final density estimate at point x is the average contributions from all these kernel functions centered at each data point X_i. In this way, the density is high, where many data points are close together, and low, where the data points are spread out.

Since we have data spatially aggregated for each area of the city, rather than the exact coordinates, we use the approach of Groß et al. [21], that introduces measurement error to produce estimates of population density while anonymising the sensitive data. Formally, the actual values $X = \{X_1, X_2, \dots, X_n\}$ are unknown, and only the aggregated values over

[1] In our study $d = 2$.

each area can be utilized, which are denoted by $W = \{W_1, W_2, \ldots, W_n\}$. They can be seen as a measurement with an introduced error of the actual coordinates of individual i, where $i = 1, \ldots, n$. The objective is to estimate the density $f(x)$, from which X is drawn, only with the values W_i.

A naive kernel density estimator, which would use the aggregated values as the real coordinates in Equation (1), may lead to a spiky density far from the actual density of the true data. This effect becomes more noticeable as the sample size increases. Therefore, a model which contemplates the measurement error must be used. Under the assumption that the anonymization process is known, a measurement error model for W can be defined as $\pi(W|X) = \prod_{i=1}^{n} \pi(W_i|X_i)$, where $\pi(W|X)$ refers to the conditional distribution of W given X, and

$$\pi(W_i|X_i) = \begin{cases} 1 & \text{for } X_i \in \text{area}(W_i), \\ 0 & \text{otherwise}, \end{cases} \tag{2}$$

with area(W_i) being the set of coordinates that lie within the area where W_i belongs. Using the Bayes theorem formulation $\pi(X_i|W_i) \propto \pi(W_i|X_i)\pi(X_i)$ (i.e. the probability of X_i given W_i is proportional to the probability of W_i given X_i times the probability of X_i), pseudo-samples of X_i can be drawn from $\pi(X_i|W_i)$, which are used to estimate the density function $f(x)$. In particular, following an iterative procedure, X_i is drawn from the known conditional distribution $\pi(W_i|X_i)$ using $\pi(X_i)$ as a weight. Since $f(X_i)$ is unknown, and thus, $\pi(X_i)$ as well, the multivariate kernel density estimator $\hat{f}_H(x)$ defined in Equation (1) is used instead. At the beginning of the procedure, an estimate $\hat{f}_H^{(0)}(x)$ is calculated according to Equation (1) from the artificial geo-coordinates W_i. After drawing the pseudo-samples, the multivariate kernel density estimator is applied to these samples to estimate the density function $\hat{f}_H^{(1)}(x)$. In the following iterations, the density estimate $\hat{f}_H^{(N+1)}(x)$ is recalculated by utilizing the drawn pseudo-samples in the previous iteration N. In this way, the pseudo-samples provide a way to fill in the information lost due to data aggregation, and the density estimate is refined in each iteration. For more details on the steps of the algorithm, see [21].

3.1.3 *Dynamic fuzzy c-means*

This dynamic clustering algorithm, presented by Crespo and Weber [44] in 2005, relies on updating the structure of the current clusters based on relevant changes in the dynamic data. The period between the creation of a cluster structure and its update is called *cycle*, and its definition makes it possible to adapt the algorithm methodology to any probabilistic clustering algorithm, i.e., any clustering algorithm that determines degrees of membership. The degree of membership of an item to the clusters is used to identify changes in the structure of the clusters.

Changes in the structure of the clusters can be the creation of new clusters, elimination of clusters, or movement of the centers of the clusters. The following are the basic steps of the Dynamic Fuzzy C-Means:

- 1. Run the fuzzy c-means algorithm using the initial data set.
- 2. Receive new data and merge it with the current data.
- 3. Look for relevant changes in the structure of clusters.
- 4. If relevant changes exist, update the structure of clusters.

- 5. Repeat until no new data arrive.

In what follows, a detailed description of the mathematical aspects of the algorithm used here is provided. Let X_0 be the initial data set and $X_1, X_2, \ldots, X_t$ be the new datasets the algorithm receives in each cycle $t > 0$. In the beginning, the traditional fuzzy c-means algorithm is run on the first data set, X_0, with $c \geq 2$ clusters and fuzzifier $m > 0$, so that it produces c clusters with its respective centers $\mathbf{v}_j$, for $j = 1, 2, \ldots, c$, and the membership matrix $\mathbf{M}^0_{n\times c}$, being n the number of data points in X_0. The components of this membership matrix are the membership degrees, i.e., its component at position (i,j), $i = 1, \ldots, n$, $j = 1, 2, \ldots, c$, is the membership degree $\mu_{i,j}$ of the data point $x_i \in X_0$ to cluster j.

Let X_t be the new data chunk arriving into the dataset at cycle $t > 0$, which could produce changes in the current structure of the clusters because it contains data points that are not well classified by the current clusters. Let c_t be the number of clusters at cycle t, n_t the number of objects in the dataset X_t and $i = 1, \ldots, n_t$ the index of the new objects. To identify the data points producing changes, the following must be calculated:

- pair-wise distance $d(\mathbf{v}_j, \mathbf{v}_k)$ between each pair of the current centers $\mathbf{v}_j$ and $\mathbf{v}_k$, for all $j, k = 1, 2, \ldots, c_t$.
- distance $d(\mathbf{x}_i, \mathbf{v}_j)$ between the new data point $\mathbf{x}_i \in X_t$ and the current centers $\mathbf{v}_j$, for all $i = 1, \ldots, n_t$ and $j = 1, 2, \ldots, c_t$.
- the membership degree $\hat{\mu}_{i,j}$ of the new object $\mathbf{x}_i \in X_t$ to the cluster j, for all $i = 1, \ldots, n_t$ and $j = 1, 2, \ldots, c_t$.

Then, conditions shown in Equation (3) and Equation (4) must be evaluated on the new data to detect objects of X_t that are incorrectly assigned to the current clusters, i.e., those objects that would involve a change in the current structure.

$$\left| \hat{\mu}_{i,j} - \frac{1}{c_t} \right| \leq \alpha \quad \forall i = 1, \ldots, n_t, j = 1, 2, \ldots, c_t, \tag{3}$$

where $\alpha > 0$ is a threshold parameter fixed beforehand by the decision maker or dynamically determined by the algorithm.

$$d(\mathbf{x}_i, \mathbf{v}_j) \geq \frac{1}{2} \min_{\mathbf{v}_k \neq \mathbf{v}_j} \left\{ d(\mathbf{v}_j, \mathbf{v}_k) \right\} \quad \forall i = 1, \ldots, n_t, j = 1, 2, \ldots, c_t. \tag{4}$$

The two above conditions are used to define the indicator function (see Equation (5)), that is equal to one if, and only if, the data point $\mathbf{x}_i \in X_t$ is correctly classified by the current structure:

$$\mathbb{1}_{\mathrm{IC}}(\mathbf{x}_i) = \begin{cases} 1, & \text{if } \mathbf{x}_i \in X_t \text{ fulfills Equation (3) and Equation (4)} \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

If at least one new data object is not well classified, the criterion defined in Equation (6) is applied to decide whether new clusters should be created or if, conversely, moving the current centers is sufficient:

$$\frac{\sum_{\mathbf{x}_i \in X_t} \mathbb{1}_{\mathrm{IC}}(\mathbf{x}_i)}{|X_t|} \geq \beta, \tag{6}$$

where $\beta \in [0,1]$ is another threshold parameter that can be fixed previously or adjusted dynamically, and $|\cdot|$ represents the number of elements of a set. Whenever the condition defined in Equation (6) fulfils, new clusters must be created, and, in other cases, it is enough to update the centers of the current clusters.

If many new objects cannot be correctly assigned to the current clusters, i.e., new clusters are to be created, the optimum number of new clusters has to be determined. To do so, we select the number of clusters that maximize the structure strength [45], as it is done in the original paper [44]. Nevertheless, any other procedure could be used to find the new number of clusters. Once the optimum number is determined, the fuzzy *c*-means algorithm is run from scratch using that number on the total dataset.

In other cases, when it is sufficient to move the current centers of the clusters, the current centers are combined with those representing the new data. The cluster centers representing only the new data are calculated using Equation (7) and combined with the previous centers as defined in Equation (8):

$$\mathbf{v}_j^* = \frac{\sum_{\mathbf{x}_i \in X_t} [1 - \mathbb{1}_{\mathrm{IC}}(\mathbf{x}_i)](\hat{\mu}_{i,j})^m \cdot \mathbf{x}_i}{\sum_{\mathbf{x}_i \in X_t} [1 - \mathbb{1}_{\mathrm{IC}}(\mathbf{x}_i)](\hat{\mu}_{i,j})^m} \quad \forall j = 1, 2, \ldots, c_t, \tag{7}$$

$$\hat{\mathbf{v}}_j = \lambda_j \mathbf{v}_j^* + (1 - \lambda_j)\mathbf{v}_j \quad \forall j = 1, 2, \ldots, c_t, \tag{8}$$

where λ_j is determined by Equation (9):

$$\lambda_j = \frac{\sum_{\mathbf{x}_i \in X_t \cap C_j} [1 - \mathbb{1}_{\mathrm{IC}}(\mathbf{x}_i)](\hat{\mu}_{i,j})}{\sum_{\mathbf{x}_i \in C_j \setminus X_t} \mu_{i,j} + \sum_{\mathbf{x}_i \in X_t \cap C_j} [1 - \mathbb{1}_{\mathrm{IC}}(\mathbf{x}_i)](\hat{\mu}_{i,j})} \quad \forall j = 1, 2, \ldots, c_t. \tag{9}$$

Note that a data point $\mathbf{x}_i$ is assigned to a cluster C_j if and only if $j = \arg\max_{k=1,2,\ldots,c_t}\{\mu_{i,k}\}$, being C_j the set of data points that belongs to cluster j, $\forall j = 1, 2, \ldots, c_t$

As a last step of the algorithm, a cluster is deleted if it has been a predefined number of cycles, T, without receiving new objects. For this purpose, each cluster has a counter that includes the number of cycles it has been active without any update. When the counter reaches the value T, it is deleted by removing its center and all the data belonging to it from the data set.

3.1.4 Cluster interpretation

To characterize a cluster with numerical variables, e.g. X, value-tests (v-test) are computed for each of those variables using the following statistic:

$$\frac{\overline{X_C} - \overline{X}}{\sqrt{(1 - \frac{n_C}{n})\frac{s^2}{n_C}}} \sim t_{n_C - 1}, \tag{10}$$

where $\overline{X}$ is the mean of the variable X in the entire dataset, $\overline{X_C}$ is the mean of X within the cluster C, n_C is the number of objects in C, and s^2 is the global variance of X. The statistic follows a Student's t-distribution with $n_C - 1$ degrees of freedom, denoted by $t_{n_C - 1}$.

The v-test allows the interpretation of which variables characterize the clusters. If the value of the statistic in Equation (10) for a variable X in a cluster C is larger than 1.96, then it is interpreted that the variable characterizes the cluster. Additionally, the larger the value of the statistic, the better that variable characterizes the cluster, and the sign of the test indicates whether the variable is underrepresented (i.e., a negative sign) or overrepresented (i.e., a positive sign) in the given cluster, in comparison with all the data available for a given year. This statistic is very intuitive, as specific subpopulations may be over- or under-represented when all are compared at the city level.

4 Results

4.1 Results of the multivariate kernel density analysis

The results of the multivariate application of the kernel density estimation method are presented in Table 2, which shows in aggregate form, over the years measured, the mean values of the variables studied (i.e., migration background, age group, gender, and socio-economic factors), the standard deviation, the minimum, the maximum, and selected percentiles.

As seen in Table 2, the residential densities of the German subpopulations have higher mean values, while those of the Chinese subpopulations have the lowest mean values. It can also be observed that the population densities of individuals with a migrant background from Turkey have a higher standard deviation over the years, meaning there have been changes in the residential densities over time. Subpopulations with a migration background from Vietnam reach the maximum residential density, which can be interpreted as these communities locating in the specific residential areas, while those from Ukraine reach the minimum residential density. These statistics also show that, over time, Berlin has an average population of young adults aged between 30 and 35, and the sub-population of elderly people aged over ninety has lower average values.

In addition, the highest values are found in the population aged 50-55 years, which is the population living in common areas of the city. Finally, there are only marginal differences in the distribution between men and women in all the descriptive statistics. However, it should be noted that in some areas of the city, the female population peaks almost twice as high as the male population. Similarly, SGB II and SGB III show similar population densities reflecting socio-economic problems, although SGB II shows slightly higher values in some descriptive statistics.

4.2 Results based on the dynamic fuzzy *c*-means

4.2.1 Cluster validation

The Bezdek partition coefficient, an indicator defined by James Bezdek, was used to validate and quantify the quality of clustering solutions on our time-varying data sets. The Bezdek partition coefficient of a fuzzy c-partition of n data points is defined as [46]:

$$F(U;c) = \sum_{k=1}^{n} \sum_{i=1}^{c} \frac{u_{ik}^2}{n}, \tag{11}$$

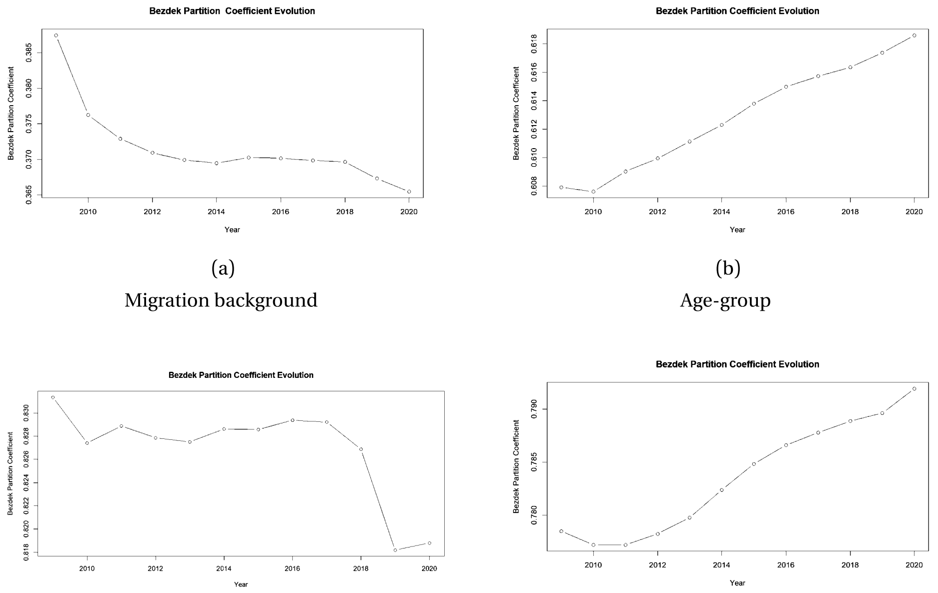
where u_{ik} is the membership of object i to cluster k, and c is the number of clusters. This index takes the value of 1 when the clusters are perfectly differentiated, and each object belongs only to a single cluster, and the value $1/c$ when each object belongs simultaneously to each cluster, so the distinction between the different clusters is undetermined. Therefore, the extreme values of the Bezdek Partition Coefficient allow evaluation of the quality of the clustering solution generated. Also, the partition coefficient depends on the number of clusters; the more clusters there are, the lower the value of the Bezdek index,

Table 2 Descriptive statistics of the estimated population densities for the dimension of migration background, age group, gender, and socio-economic variables between 2009 and 2020. The values in blue color indicate the maximum and the values in red indicate the minimum for each dimension

Dimension	Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Migration background	Germany	381,206	15.083	13.094	0	5.885	20.682	431.415
	France	381,206	12.636	23.468	0	1.144	9.679	165.712
	Greece	381,206	11.926	18.624	0	1.418	12.947	177.716
	Italy	381,206	12.568	20.972	0	1.26	11.187	149.503
	Austria	381,206	12.079	17.773	0	2.029	11.742	118.364
	Spain	381,206	12.267	23.261	0	1.085	8.931	162.154
	Poland	381,206	12.536	14.915	0	2.387	17.903	147.334
	Bulgaria	381,206	11.716	23.235	0	1.198	12.919	322.188
	Romania	381,206	11.514	18.308	0	1.737	14.081	334.725
	Croatia	381,206	11.735	16.593	0	1.238	14.481	118.275
	UK	381,206	12.485	21.817	0	1.215	10.86	158.278
	Bosnia and Herzegovina	381,206	11.484	18.942	0	1.029	13.764	288.607
	Serbia	381,206	11.574	18.398	0	1.043	14.347	257.834
	Russian Federation	381,206	11.743	16.09	0	1.853	15.023	234.288
	Ukraine	381,206	11.438	15.041	0	2.338	14.721	133.537
	Turkey	381,206	14.058	33.662	0	0.573	10.426	488.638
	USA	381,206	12.891	21.724	0	1.066	14.168	168.272
	Kazakhstan	381,206	11.392	21.542	0	1.506	11.066	284.247
	Iran	381,206	11.395	16.999	0	1.425	14.565	169.548
	Lebanon	381,206	12.207	30.851	0	0.255	8.543	347.993
	Syria	381,206	10.884	14.38	0	1.493	14.56	178.049
	Vietnam	381,206	11.279	25.054	0	0.789	10.736	596.92
	China	381,206	10.578	16.869	0	1.856	11.827	250.968
	Minorities	381,206	13.445	19.377	0	1.969	15.818	345.677
Age-group	Under_5	381,206	12.815	12.741	0	3.635	18.394	127.91
	From_05_to_10	381,206	12.677	11.284	0	4.574	17.365	93.586
	From_10_to_15	381,206	12.609	10.56	0	5.087	16.851	95.901
	From_15_to_20	381,206	12.519	10.488	0	5.123	16.572	92.951
	From_20_to_25	381,206	12.699	14.196	0	3.131	17.919	157.927
	From_25_to_30	381,206	13.368	17.784	0	2.118	18.189	207.023
	From_30_to_35	381,206	13.564	17.982	0	2.43	17.802	223.511
	From_35_to_40	381,206	13.426	16.044	0	3.255	17.251	191.242
	From_40_to_45	381,206	13.238	13.553	0	4.214	17.002	206.505
	From_45_to_50	381,206	13.167	11.568	0.004	5.05	17.4	224.742
	From_50_to_55	381,206	13.068	10.698	0	5.341	17.764	277.409
	From_55_to_60	381,206	12.927	10.54	0	5.301	17.66	262.978
	From_60_to_65	381,206	12.884	10.4	0	5.315	17.288	185.149
	From_65_to_70	381,206	12.827	10.104	0	5.403	17.14	196.703
	From_70_to_75	381,206	12.858	9.582	0	5.542	17.507	168.438
	From_75_to_80	381,206	12.677	9.372	0	5.338	17.646	194.289
	From_80_to_85	381,206	12.175	8.884	0	5.168	17.307	118.633
	From_85_to_90	381,206	11.706	8.281	0	5.112	16.636	86.574
	Over_90	381,206	11.225	8.347	0	4.729	15.685	59.25
Gender	Male	381,206	14.688	14.207	0.001	4.806	20.325	326.966
	Female	381,206	14.623	14.722	0	4.679	20.307	686.73
Socio-economic	SGB_II	381,206	10.828	12.171	0	2.405	14.766	174.377
	SGB_III	381,206	10.85	11.496	0	2.665	15.044	174.377

which means that clustering is fuzzy since its value is close to $1/c$. Figure 2 plots the evolution of the Bezdek Partition Coefficient over the years 2009 to 2020 for the dimensions considered.

The dynamic clustering algorithm tries to make decisions that do not worsen the Bezdek partition coefficient too much so that the partition continues to have a good level of quality. The initial cluster number of the algorithm was chosen considering the best Bezdek partition index obtained. The period of cycle T has been set to 20 in order to keep the data



(a)
Migration background

(b)
Age-group

(c)
Gender

(d)
Socio-economics

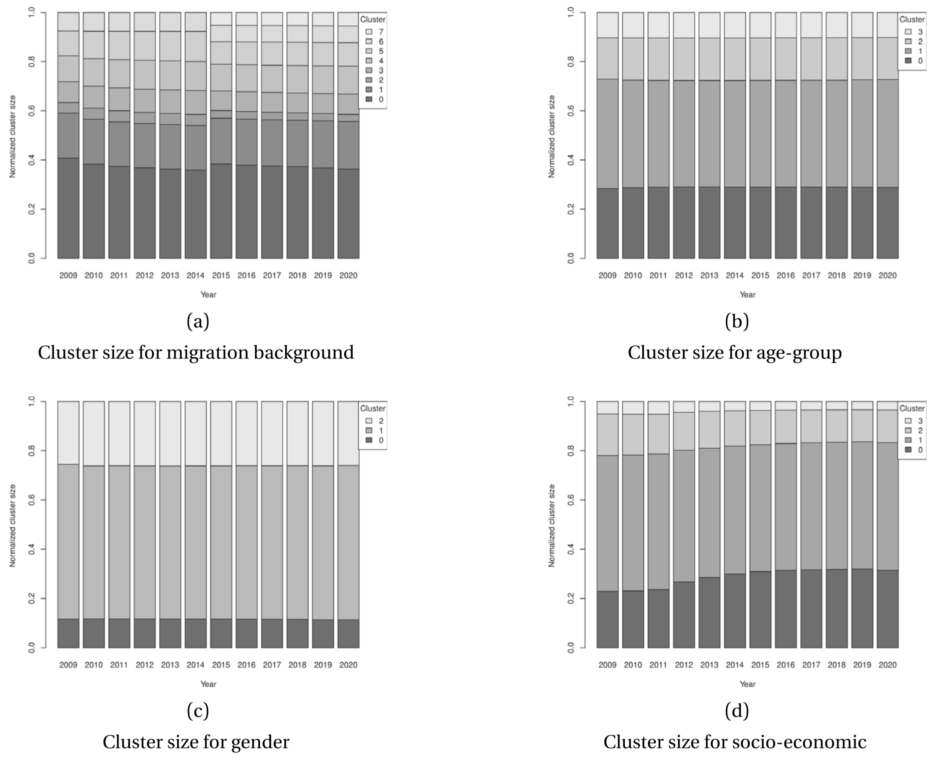
Figure 2 Cluster validation using Bezdek's partition coefficient for the dynamic clustering based on migration background, age group, gender, and socio-economic variables

up to date. At the same time, the different variables are updated since, as explained above, the algorithm deletes the classes that are not updated. This parameter avoids deleting data in the classes that are not updated, and this allows keeping the data for the analysis of new incoming data blocks.

The Bezdek partition coefficient indicates that the dynamic clustering solutions for the age group dimensions and the socio-economic variables improve over time. However, the coefficient behaves differently in the case of dynamic clustering based on migration background and clustering based on gender variables. In the case of dynamic clustering based on variables describing migration origin, it can be observed that the coefficient decreases until 2015 when the dynamic clustering algorithm detects the emergence of a new cluster, which reflects a new cluster structure. After this year, the coefficient improved and remained relatively stable but declined after 2018.

In the case of dynamic clustering based on gender variables, the Bezdek partition coefficient remains relatively stable over the years. It was only in 2017 that the coefficient values started to fall, but this reflects only a certain instability of the clusters. As we will see below, those clusters arise due more to changing population densities across the city than to uneven differences in the density of men and women in Berlin's residential areas, which are often remarkably similar.

The Bezdek partition coefficient generally shows that the dynamic clustering solutions obtained improve over time. It also shows us that when a decrease in the coefficient is observed, the emergence of a new cluster structure can be expected, in this case, when the migration background variables are considered. Therefore, the solutions obtained from the data represent valid cluster solutions.



(a)
Cluster size for migration background

(b)
Cluster size for age-group

(c)
Cluster size for gender

(d)
Cluster size for socio-economic

Figure 3 Normalized cluster sizes over time

4.2.2 Clustering results

To characterize the change of the clusters considering migration background, age group, gender, or socio-economic variables, we will present the clustering results in the years 2009 and 2020 in terms of their Mean in Cluster (MIC, the mean of a given variable in a given cluster) and the v-test, both indicating whether a variable is under- or over-represented in a given cluster and year and their corresponding statistical significance. Finally, the normalized size of the clusters (see, Fig. 3) and in absolute terms (see, tables in the Appendix) are presented.[2]

To visualize the micro changes and trends in the clusters over time, we generated a series of bump charts to describe which categories of variables were significant over time in each cluster. A Bump Chart "shows how quantitative category rankings have changed over time. They are typically structured around a temporal x-axis with equal intervals from the earliest to the latest. Quantitative rankings are plotted using joined-up lines that effectively connect consecutive points positioned along a y-axis (typically top = first)" [47]. After evaluation, each v-test value is assigned a rank, and each variable's ranks for a given year are plotted in descending order. The graph also shows that the values are grouped into different clusters based on a threshold to determine whether the variable is over-represented, underrepresented, or significant in a given cluster. To do this, the graphs use the critical values (i.e., 1.96 and –1.96 for a two-tailed test at a 5% significance level). Therefore:

[2] In this Section, we have chosen to present the results with additional textual elaboration. This decision was made to make the results accessible to all readers, including those who are visually impaired or unable to perceive visual representations such as maps. By providing detailed written descriptions, we hope to improve the comprehensibility and completeness of the information presented.

- If the v-test value is greater than 1.96, a variable is considered to be over-represented in a given cluster.
- If the v-test value is less than –1.96, a variable is considered to be underrepresented.
- If the v-test value is between –1.96 and 1.96, then a variable is not considered significant.

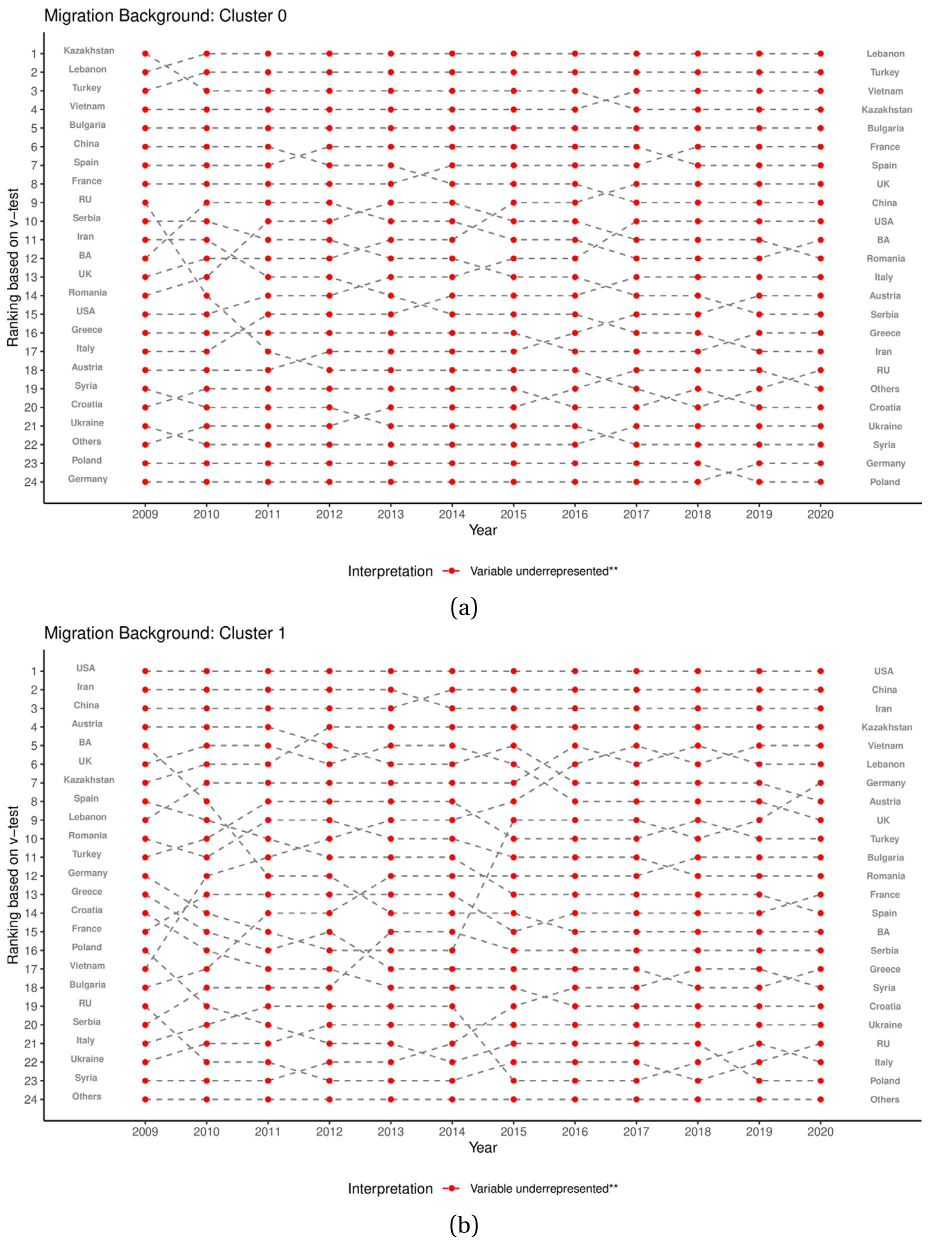
The bump chart is used here to visualize the micro-dynamics of residential segregation. On the Y-axis, the names of the variables are listed according to the value obtained for each year and cluster. This provides very informative visual information about the composition of the clusters because while the y-axis represents the ranking of the variables, the x-axis represents the years, and the connecting different lines show how the ranking of the different categories changes over time.

Results based on the migratory background. In terms of migrant background and clustered residential areas, the city of Berlin has a diverse and mixed population. The dynamic cluster analysis shows seven clusters from 2009 to 2014 and a total of eight clusters from 2015 to 2020. In 2015 a change in the clusters' structure was detected. The change corresponds to the emergence of `Cluster 7`. From a qualitative point of view, it can be seen that the change in the cluster structure occurred in the same year as the so-called European migration crisis.

`Cluster 0` is characterized by all migrant-related variables being underrepresented (see, Fig. 4a). In the year 2009, the three most under-represented variables correspond to Germans without a background of migration (MIC = 7.165; v-test = –96.753; p = 0.000), Poland (MIC = 2.181; v-test = –96.19; p = 0.000) and other subpopulations, while for the year 2020, the most underrepresented variables correspond to Poland (MIC = 2.052; v-test = –313.965; p = 0.000), Germans with no migration history (MIC = 6.301; v-test = –311.816; p = 0.000) and Syria (MIC = 1.702; v-test = –283.379; p = 0.000).

`Cluster 1` is also characterized by the under-representation of all variables related to the migrant background (see Fig. 4b). In 2009, the three most underrepresented variables corresponded to other minorities (MIC = 6.332; v-test = –30.513; p = 0.000), Syrians (MIC = 5.136; v-test = –29.968 p = 0.000) and Ukrainian subpopulations (MIC = 5.981, v-test = –29.255, p = 0.000). The least underrepresented are the USA (MIC = 12.166; v-test = –3.303; p = 0.001), Iran (MIC = 9.331; v-test = –10.596; p = 0.000) and China (MIC = 6.855; v-test = –16.39; p = 0.000). For 2020, the most underrepresented variables are other minorities category (MIC = 5.132; v-test = –129.563; p = 0.000), Poland (MIC = 6.504; v-test = –122.126; p = 0.000), and Italy (MIC = 4.154; v-test = –121.158; p = 0.000).

In `Cluster 2`, all variables are overrepresented, except for Kazakhstan, which ranks last and is underrepresented for all measured years (see Fig. 4c). The most overrepresented subpopulations in 2009 are Iran (MIC = 76.504; v-test = 123.307; p = 0.000), Ukraine (MIC = 56.573; v-test = 104.875; p = 0.000), China (MIC = 54.082; v-test = 84.846; p = 0.000), USA (MIC = 60.44; v-test = 84.445; p = 0.000) and Austria (MIC = 51.335; v-test = 83.014; p = 0.000), and other subpopulations. Similarly, the most overrepresented variables for the year 2020 correspond to those of Iran (MIC = 74.076; v-test = 394.988; p = 0.000), Ukraine (MIC = 62.733; v-test = 365.308; p = 0.000), China (MIC = 56.408; v-test = 291.015; p = 0.000), Greece (MIC = 54.289; v-test = 243.657; p = 0.000) and Austria (MIC = 52.133; v-test = 241.398; p = 0.000), among other variables that characterize this cluster.


Migration Background: Cluster 0

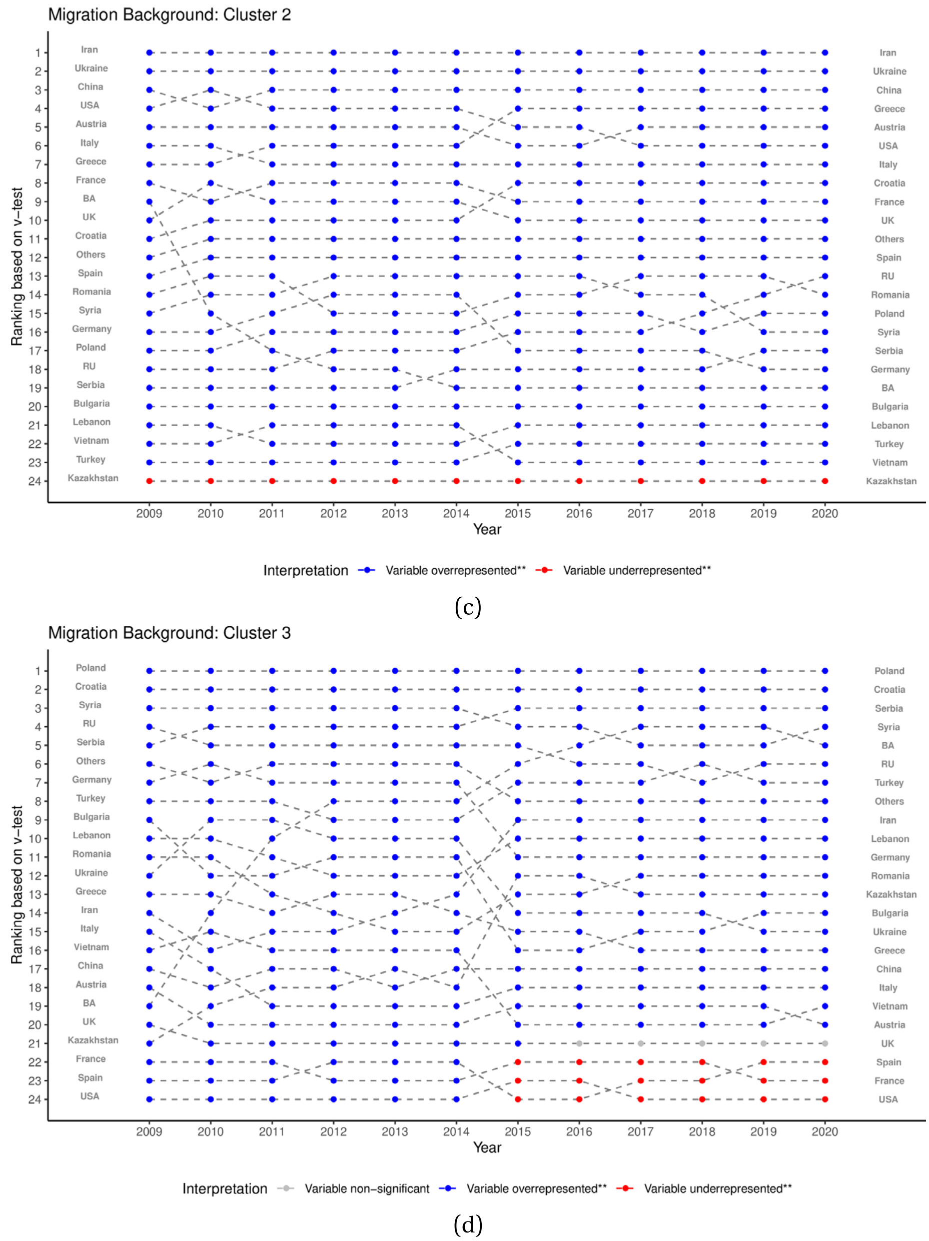
(a)

Migration Background: Cluster 1

(b)

Figure 4 Ranking of migration background variables by cluster and year

In Cluster 3, as all variables are statistically significant, all variables characterize this cluster (see Fig. 4d). In 2009, all the variables of the migratory background were overrepresented, as in the case of Poland (MIC = 35.064; v-test = 73.191; p = 0.000), Croatia (MIC = 31.789; v-test = 56.777; p = 0.000), Syria (MIC = 27.023; v-test = 49.38; p = 0.000), RU (MIC = 32.033; v-test = 49.372; p = 0.000) and Serbia (MIC = 32.033; v-test = 49.372; p = 0.000). From 2015 to 2020, a change was observed as the USA, France, and Spain subpopulations became underrepresented. It is also observed that between 2015 and 2020, the United Kingdom no longer represents this cluster. For the year 2020, it is observed that the subpopulations of Poland (MIC = 32.751; v-test = 251.125; p = 0.000), Croatia (MIC =



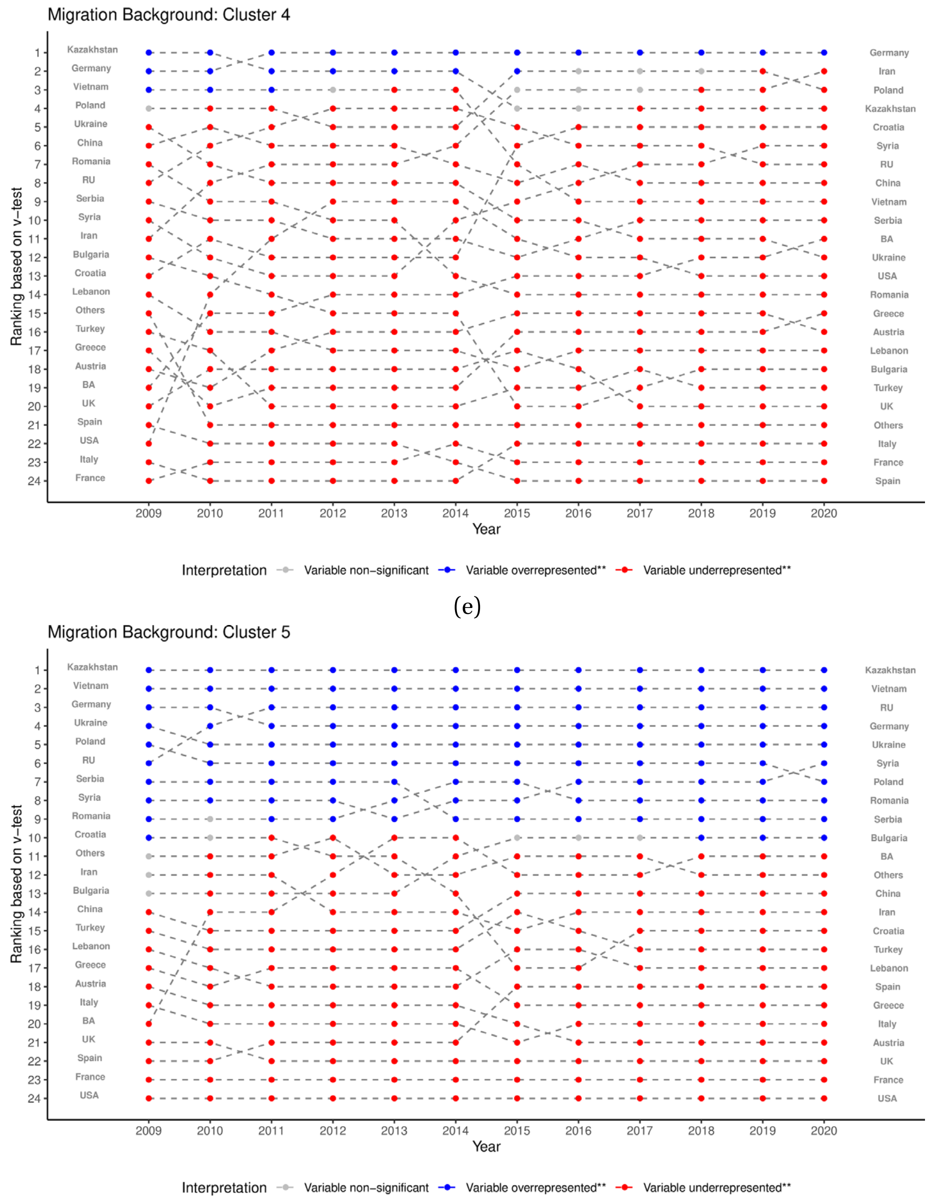
(c)

(d)

Figure 4 Continued

27.117; v-test = 171.742; p = 0.000), Serbia (MIC = 25.932; v-test = 144.585; p = 0.000), Syria (MIC = 20.281; v-test = 121.066; p = 0.000), and BA (MIC = 23.835; v-test = 120.809; p = 0.000) are the five most overrepresented subpopulations in this cluster.

Most of the variables in `Cluster 4` are under-represented, although there are a few over-represented variables (see, Fig. 4e). In 2009, the variables Kazagastan (MIC = 19.642; v-test = 21.459; p = 0.000), Germans without a migration background (MIC = 18.297; v-test = 14.523; p = 0.000), and Vietnam (MIC = 16.58; v-test = 14.282; p = 0.000) are over-represented. For the same year, the most underrepresented variables are the subpopulations of France (MIC = 6.218; v-test = –17.707; p = 0.000), Italy (MIC = 6.797; v-test =



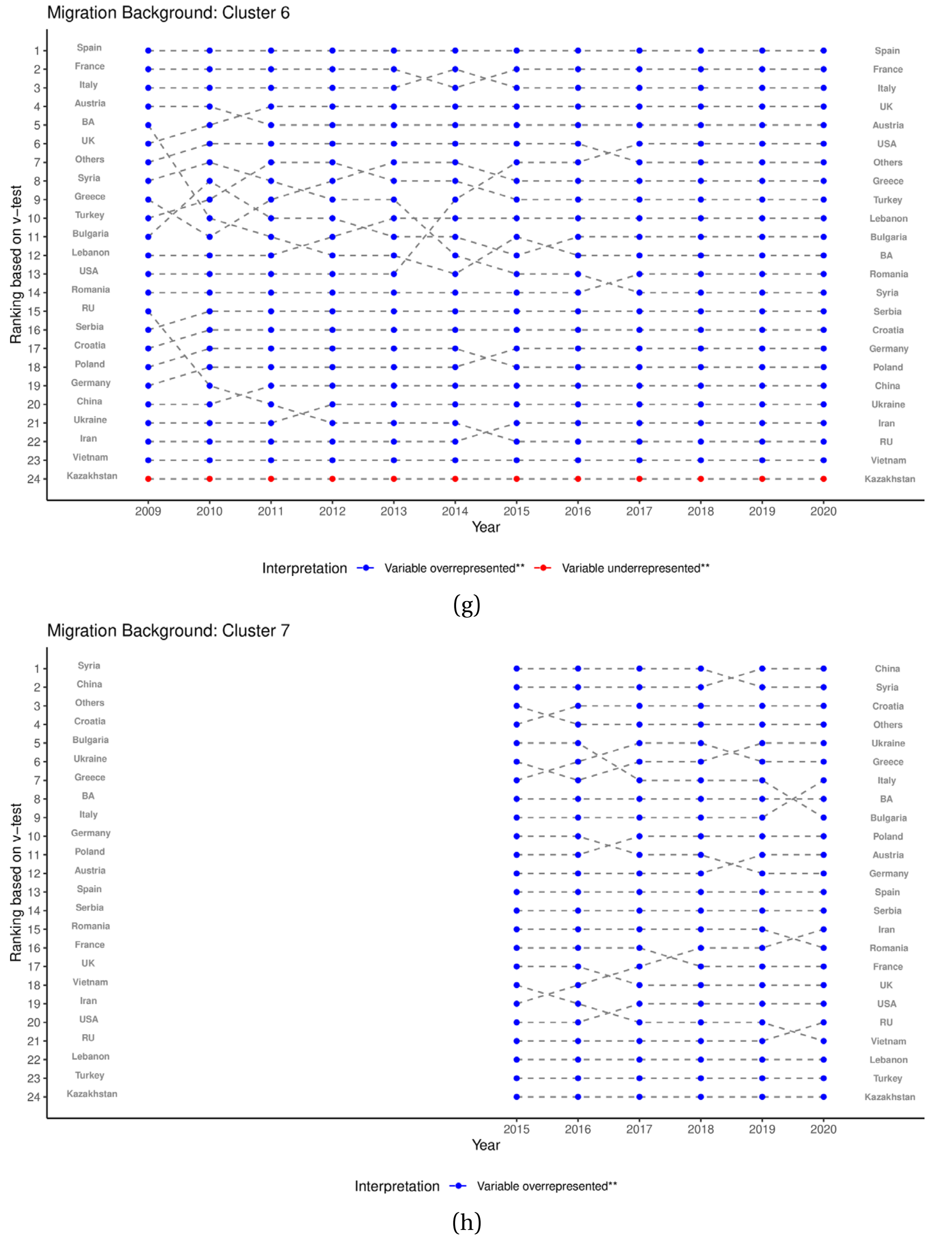
(e)

(f)

Figure 4 Continued

−17.258; p = 0.000), USA (MIC = 7.334; v-test = −16.322; p = 0.000), Spain (MIC = 5.953; v-test = −16.258; p = 0.000) and UK (MIC = 7.218; v-test = −15.911; p = 0.000), among others. By 2020, the only over-represented sub-population is the German sub-population without a migration background (MIC = 15.596; v-test = 8.687; p = 0.000), while the Spanish (MIC = 5.938; v-test = −60.339; p = 0.000), French (MIC = 5.938; v-test = −58.254; p = 0.000) and Italian subpopulations (MIC = 7.178; v-test = −56.994; p = 0.000) are the most underrepresented.

There is a mixture of over- and under-represented variables in `Cluster 5` (see Fig. 4f). The most overrepresented subpopulations in 2009 are Kazakhstan (MIC = 46.609; v-test =

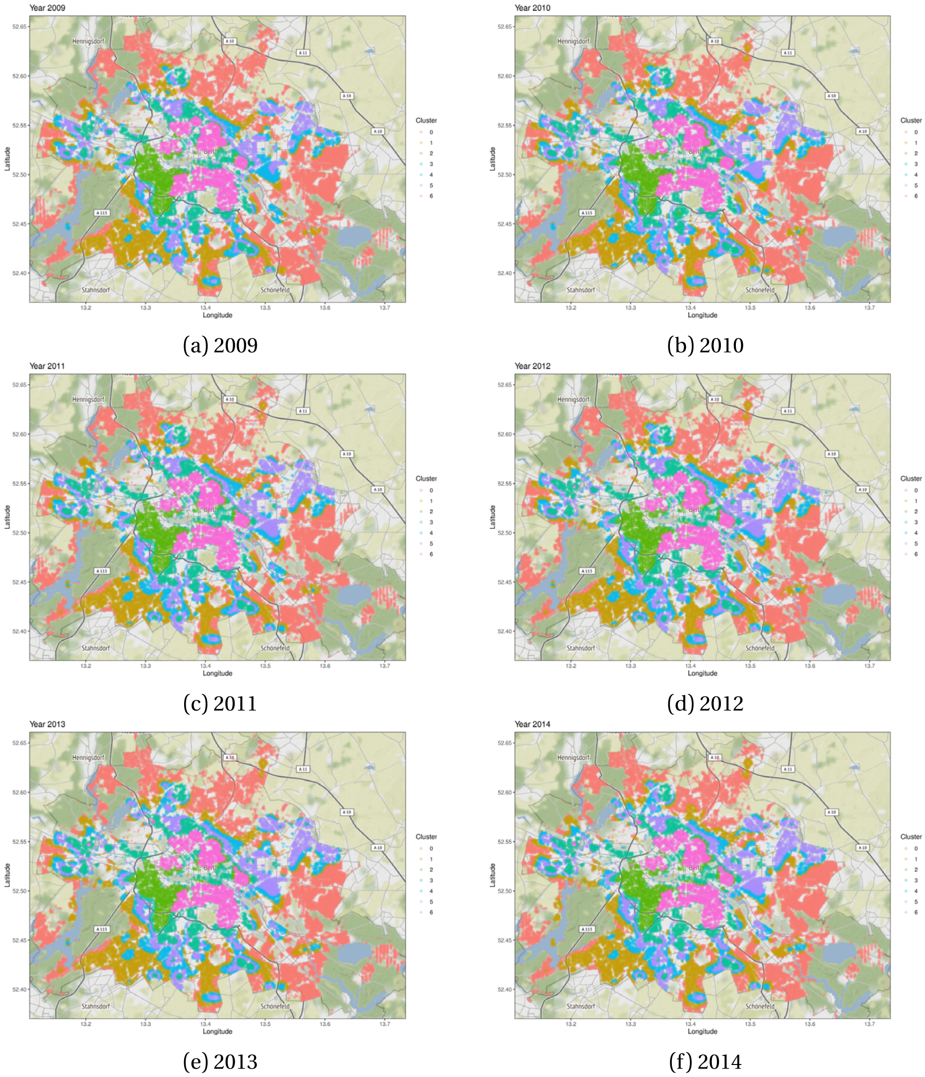


(g)

(h)

Figure 4 Continued

91.192; p = 0.000), Vietnam (MIC = 35.34; v-test = 62.19; p = 0.000) and Germans without a migration background (MIC = 23.93; v-test = 41.097; p = 0.000), among others. In the same year, the most underrepresented subpopulations are the USA (MIC = 8.197; v-test = −13.493; p = 0.000), France (MIC = 8.265; v-test = −11.823; p = 0.000) and Spain (MIC = 8.126; v-test = −10.444; p = 0.000), among other subpopulations. For the year 2020, the three most overrepresented subpopulations are Kazakhstan (MIC = 48.736; v-test = 346.537; p = 0.000), Vietnam (MIC = 45.111; v-test = 269.945; p = 0.000) and RU (MIC = 31.451 v-test = 244.861 p = 0.000), while the USA is the most underrepresented (MIC = 5.524; v-test = −67.792; p = 0.000), followed by France (MIC = 5.78; v-test = −58.401; p =



(a) 2009

(b) 2010

(c) 2011

(d) 2012

(e) 2013

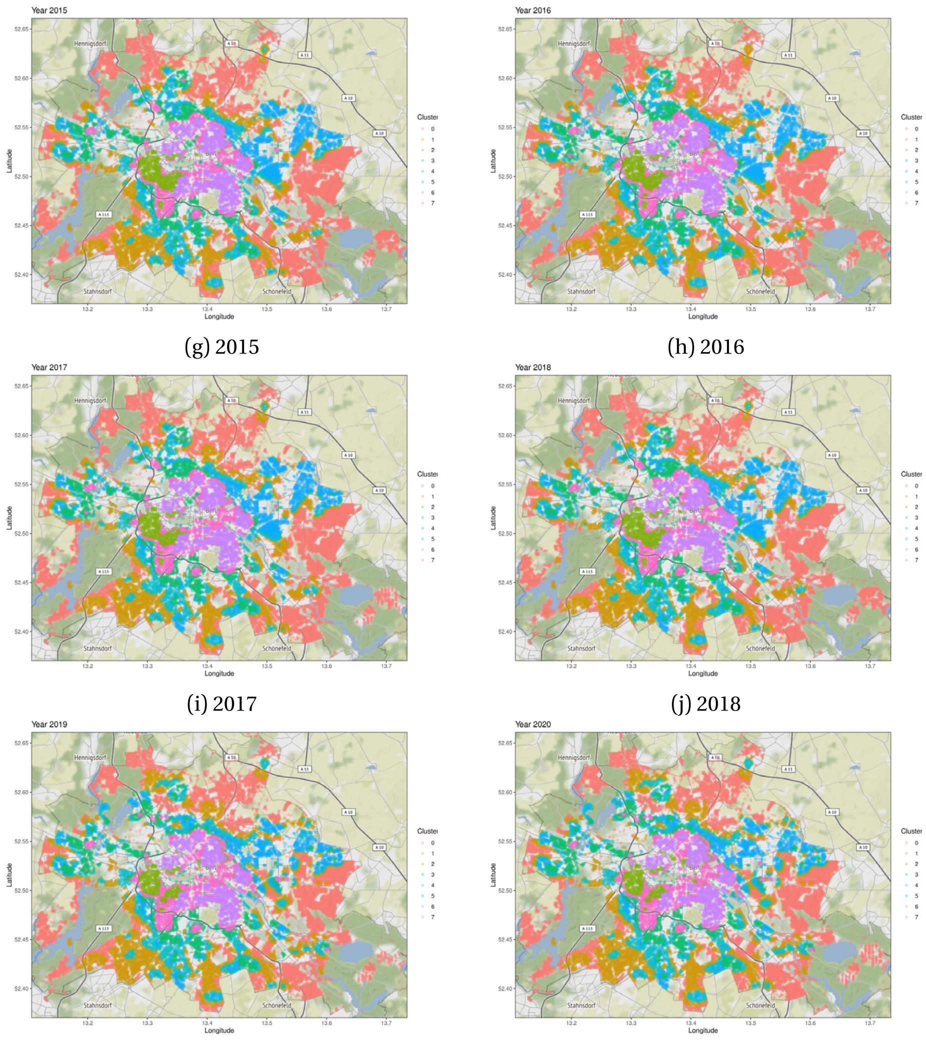
(f) 2014

Figure 5 Dynamic clustering results visualized according to migration background variables

0.000) and UK (MIC = 6.468; v-test = –55.138; p = 0.000). Finally, the variables of Romania, Croatia, Iran, Bulgaria, and other minorities are not always characteristic of this cluster over time.

During 2009, `Cluster 6` was characterized by almost all subpopulations being overrepresented, except for Kazakhstan (MIC = 10.477; v-test = –7.066; p = 0.000), which was the only one underrepresented (see Fig. 4g). In 2020, the three most overrepresented subpopulations were Spain (MIC = 79.323; v-test = 479.471; p = 0.000), France (MIC = 80.279; v-test = 479.408; p = 0.000), and Italy (MIC = 71.934; v-test = 470.824; p = 0.000), along with other subpopulations.

Most interestingly, the emergence of `cluster 7` in 2015 was revealed by the dynamic cluster analysis. In this cluster, all subpopulations are representative and overrepresented. In 2015, this cluster has Syria (MIC = 30.966; v-test = 142.006; p = 0.000) as the most overrepresented variable, and the second most overrepresented nation is China (MIC = 31.461; v-test = 131.278; p = 0.000) and the third most overrepresented variable is other



(g) 2015

(h) 2016

(i) 2017

(j) 2018

(k) 2019

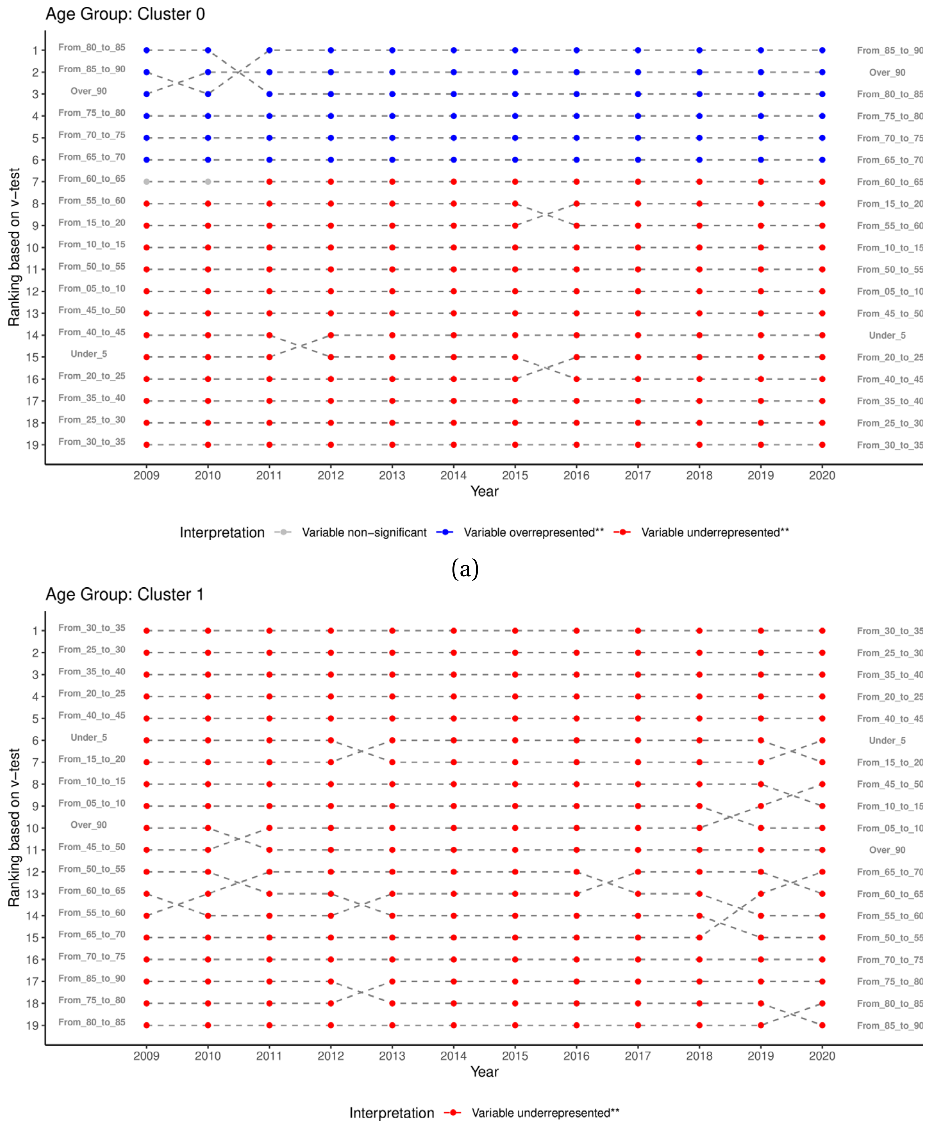
(l) 2020

Figure 5 Continued

minorities (MIC = 37.489; v-test = 130-72; p = 0.000). The least overrepresented variable is Kazakhstan (MIC = 13.077; v-test = 7.84; p = 0.000). In the year 2020, this cluster is characterized by China as the most overrepresented variable (MIC = 32.429; v-test = 193.024; p = 0.000). The second most overrepresented variable is Syria (MIC = 28.793; v-test = 185.59; p = 0.000), and the third most overrepresented variable is Croatia (MIC = 32.348; v-test = 185.12; p = 0.000). The least overrepresented variable is again Kazakhstan (MIC = 13.718; v-test = 16.089; p = 0.000).

Finally, the visualization of the clusters is shown in the maps in Fig. 5, and the normalized size of clusters over time is shown in Fig. 3a.

Results based on the age group. The dynamic cluster analysis revealed that the population of Berlin is grouped in residential areas in a structure of four different clusters of age groups. Qualitatively, `Cluster 3` is located in the city centre, `Cluster 2` is located



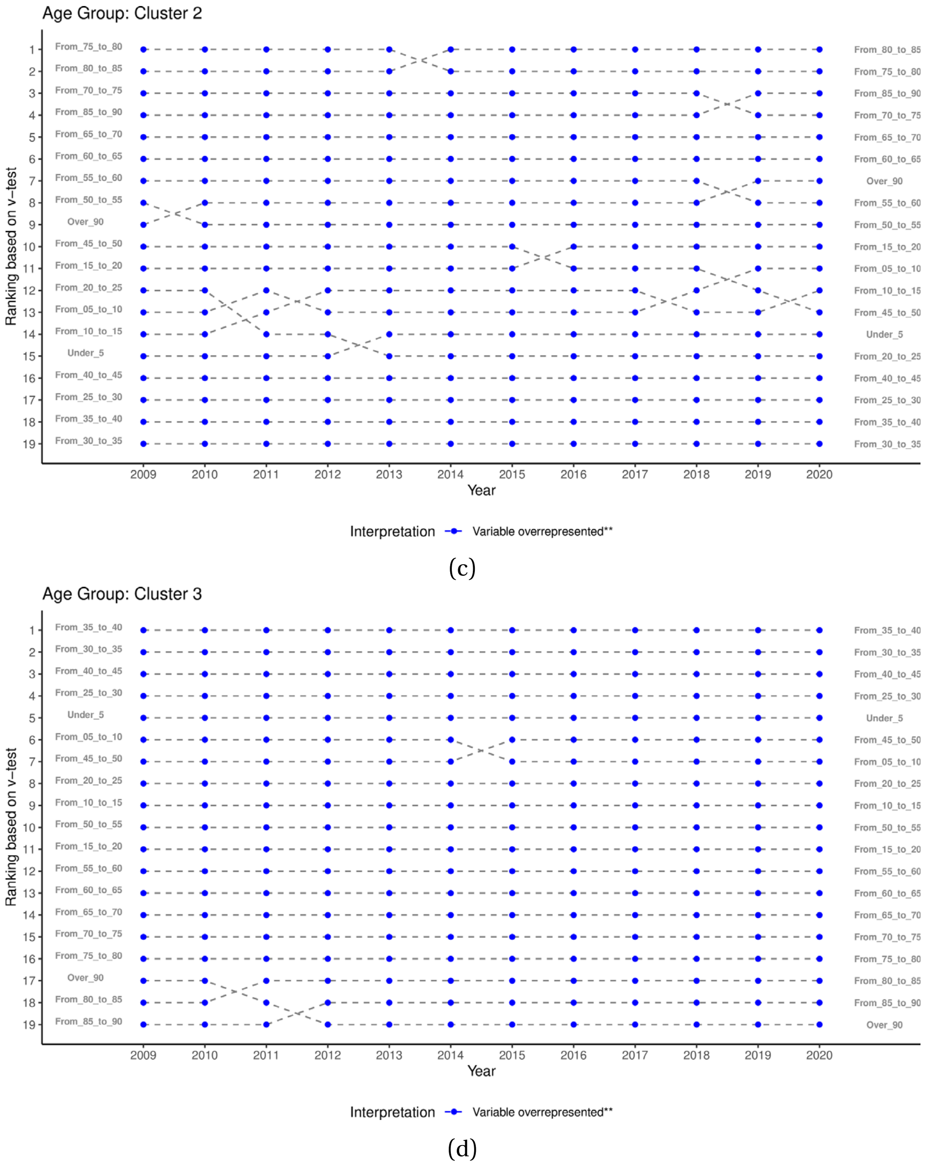
(a)

(b)

Figure 6 Ranking of age group variables by cluster and year

around the city centre, surrounded by `Cluster 0`. Finally, `Cluster 1` is located on the city's outskirts.

`Cluster 0` has only a few overrepresented variables and several underrepresented ones (see Fig. 6a). Analysis using the value test shows that in 2009, subpopulations in `Cluster 0` ranging from 80 to 85 are the most overrepresented (MIC = 14.578; v-test = 31.307; p = 0.000), and subpopulations ranging from 30 to 35 are the most underrepresented (MIC = 9.79; v-test = –25.043; p = 0.000), and subpopulations aged 60 to 65 were not significant (MIC = 13.201; v-test = 0.491; p = 0.623). For the year 2020, subpopulations between 85 and 90 years are the most overrepresented (MIC = 13.909; v-test = 104.682; p = 0.000), and subpopulations between 30 and 35 years are the most underrepresented (MIC = 9.327; v-test = –92.725; p = 0.000), showing ageing of the cluster compared to 2009.



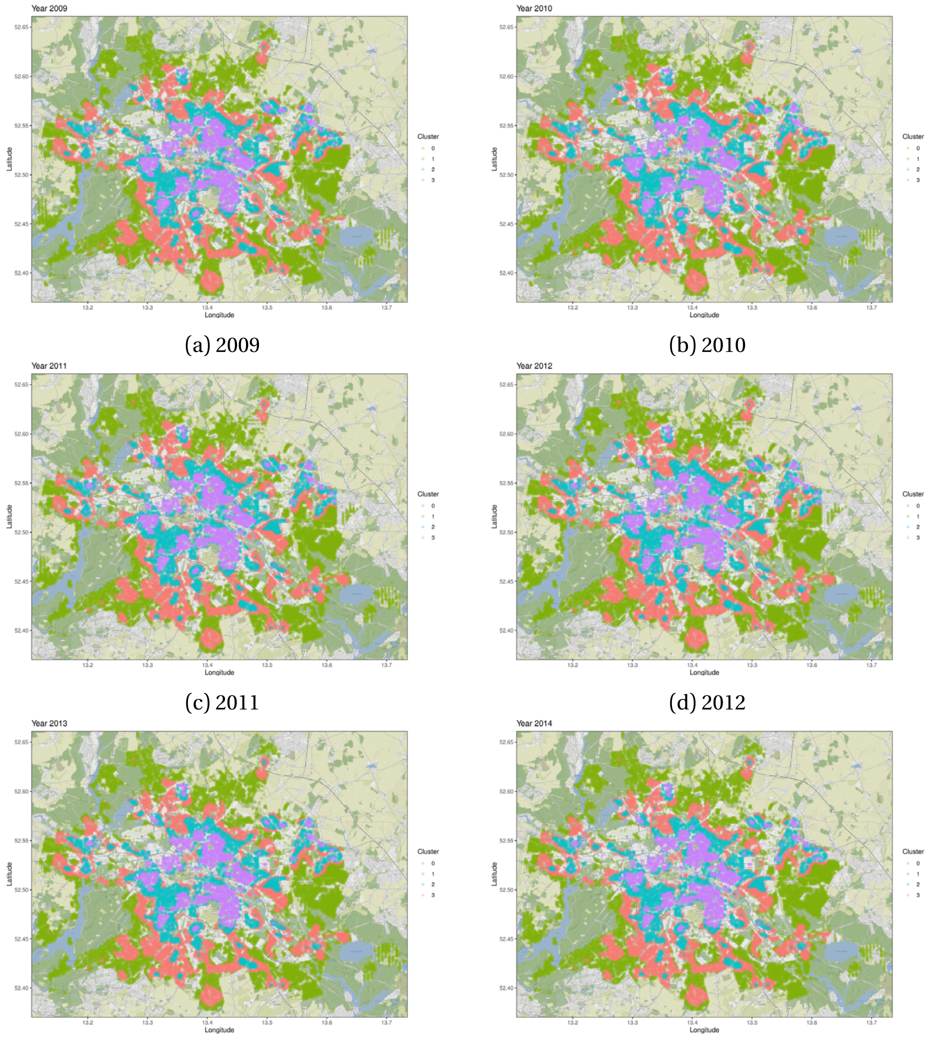
(c)

(d)

Figure 6 Continued

Cluster 1 has all variables underrepresented (see Fig. 6b). For the year 2009 in Cluster 1, the subpopulations between 30 and 35 are the least underrepresented (MIC = 2.804; v-test = –96.097; p = 0.000), and the most underrepresented are the subpopulations between 80 and 85 (MIC = 5.244; v-test = –130.039; p = 0.000). By 2020, the least underrepresented subpopulations in Cluster 1 are those between 30 and 35 (MIC = 2.541; v-test = –334.159; p = 0.000), and the most underrepresented are those between 85 and 90 (MIC = 5.062; v-test = –437.36; p = 0.000).

Cluster 2 has all variables overrepresented during 2009 (see Fig. 6c). For the year 2020, the most overrepresented age groups are the 80 to 85-year-olds (MIC = 22.479; v-test = 325.474; p = 0.000), and the least overrepresented groups are the 30 to 35-year-olds (MIC = 23.894; v-test = 161.203; p = 0.000).



(a) 2009

(b) 2010

(c) 2011

(d) 2012

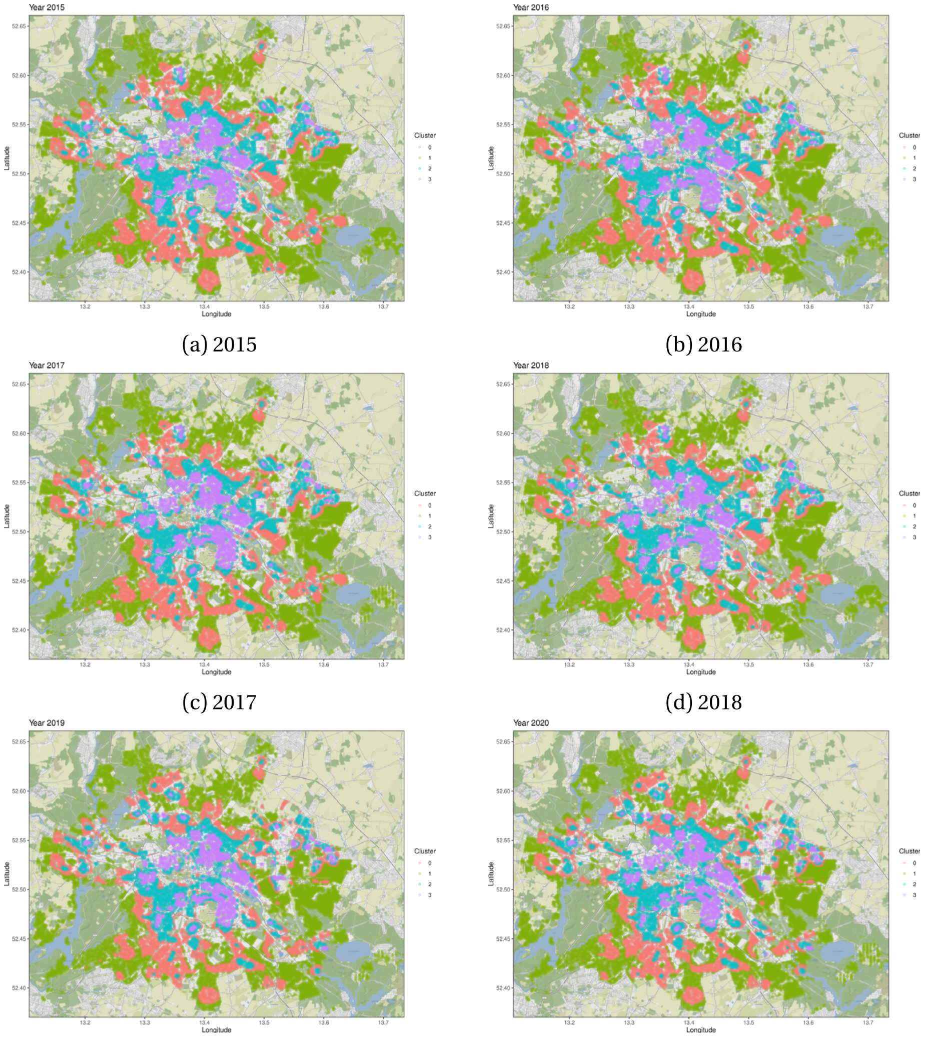
(e) 2013

(f) 2014

Figure 7 Dynamic clustering results visualized according to age-group variables

Similarly, in `Cluster 3`, all variables are overrepresented (see Fig. 6d). For the year 2009, the subpopulations from 35 to 40 are the most overrepresented (MIC = 51.23; v-test = 142.501; p = 0.000), and the subpopulations from 85 to 90 are the least overrepresented (MIC = 18.939; v-test = 49.795; p = 0.000), showing that it is a representative cluster of adults. Similarly, in the year 2020, the subpopulations ranging from 35 to 40 are the most overrepresented (MIC = 51.42; v-test = 492.401; p = 0.000), and subpopulations over 90 are the least overrepresented (MIC = 17.28; v-test = 150.838; p = 0.000).

The visualization of the clusters on the map of Berlin is shown in Fig. 7, and the normalized size of clusters over time is shown in Fig. 3b. Interestingly, there is no change in the number of clusters over time, but there is an increase in the overall population density. This means that the cluster structure based on the age group dimension remains stable over the period observed.



(a) 2015

(b) 2016

(c) 2017

(d) 2018

(e) 2019

(f) 2020

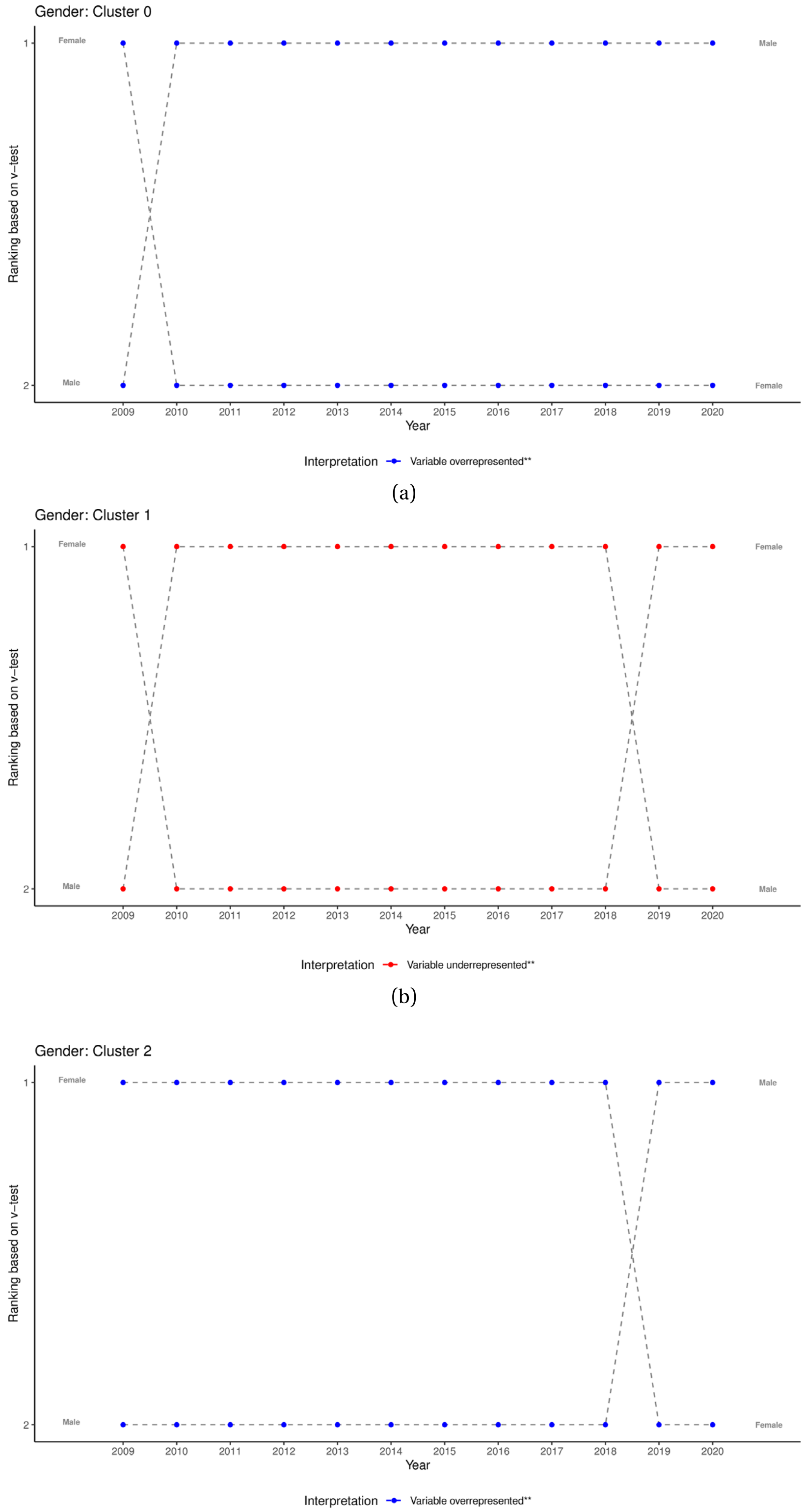
Figure 7 Continued

Results based on gender. The results of the cluster analysis allowed the identification of 3 clusters.

In general, the cluster analysis shows that the clusters represent the population density in residential areas. In other words, the clusters divided into male and female population densities correspond to Berlin's more or less densified areas. In the case of the clusters, the marginal differences over time are reported below.

In `Cluster 0`, both variables are overrepresented (see Fig. 8a). For the year 2009, both male (MIC = 14.92; v-test = 139.279; p = 0.000) and female (MIC = 14.92; v-test = 139.279; p = 0.000) subpopulations are equally overrepresented in this cluster. For the year 2020, the male population (MIC = 45.058; v-test = 471.478; p = 0.000) is more overrepresented than the female population (MIC = 461.205; v-test = 461.205; p = 0.000).

`Cluster 1` has both variables underrepresented (see, Fig. 8b). Both male and female residential population densities reached the same values in 2009 (MIC = 6.55; v-test = –137.675; p = 0.000). However, for 2020, the male population (MIC = 6.316; v-test =

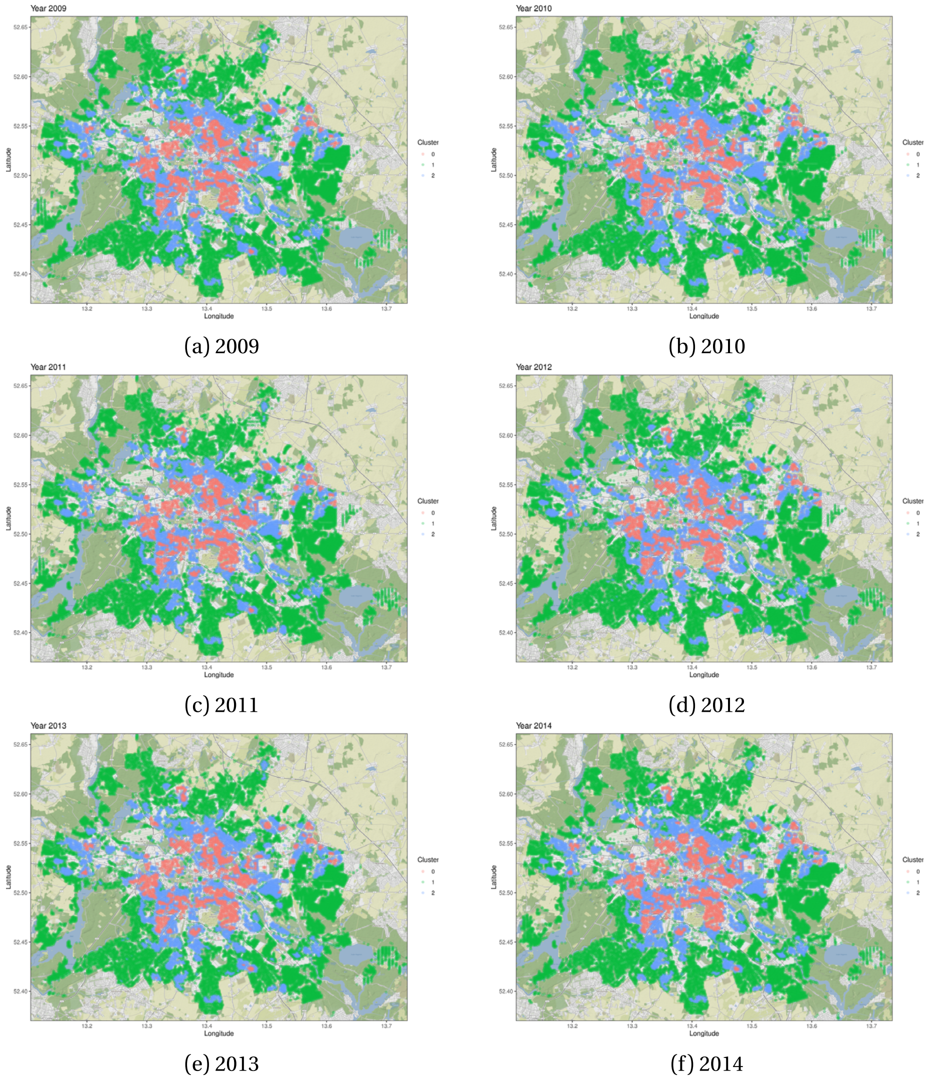


(a)

(b)

(c)

Figure 8 Ranking of gender variables by cluster and year



(a) 2009 (b) 2010

(c) 2011 (d) 2012

(e) 2013 (f) 2014

Figure 9 Dynamic clustering results visualized according to gender variables

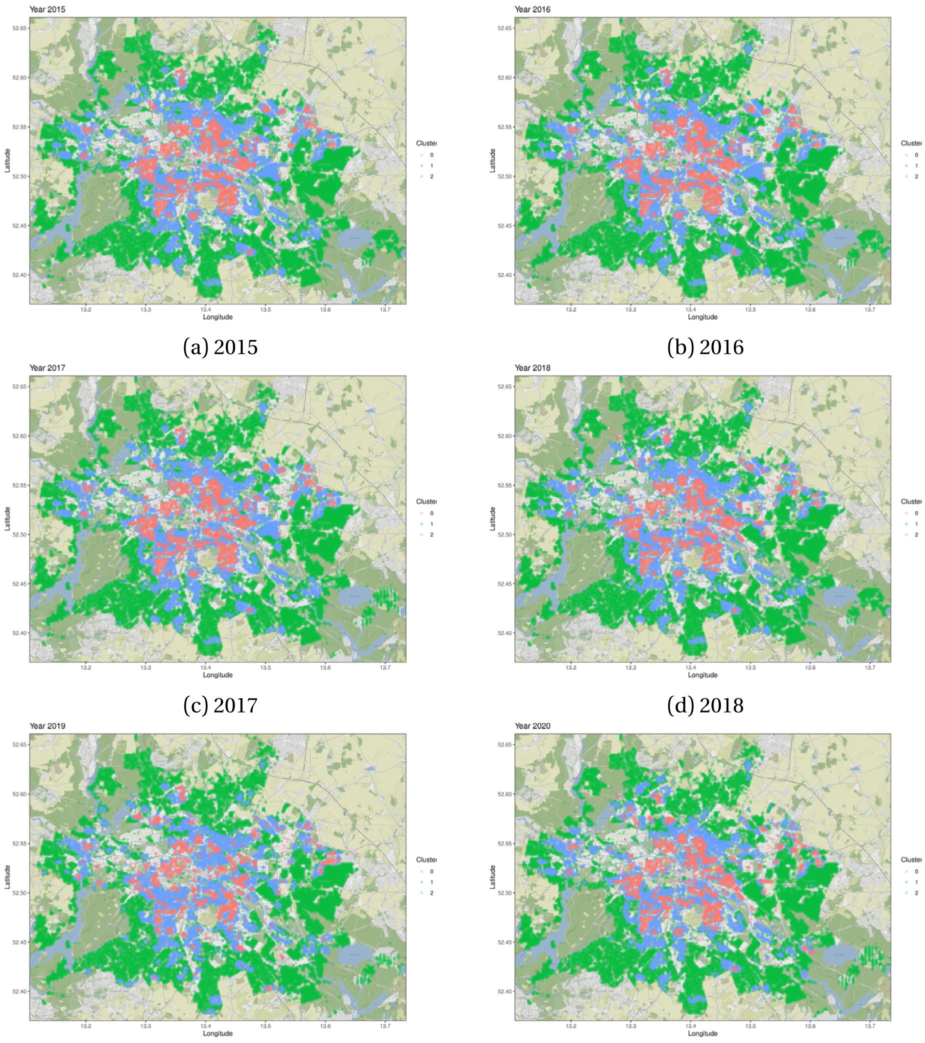
–471.675; p = 0.000) is only slightly more underrepresented than the female population (MIC = 6.233; v-test = –456.146; p = 0.000).

In `Cluster 2`, both variables are overrepresented (see Fig. 8c). Both male and female populations had the same spatial density for 2009 (MIC = 21.696; v-test = 50.164; p = 0.000). However, for 2020, male populations (MIC = 21.66; v-test = 179.541; p = 0.000) are more overrepresented than female populations (MIC = 21.458; v-test = 169.838; p = 0.000).

The visualization of the clusters on the map of Berlin is shown in Fig. 9, and the normalized size of clusters over time can be seen in Fig. 3c.

Results based on socio-economics. For the socio-economic dimension, the cluster analysis resulted in the identification of four clusters.

From a qualitative point of view, `Cluster 3` represents the places with the most significant socio-economic problems. It can be seen that the areas corresponding to clusters



(a) 2015

(b) 2016

(c) 2017

(d) 2018

(e) 2019

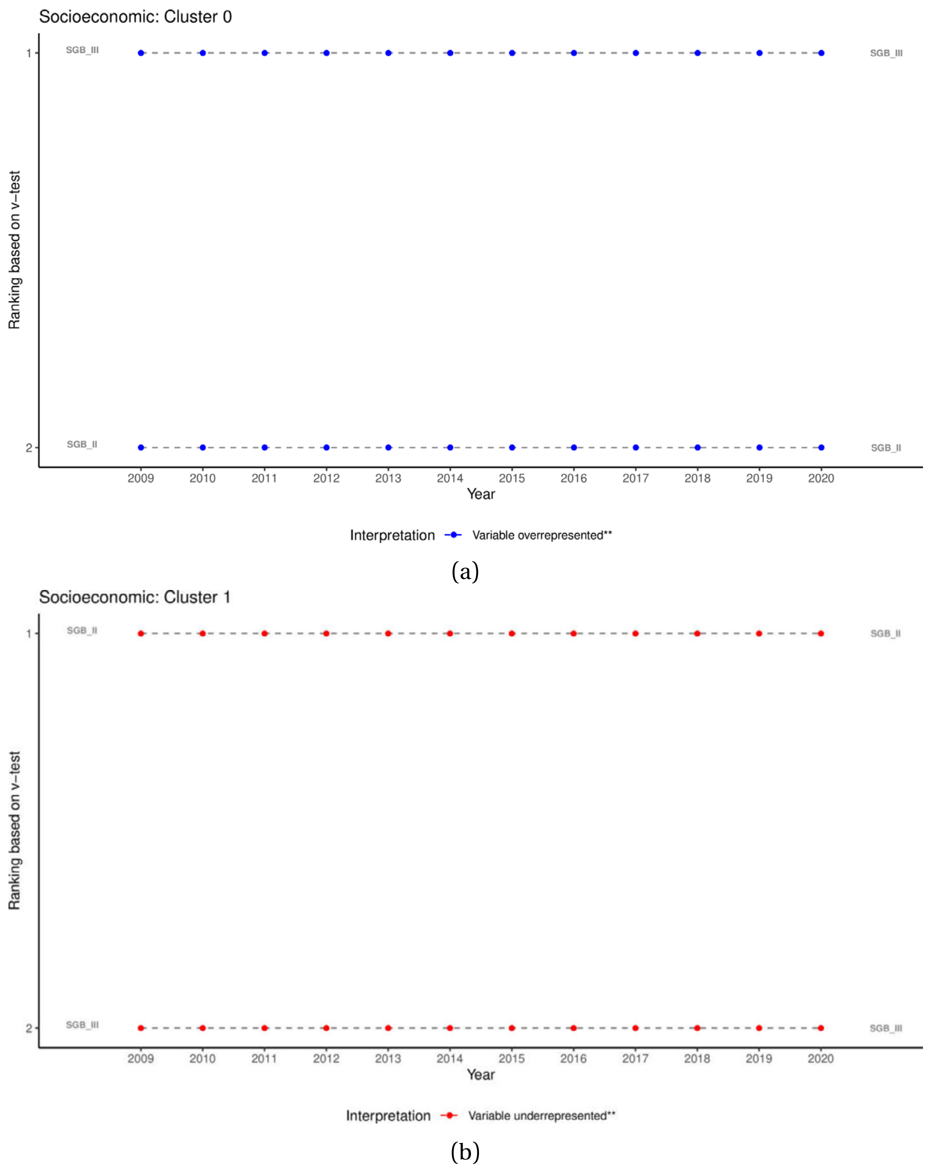
(f) 2020

Figure 9 Continued

2 and 3 have a larger area in 2009, after the global subprime crisis, and the onset of the COVID-19 pandemic in 2020 these clusters have slightly different shapes.

In `Cluster 0`, both variables are overrepresented and statistically significant (see Fig. 10a). In 2009, SGB III was the most overrepresented (MIC = 13.564; v-test = 15.813; p = 0.000), and SGB II was the least overrepresented (MIC = 12.404; v-test = 6.638; p = 0.000). For the year 2020, SGB III is the most overrepresented (MIC = 12.727; v-test = 68.251; p = 0.000), and SGB II is the least overrepresented (MIC = 12.423; v-test = 54.825; p = 0.000), with a decrease in the former group, and an increase in the latter since 2009.

On the contrary, in `Cluster 1`, the variables are underrepresented and statistically significant (see Fig. 10b). For 2009, `Cluster 1` had SGB II as the least underrepresented socio-economic variable (MIC = 2.388; v-test = –123.634; p = 0.000), and SGB III as the most underrepresented (MIC = 3.107; v-test = –130.744; p = 0.000). Similarly, for 2020, SGB II was the least underrepresented (MIC = 3.012; v-test = –411.076; p = 0.000), and SGB III was the most underrepresented (MIC = 3.227; v-test = –424.48; p = 0.000).



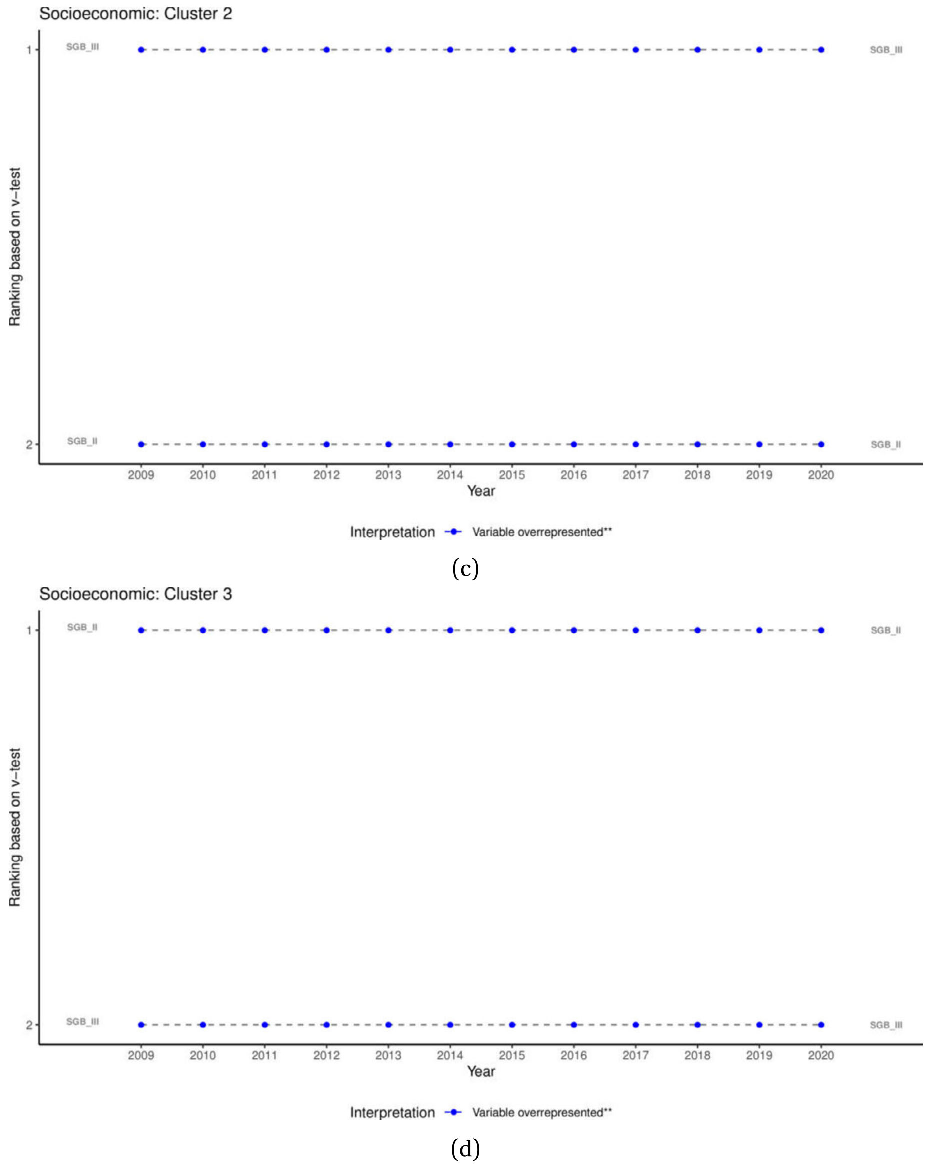
(a)

(b)

Figure 10 Ranking of socio-economic variables by cluster and year

In `Cluster 2`, both variables are overrepresented and statistically significant (see Fig. 10c). For 2009, SGB III was the most overrepresented (MIC = 25.641; v-test = 89.369; p = 0.000), and SGB II was the least overrepresented (MIC = 26.481; v-test = 83.642; p = 0.000). Similarly, in 2020, SGB III was the most overrepresented (MIC = 25.985; v-test = 318.287; p = 0.000), and SGB II was the least overrepresented (MIC = 26.169; v-test = 304.748; p = 0.000).

Finally, in `Cluster 3`, both variables are statistically significant and overrepresented (see Fig. 10d). In 2009, SGB II was the most overrepresented (MIC = 55.375; v-test = 125.056; p = 0.000) and SGB III was the least overrepresented (MIC = 46.651; v-test = 113.809; p = 0.000). The same situation occurred in 2020, where SGB II was the most overrepresented (MIC = 54.557; v-test = 419.17; p = 0.000), and SGB III was the least



(c)

(d)

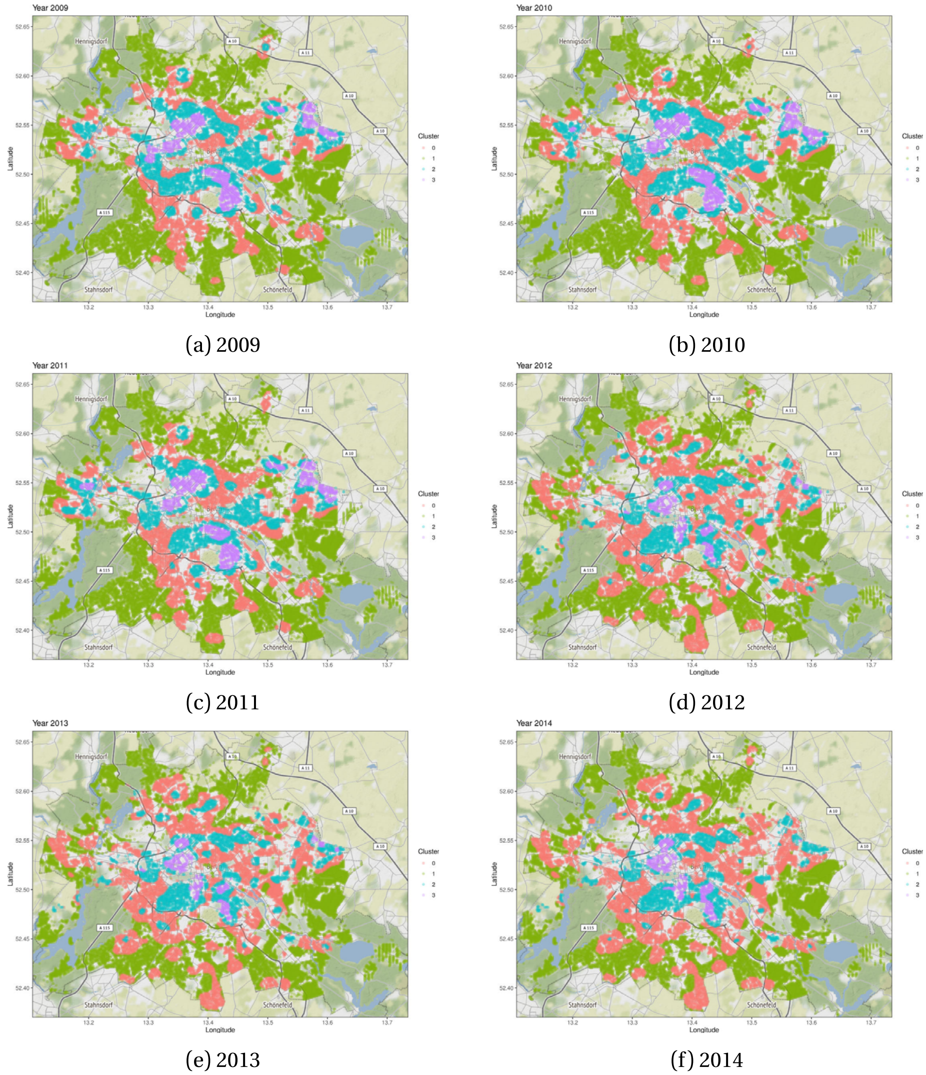
Figure 10 Continued

overrepresented (MIC = 49.921; v-test = 396.516; p = 0.000). The map of Berlin is shown in Fig. 11, and Fig. 3d shows the normalized size of clusters over time.

In summary, the maps show that residential segregation in Berlin is a phenomenon that can be visualized on a geographical level. The analysis also detected the emergence of a cluster when analysing the migration background of Berlin's populations. Finally, the results show that the clusters have small movements because the composition of the clusters changes over time and space.

5 Discussion and conclusion

This study aimed to examine the phenomenon of residential segregation from a dynamic point of view. According to our approach, residential segregation can be explored from



(a) 2009

(b) 2010

(c) 2011

(d) 2012

(e) 2013

(f) 2014

Figure 11 Dynamic clustering results visualized according to socio-economic variables

different angles, for example, from the side of migration background, age group, gender, or other variables describing the economic situation of the population under study.

To open the discussion, we would like to recall that several studies have been carried out on the spatial distribution of Berlin's subpopulations. In particular, we believe that the reporting of spatial densities excluding non-residential areas, the separate analysis of dimensions that has already been documented by several researchers, and the use of dynamic rather than static cluster analysis are aspects that can help different disciplines, especially those that are looking for novel methodological new methodological ways to identify changes in population structure from cohort data. In this context, we briefly discuss some of the findings and then and then summarise the research undertaken.

5.1 Comparison with previous research

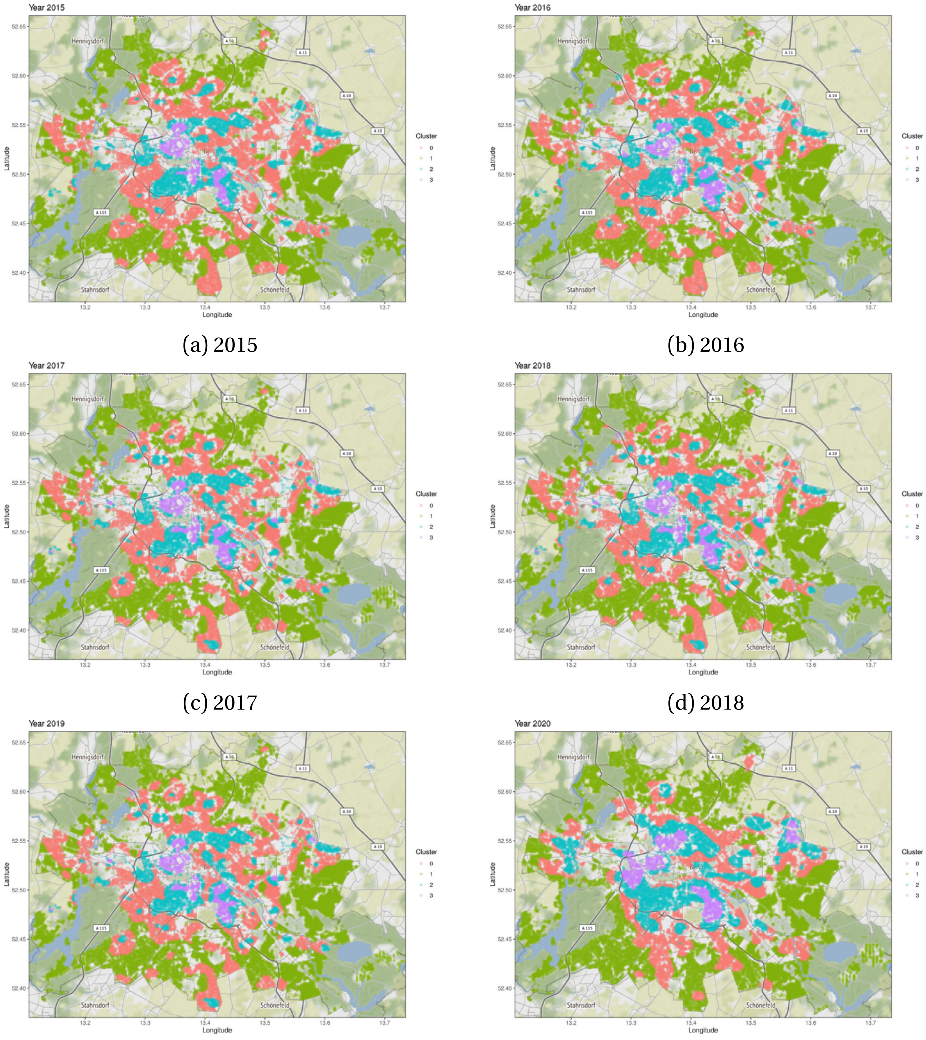
To provide a more comprehensive overview of the results, we will compare our findings with those of other researchers who have independently addressed the issue of residential

Table 3 Summary of our results in comparison with selected previous studies

Author	Cohort	Method	Dimension explored	Main findings
Yamamoto [5]	{1973, 1975, 1990}	Plotting segregation indexes (location quotient)	Ethnic segregation of Turkish inhabitants in Berlin West (7 color-mapped areas)	In 1973, more than half of all Turks in West Berlin lived in Kreuzberg and Wedding. The research reports that in 1975 Turks were the most segregated compared to Germans, Italians, Greeks, Yugoslavs and other groups. By 1990, segregation between Turks and Germans had largely decreased.
Nakagawa [49]	{1965, 1970, 1975, 1980, 1985}	Hierarchical cluster analysis	Age-group segregation (the area was divided in a two concentric zone model of West Berlin)	The age groups 0-19 and 35 and over are more densely distributed in more densely in Outer Berlin than in Inner Berlin, and the age groups 20 to 34 tend to be more densely distributed in Inner Berlin.
Kemper [7]	{1991, 1995}	Plotting segregation indexes	• Ethnic segregation (2 zones, West and East, classified in a total of 6 colored areas) • Age-group segregation (Comparison of West and East Berlin, classified in a total of 7 colored areas)	The study notes that age segregation was more pronounced in East Berlin before unification, while socio-economic segregation was more pronounced in West Berlin. After unification, there was a decrease in the age group of children under 6 in the former East Berlin. Also, segregation rates of the foreign population decreased in both former West and East Berlin.
Kröhnert and Vollmer [15]	{1992, 1994, 1996, 1998, 2000, 2002, 2004}	Cluster analysis	• Gender segregation (5 clusters, at country level)	Berlin is part of a cluster of German geographical areas segregated by gender in which "the sex ratio is above average (…) and the share of students is the second highest among all clusters. The cities have strong service and tourism sectors. Unemployment among young people is low. The proportion of people employed in service sectors is among the highest of all clusters" [15, p. 9].
Blokland and Vief [27]	{2007, 2012, 2016}	Plotting segregation indexes (location quotient)	• Ethnic segregation (5 color-mapped areas) • Socio-economic (5 color-mapped areas)	Ethnic indicators: • Foreigners (strong decrease) • Persons with migration background (fair decrease) • Migration background: Turkey and Arabic states (strong decrease) • Migration background: European Union (stable) Socio-economic indicators: • Unemployed persons (stable) • Long-term unemployed persons (stable) • Non-unemployed persons receiving state subsidies (slight increase) • Child poverty (slight increase)
Marcińczak and Bernt [28]	{2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019}	Regression trees; Hierarchical cluster analysis	• Ethnic segregation (7 clusters)	This research found the following clusters: • Rising pluralist enclaves • Non-isolated host communities I • Stable pluralist areas • Non-isolated host communities II • Established and increasingly pluralist areas • Stable non-isolated host communities • Persistent host communities

Table 3 (*Continued*)

Author	Cohort	Method	Dimension explored	Main findings
Masías et al. [29]	{2020}	Multivariate Kernel Density Estimation; Non-Negative Matrix Factorization. Maps are provided for each dimension.	• Ethnic segregation (4 clusters) • Age-group segregation (3 clusters) • Socio-economic segregation (3 clusters)	Using a data science approach, it was possible to reveal highly interpretable patterns in the data, confirming the existence of the phenomena of ethnic segregation, age-group segregation and socio-economic segregation.
Present work	{2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020}	Multivariate Kernel Density Estimation; Dynamic Fuzzy C-Means; Maps and Bump charts are provided for each dimension	• Ethnic segregation (Changes from 7 to 8 clusters) • Age-group segregation (4 clusters) • Socio-economic segregation (3 clusters) • Gender Segregation (3 clusters)	*Macro dynamics* • The identification of a new cluster was determined. *Microdynamics* Migration background: • Cluster 0: Lebanon and Turkey become since 2010 the most overrepresented in this cluster • Cluster 1: The most overpreserented is USA • Cluster 2: subpopulation from Iran is the most overrepresented • Cluster 3: subpopulation from Poland is the most overrepresented • Cluster 4: Only Germans are overreprested and the rest groups become underrepresented • Cluster 5: Kazakhstan, Vietnam, former Soviet Republic, among others, are over-represented. • Cluster 6: Spain, France and Italy, among others, remain overrepresented over time. • Cluster 7: subpopulations with migratory backgrounds from Syria and China become the most overrepresent sup-population in the emergent cluster. Age-group segregation: • Cluster 0 only over-represents subpopulations the 65 and 90 subpopulations. • Cluster 1 over-represents young adults, adolescents and teenagers • Cluster 2 over-represents all groups, especially those aged 65-90 • Cluster 3 over-represents young adults and children. Gender segregation: • There is no residential segregation by gender. The clusters appear to mirror changes in population density across the city. Socio-economic segregation: • The clusters can identify areas where there is a higher density of people applying for unemployment benefits. Qualitatively, it can also be observed that there was a change in the distribution of these residential densities across the city in 2009 and 2020.



(a) 2015

(b) 2016

(c) 2017

(d) 2018

(e) 2019

(f) 2020

Figure 11 Continued

segregation. First, we would like to stress that the proposed methodology allows us to identify changes that we can label macro and micro. By macro changes, we refer to the possibility of clusters appearing, moving, or disappearing over time. By micro changes, we refer to the internal changes that can occur in the composition of each cluster, which we have operationalized and visualized using bump charts. Table 3 summarizes the main results of our approach, together with selected previous studies.

5.1.1 Macro dynamics

The results allowed us to establish that there is evidence of a structural change over the period analyzed. Within this structural change, a new cluster emerged in 2015, coinciding with the peak of the migration wave in the context of the European migration crisis. Given the nature of the dynamic clustering algorithm we use, which uses all past data to assess whether a change in cluster structure is taking place, identifying the emergence of a new cluster structure requires an event at the demographic level that makes it possible. The migration crisis in Europe and Germany's unprecedented refugee policy make the structural change we detect in Berlin a plausible interpretation of the data analysis results. The ability to detect the presence of residential segregation is the most salient finding of this study, as it demonstrates that the methodology can help to identify new patterns of residential segregation.

5.1.2 Micro dynamics

At the micro level, the bump charts show that some clusters have developed overrepresented subpopulations over time, others only underrepresented subpopulations, and a combination of both. The main trends identified can be summarized as follows:

- *Concerning ethnic residential segregation*: In terms of micro-dynamic changes, the proposed method allows us to study the changes within each cluster. The richness of the results allows us to observe the overrepresented subpopulations in each cluster and the changes in the classification of each cluster, allowing us to observe the dynamics over time. The results are consistent. They continue to show the results of the now long-past migration waves of "temporary" guest workers (i.e. the so-called *Gastarbeiter*) from Turkey and Lebanon. However, it is only in the present work that we can observe the positioning of the Syrian and Chinese migrant subpopulations as the most over-represented subpopulations as part of Cluster 7. The fact that both Syrian refugees and asylum seekers from China are known reality of recent immigration to Berlin. For example, Kate Martyr [48], an editor and video producer at DW's Asia desk, reports on the surge in asylum applications from China to Germany, particularly from the oppressed Uighur minority. Finally, we observe the increase or decrease of the spatial areas occupied by the clusters in Berlin as the normalized cluster size changes, which was noticeable in 2015 due to the structural change of the clusters.
- *Concerning age-group segregation*: In general, the bump charts show slight changes in the ranking of the categories of variables describing the phenomenon of age segregation. Age segregation is a demographic phenomenon characterized in detail by Yamamoto, Kemper, and Nakagawa, who used data available before and after the fall of the Berlin Wall. Nakagawa found two clusters in West Berlin, characterized by higher adult densities in outer Berlin compared to populations in inner Berlin. Kemper compared East and West Berlin before and after reunification and found different degrees of segregation in these two areas. Finally, Masías et al. [29] find four clusters with different age group distributions in the city.

  Through the application of dynamic analysis, our study confirms the existence of age group segregation phenomena, which is materialised in the four clusters we have found. The maps we present do not show idealized concentric zones, as suggested by earlier studies such as Nakagawa's, but more complex-shaped clusters that can be observed visually. We find that older people are concentrated in the peripheral areas of Berlin, spatially surrounding the other groups within the city, as seen in the maps provided. We also identified areas where young adults are found and clusters where children are over-represented. We also observe that the standardized size of the clusters does not change significantly over time, which can be interpreted to mean that the spatial areas these clusters occupy in space remain relatively stable. This is highly consistent with the observation of Nakagawa, who stated that "residential segregation by age group is a very real phenomenon" [49, p. 134]. In our results, we show with greater detail that the phenomenon of residential age segregation is present in Berlin.

- *On socio-economic residential segregation*: We observed that the ranking of the variables remained stable, i.e. in the same ranking position in all clusters during the years studied. Compared to previous research, some similarities can be observed in the locations with the highest rate of people claiming state subsidies (see the maps published by Blokland [27], Fig. 13.3 in p. 257). Finally, we would like to report that we have observed a qualitative change that can be seen in the 2020 map, where cluster areas take on new shapes. The results of the method show that, at least visually, there are socio-economically disadvantaged areas that only expanded in the years 2009 and 2020, which is reflected in the size of the clusters. We should bear in mind that 2009 was part of the subprime financial crisis and in 2020 the economy was under the stress of the COVID-19 outbreak. We believe that the change in cluster shapes may be related to the event of the global COVID-19 pandemic when many individuals in Berlin started to apply for social welfare. However, more research is needed to link this qualitative observation to a cause-effect relationship.
- *On residential segregation by gender*: We found changes in the variables describing population densities by gender. The data analysis shows 3 clusters representing different densities of male and female individuals. However, we observe cluster densities that reflect a slight imbalance between females and males. Finally, the normalized cluster size does not vary significantly over time, which means that the spatial areas of the clusters have neither shrunk nor expanded spatially throughout observation.

  Kröhnert and Vollmer [15] have argued that women from rural areas in Germany migrate to large cities such as Berlin more than men who remain in rural areas. Under this hypothesis, one might expect the possible emergence of clusters in which groups of internal migrants of women form clusters reflecting this phenomenon, which is still unknown to us. However, the variation in high, medium and low population density described by the clusters seems to reflect the variation in population density as a whole. Some changes are numerically small, but qualitatively significant for monitoring the expansion of gender residential segregation observed in other geographical regions (e.g. for examining population sex ratios in China and Saudi Arabia over time and space). Perhaps because the sex ratios in Germany are mostly balanced, the phenomenon can be observed when comparing rural areas with urban areas or between eastern and western Germany, i.e. when looking at data at the country level.

In this study, an analysis was conducted using a dynamic approach to describe the phenomenon of residential segregation in Berlin. As described in this paper, residential segregation is more of a complex dynamic phenomenon where different facets of Berlin's subpopulations are over- or under-represented in clusters across the city. We believe that the use of dynamic cluster analysis may be of particular interest to researchers who would like to find patterns that emerge from the data rather than trying to explain or predict a variable from a survey or a multivariate index, as in both cases, it can be understood as a supervized analysis problem, which by definition involves the creation of a variable or index that directly represents residential segregation. In our methodological and theoretical approach, patterns emerge from data based on multivariate and non-black box analysis methods.

Thus, at a high conceptual level, the analysis shows that there is no such thing as a subpopulation that isolates itself in residential areas. Instead, it can be represented as a multivariate phenomenon where clusters can be observed on the dimension of migration background, age groups, socio-economic groups, or the dimension of gender. These dimensions may have causal relationships with each other. However, in this study, we have taken a more focused approach to represent the phenomenon of residential segregation, which has been extensively documented in Berlin, rather than generating an explanatory model, as is commonly attempted.

5.2 Future research

Future research would aim to apply this approach to data from other cities in Germany and worldwide. In particular, it would be exciting to study demographic changes in migration crises or birth shortages and case diffusion processes in times of pandemics. It would also be interesting to include other variables that represent the neighbourhood, the quality of life of people, or the transport systems. In this way, the representation of residential segregation would also have associated elements of the city's infrastructure. Another idea is to analyze the clusters by looking at several dimensions together, for example, age and migration background, rather than studying them independently. But we explore these dimensions separately to illustrate the general thrust of our approach and also to contribute to protecting geo-privacy.

5.3 Practical applications

In this context, we believe that the approach followed in this study has multiple practical applications. Several tools describe the demography of Berlin, and some of them focus on measuring the integration of migratory subpopulations, such as the so-called "Integration Indicator Report", which is based on data provided by Der Mikrozensus, das Sozioökonomische Panel, and the Programme for International Student Assessment [50]. Also, the annual indicators published by the Federal Statistical Office (Destatis) [51], the German Expert Council on Integration and Migration (https://www.svr-migration.de/jahresgutachten/) and the annual reports of the *Organisation for Economic Co-operation and Development Integration Monitoring* [52] provide an overview of the migration situation of the diverse communities in the different countries.

The differential aspect of our approach is that it allows us to observe demographic changes in residential areas over time from a global perspective. In addition, different sets of variables can be analyzed separately or together. For example, as shown in Table 1, little is known about digital segregation. The possibility of analyzing the dynamics of clusters allows for a better understanding of the impacts of territorial policies and social interventions. We believe that the future availability of spatial databases describing the information and communication technologies used will make it possible to generate new representations of the relationship between the different aspects describing the phenomenon of residential segregation.

5.4 Limitations

The clustering algorithm makes it possible to detect structural changes, but it does not provide direct knowledge of the exact size of the clusters in each period. Instead, it only gives an idea of the size of the clusters in terms of proportions, as the algorithm aggregates data from previous periods in its updating process in each new re-evaluation, which can be seen as a limitation of the analytical approach. However, a more robust assessment of dynamic changes can be obtained with this strategy. The dynamic fuzzy clustering algorithm updates the clusters by incorporating the data previously evaluated in previous cluster updates. This means that it treats all previously analyzed data at each stage as entirely new. Using this strategy, the algorithm monitors structural changes rather than assessing year-on-year changes in the cluster. In this way, it detects changes in cluster composition when new data that differ from previously observed classes appear. Instead, the important thing is to assess how different the distribution of the data being aggregated is at each point in time. We can assess that this does not correspond to a local change but to a significant one because the algorithm detected this change considering all the previously available and evaluated data.

5.5 Conclusion

In this paper, we have proposed a methodology to explore and describe the demography of Berlin in residential areas. The proposed methodology allows us to make new observations on how different subpopulations are distributed in residential areas. In addition, as the analysis is carried out over time, new insights were gained into the changing internal composition of clusters, a rich diversity, and structural changes. We conclude that this novel approach, based on data science principles, allows us to identify patterns of residential segregation in Berlin in a more unified way. We encourage other researchers to develop new hypotheses about the demographic changes observed in residential areas and the factors that might explain them.

Appendix: Cluster size over time

Table 4 Cluster size over time-based on migration background variables

Cluster	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	12,743	23,883	34,978	46,030	56,991	67,600	84,643	95,928	106,916	118,175	128,286	138,482
1	5721	11,449	17,020	22,549	28,284	34,246	41,158	47,423	53,745	60,006	66,851	73,588
2	1345	2786	4218	5648	7088	8530	6942	7825	8654	9543	10,341	11,139
3	2653	5592	8712	11,851	15,057	18,218	17,475	20,200	22,977	25,617	28,489	31,491
4	3296	6947	10,721	14,642	18,486	22,268	24,064	27,852	31,527	35,363	39,444	43,558
5	3152	6974	10,822	14,749	18,878	23,156	20,102	23,356	26,721	30,034	33,110	36,171
6	2353	4750	7173	9603	12,031	14,509	14,914	17,145	19,372	21,608	23,718	25,792
7	–	–	–	–	–	–	11,442	13,235	15,083	16,862	18,909	20,985

Table 5 Cluster size over time-based on age-group variables

Cluster	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	8878	17,953	27,129	36,355	45,590	54,716	64,048	73,398	82,748	92,108	101,089	110,130
1	13,908	27,265	40,671	54,131	67,853	81,643	95,751	109,839	123,702	137,722	152,316	166,976
2	5257	10,702	16,113	21,568	27,042	32,535	38,043	43,616	49,216	54,806	60,027	65,268
3	3220	6461	9731	13,018	16,330	19,633	22,898	26,111	29,329	32,572	35,716	38,832

Table 6 Cluster size over time-based on gender variables

Cluster	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	3637	7307	10,959	14,638	18,341	22,010	25,693	29,374	33,017	36,556	39,536	43,144
1	19,647	38,745	58,295	77,700	97,361	117,216	137,320	157,678	177,568	198,016	218,377	238,998
2	7979	16,329	24,390	32,734	41,113	49,301	57,727	65,912	74,410	82,636	91,235	99,064

Table 7 Cluster size over time-based on socio-economic variables

Cluster	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
0	7158	14,389	22,112	33,393	44,725	56,454	68,216	79,462	90,243	101,038	111,819	119,882
1	17,259	34,457	51,589	66,840	82,386	97,864	113,824	130,342	146,901	163,631	180,108	197,498
2	5267	10,328	15,133	19,302	23,426	27,108	30,765	34,346	38,060	41,769	45,473	50,683
3	1579	3207	4810	5537	6278	7101	7935	8814	9791	10,770	11,748	13,143

Acknowledgements
We would like to thank the Department *Migration, Integration, Transnationalization* at the Social Science Research Center Berlin (WZB) for their academic support. Fernando Crespo acknowledges the support of Fondecyt Chile 1221562. We acknowledge the four anonymous reviewers. Any remaining errors are our own.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Abbreviations
LOR, spatial planning area (i.e., Die Lebensweltlich orientierten Räume; MIC, Mean in cluster; SGB II, Sozialgesetzbuch Zweites Buch; SGB III, Sozialgesetzbuch Drittes Buch; v-test, value-tests.

Data availability
The data that support the findings of this study are available from the register of residents (Einwohnerregister) from 2009 to 2020 available at the Statistical Office of Brandenburg (see www.statistik–berlin–brandenburg.de). However, restrictions apply to the availability of these data, which were used under license for the current study and are not publicly available.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author contributions
The collaborative project included contributions from Víctor H. Masías H., who worked on the conceptualization, methodology, formal analysis, investigation, interpretation of results, writing - original draft, visualization and project management. Julia Stier contributed to the conceptualization, investigation, writing - original draft and project management. Pilar Navarro R. focused on the methodology and writing - original draft. Mauricio A. Valle, Augusto Vargas and Sigifredo Laengle worked on researching, writing - original draft and writing - proofreading and editing. Lastly, Fernando Crespo worked on the writing - original draft, methodology and formal analysis. The authors read and approved the final manuscript.

Author details
[1]WZB Berlin Social Science Center, Reichpietschufer 50, 10785, Berlin, Germany. [2]E.T.S. of Computer and Telecommunication Engineering, University of Granada, Granada, Spain. [3]Facultad de Economía y Negocios, Universidad Finis Terrae, Santiago, Chile. [4]Faculty of Economics and Business, University of Chile, Diagonal Paraguay 257, 8320000, Santiago, Chile. [5]Departamento de Diseño y Manufactura, Universidad Técnica Federico Santa María, Viña del Mar, Chile. [6]Facultad de Economía y Negocios, Departamento de Gestión y Negocios, Universidad Alberto Hurtado, Santiago, Chile.

Received: 12 July 2023 Accepted: 19 February 2024 Published online: 12 March 2024

References

1. Nebe JM (1988) Residential segregation of ethnic groups in West German cities. Cities 5(3):235–244
2. Häußermann H (2013) Berlin: von der geteilten zur gespaltenen Stadt?: sozialräumlicher Wandel seit 1990. Springer, Berlin
3. Arin C (1991) The housing market and housing policies for the migrant labor population in West Berlin. In: Huttman E (ed) Urban housing segregation of minorities in western Europe and the United States. Duke University Press, Durham, pp 199–214
4. Nakagawa S (1993) Applying cohort analysis to residential segregation by age group in Berlin (West). Geogr Pol 61:133–142
5. Yamamoto K (1993) Spatial segregation of ethnic minorities in German cities. Geogr Rev Jpn, Ser B 66(2):127–155

6. Kemper FJ (1998) Restructuring of housing and ethnic segregation: recent developments in Berlin. Urban Stud 35(10):1765–1789
7. Kemper FJ (1998) Residential segregation and housing in Berlin: changes since unification. GeoJournal 46(1):17–28
8. Nakagawa S (1999) Internal migration in the territory of the former German democratic republic before German unification. Regional Views 12:15–26
9. Friedrichs J (2000) Ethnische Segregation im Kontext allgemeiner Segregationsprozesse in der Stadt. In: Harth A, Scheller G, Tessin W (eds) Stadt und soziale Ungleichheit. VS Verlag für Sozialwissenschaften, pp 174–196
10. Berlin HH (2007) From divided into fragmented city. HAGAR: Studies in Culture, Polity & Identities 7(1)
11. Schönwälder K, Söhn J (2009) Immigrant settlement structures in Germany: general patterns and urban levels of concentration of major groups. Urban Stud 46(7):1439–1460
12. Häußermann H, Kronauer M, Gornig M (2008) Desintegration und soziale Kohäsion in Berlin. Düsseldorf: Hans-Böckler-Stiftung
13. Friedrichs J, Gespaltene TS (2009) Städte?: soziale und ethnische Segregation in deutschen Großstädten. Springer, Wiesbaden
14. Geraedts J (2009) Döner versus Curry Wurst-Segregation versus integration: Comparing two neighbourhoods in Multi Cultural Berlin [Master Thesis]. Available from https://theses.ubn.ru.nl/bitstream/handle/123456789/3081/Geraedts%2c_Joske_1.pdf?sequence=1
15. Kröhnert S, Vollmer S (2012) Gender-specific migration from eastern to western Germany: where have all the young women gone? Int Migr 50(5):95–112
16. Zimmermann KF, Constant A, Schüller S (2014) Ethnic Spatial Dispersion and Immigrant Identity. In: Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Migration II. ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg; 2014. p. 0–25. No. E05-V1
17. Kapphan A (2013) Das arme Berlin: Sozialräumliche Polarisierung, Armutskonzentration und Ausgrenzung in den 1990er Jahren, vol 18. Springer, Wiesbaden
18. Mayer M (2013) New lines of division in the new Berlin. In: The Berlin reader. Transcript verlag, pp 95–106
19. Glitz A (2014) Ethnic segregation in Germany. Labour Econ 29:28–40
20. Jaczewska B, Grzegorczyk A (2016) Residential segregation of metropolitan areas of Warsaw, Berlin and Paris. Geogr Pol 89(2):141–168
21. Groß M, Rendtel U, Schmid T, Schmon S, Tzavidis N (2017) Estimating the density of ethnic minorities and aged people in Berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. J R Stat Soc, Ser A, Stat Soc 180(1):161–183
22. Working OECD (2018) Together for Local Integration of Migrants and Refugees in Berlin. OECD Publishing, Available from https://books.google.de/books?id=BiVtDwAAQBAJ
23. 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
24. Kurtenbach S (2019) Digitale Segregation. Sozialräumliche Muster der Nutzung digitaler Nachbarschaftsplattformen. In: Heinze R, Kurtenbach S, Üblacker J (eds) Digitalisierung und Nachbarschaft. Erosion des Zusammenlebens oder neue Vergemeinschaftung, Baden-Baden: Nomos, pp 115–142
25. Heider B, Stroms P, Koch J, Siedentop S (2020) Where do immigrants move in Germany? The role of international migration in regional disparities in population development. Popul Space Place 26(8):1–19
26. Bartzokas-Tsiompras A, Photis Y (2020) Does neighborhood walkability affect ethnic diversity in Berlin? Insights from a spatial modeling approach. Eur. J. Geogr. 11(1):163–187
27. Blokland T, Vief R (2021) Making Sense of Segregation in a Well-Connected City: The Case of Berlin. Urban Socio-Economic Segregation and Income Inequality, 249
28. Marcińczak S, Bernt M (2021) Immigration. segregation and neighborhood change in Berlin Cities, 103417
29. Masías V, Stier J, Navarro P, Valle MA et al (2023) A novel methodological approach for analyzing the multifaceted phenomenon of residential segregation: the case of Berlin. Cities 141:104465
30. Arandelovic B, Bogunovich D (2014) City profile: Berlin. Cities 37:1–26
31. Gosnell HF, Schmidt MJ (1936) Factorial and correlational analysis of the 1934 vote in Chicago. J Am Stat Assoc 31(195):507–518
32. Price DO (1941) Factor analysis in the study of metropolitan centers. Soc Forces 20:449
33. Sweetser FL (1965) Factorial ecology: Helsinki, 1960. Demography 2(1):372–385
34. Benassi F, Bonifazi C, Heins F, Lipizzi F, Strozza S (2020) Comparing residential segregation of migrant populations in selected European urban and metropolitan areas. Spat Demogr 8:269–290
35. Olteanu M, Hazan A, Cottrell M, Randon-Furling J (2020) Multidimensional urban segregation: toward a neural network measure. Neural Comput Appl 32:18179–18191
36. Dmowska A, Stepinski TF (2023) Spatio-temporal changes in racial segregation and diversity in large US cities from 1990 to 2020: a visual data analysis. EPJ Data Sci 12(1):30
37. Yamamoto K (1983) Dynamics of population and spatial segregation in Munich. Keizai Shirin (The Hosei University Economic Review) 50(3/4):1–59
38. Bundesamt S (2005) In: Bevölkerung und Erwerbstätigkeit: Bevölkerung mit Migrationshintergrund, Wiesbaden, Germany
39. Groß M, Kreutzmann AK, Rendtel U, Schmid T, Tzavidis N (2020) Switching Between Different Non-Hierachical Administrative Areas via Simulated Geo-Coordinates: A Case Study for Student Residents in Berlin. Journal of Official Statistics JOS(2). 36
40. Rendtel U, Die RM (2018) Konstruktion von Dienstleistungskarten mit Open Data am Beispiel des lokalen Bedarfs an Kinderbetreuung in Berlin. AStA Wirtsch Sozialstat Arch 12(3):271–284
41. Erfurth K, Groß M, Rendtel U, Schmid T (2022) Kernel density smoothing of composite spatial data on administrative area level. AStA Wirtsch Sozialstat Arch 16(1):25–49
42. Erfurth K, Groß M, Rendtel U, Schmid T (2022) Kernel density smoothing of composite spatial data on administrative area level: a case study of voting data in Berlin. AStA Wirtsch Sozialstat Arch 16(1):25–49

43. Geofabrik GmbH. OpenStreetMap Daten für Berlin; 2020. http://download.geofabrik.de/europe/germany/berlin-latest-free.shp.zip
44. Crespo F, Weber R (2005) A methodology for dynamic data mining based on fuzzy clustering. Fuzzy Sets Syst 150(2):267–284
45. Li RP, Mukaidono M (1995) A maximum-entropy approach to fuzzy clustering. In: Proceedings of IEEE international conference on fuzzy systems, vol 4. IEEE, pp 2227–2232
46. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
47. Kirk A (2016) Bump chart. SAGE Publications, Newbury Park
48. Martyr K (2020) China asylum claims to Germany more than double. DW News. Available from https://www.dw.com/en/china-asylum-claims-to-germany-more-than-double/a-52396720
49. Nakagawa S (1990) Changing segregation patterns by age group in the Tokyo metropolitan area-from the viewpoint of migration with cohort analysis. Geogr Rev Jpn, Ser B 63(1):34–47
50. Dietrich E, Köller R, Koopmans R, Höhne J (2011) Zweiter Integrationsindikatorenbericht. erstellt für die Beauftragte der Bundesregierung für Migration, Flüchtlinge und Integration (Stand Dezember 2011). Köln/ Berlin: Die Beauftragte der Bundesregierung für Migration, Flüchtlinge und Integration
51. Seuberlich M (2021) Statistisches Bundesamt/Statistische Landesämter. In: Andersen U, Bogumil J, Marschall S, Woyke W (eds) Handwörterbuch des politischen Systems der Bundesrepublik Deutschland. Springer, Wiesbaden, pp 880–884
52. OECD. International Migration Outlook 2021. OECD; 2021. Available from https://doi.org/10.1787/29f23e9d-en

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.