Customer Segmentation in the Power of Attorney Business by Means of Fuzzy Clustering

Peter Rausch
Nuremberg Institute of Technology Georg Simon Ohm
Keßlerplatz 12
90489 Nuremberg, Germany
peter.rausch@th-nuernberg.de

Michael Stumpf
Nuremberg Institute of Technology Georg Simon Ohm
Keßlerplatz 12
90489 Nuremberg, Germany
michael.stumpf@th-nuernberg.de

Keywords

Customer Segmentation, Fuzzy Clustering, Modeling, Machine Learning, Data Mining, Targeting

ABSTRACT

Over the past years, a large market of service providers for powers of attorney has emerged in Germany. The number of officially registered general powers of attorney raised from almost 326,000 in 2005 to more than 5.3 million in 2021, according to the German Federal Chamber of Notaries. Service offers include consulting services on powers of attorney, text blocks and templates, annual update services and emergency call services. In general, the market contains all segments of the population, irrespective of their demographic characteristics. In this difficult and competitive environment, marketing and sales play an important role. In particular, not much research in the field of customer segmentation and marketing campaign preparation on this industry can be found. Thus, this research had the following goals which could all be achieved: The first goal was to compare the actual regional demographic structures of the German population with the regional structures of the customer base of a service provider in order to identify underrepresented population groups. The data on underrepresented groups was used to identify customer segments which can be used for targeting when future marketing campaigns are prepared. To solve this issue popular deterministic and fuzzy cluster approaches were successfully transferred to the use case and suitable configurations were explored as a second goal. As a third goal a recommendation, which of the analyzed approaches is most suitable in the present case, was derived. The approaches were evaluated on a real-world example of a service provider with more than 80,000 customers. It is important to note that despite of the fact that the regarded industry partner addresses a special market, the promising findings can be transferred to many other areas.

1. INTRODUCTION

Any person can get into a situation of being unable to articulate the own will by an accident, an illness or even slowly by old age. In some cases decisions which need immediate action, for instance, concerning medical treatments, important personal economic matters or urgent business matters, can not be made or communicated anymore. The number of people recorded in Germany who mainly needed care for the reasons listed above has risen from about 625,000 since 1995 [12] to currently approximately 1.3 million people [15]. Thus, it is not surprising that general, durable and medical powers of attorney are attracting increasing attention due to personal experiences among family and friends, but also because of media reports. According to the Federal Chamber of Notaries in Germany, the number of officially registered general powers of attorney raised from 325,637 in 2005 to 5,366,795 in 2021 in Germany [5]. The market for related products includes all parts of the population, independent from their demographic characteristics such as age, gender or place of residence. In this competitive environment, marketing and sales play an important role, since the above-mentioned products are not self-explanatory and, similarly to insurance products, the need has to be aroused. This is complicated by the fact that different demographic groups have to be targeted individually via campaigns, in order to achieve the highest possible impact. For instance, social media campaigns for younger and older people should be different. In order to support customer segmentation as a basis for preparing industry specific campaigns, very few well-founded documented findings can be found in science and reports from the field.

By means of this research this gap should become smaller and multiple goals have to be achieved. The first aim is to compare the actual regional demographic structures of the German population with the regional structures of the customer base in order to identify underrepresented population groups of the regarded service provider's customers. Knowledge about underrepresented groups is important for targeting when future marketing campaigns, for instance, via social media channels, are prepared. In case of the regarded service provider with many thousands of customers it is impossible to generate this knowledge without automated approaches. In particular, we suggest a cluster approach to further process the results of the comparison of the regional actual and customer structures. It is intended to find underrepresented customer segments in the customer base. The demographic characteristics of the related clusters play an important role in social media campaigns. The results of an appropriate cluster approach are important, because different customer segments should be differently addressed by means of marketing campaigns on social media channels. In our case, it has to be taken into account that we have to deal with mixed categorical and continuous data. We will see that cases exist in which the customers do not have a homogeneous structure. This means, that individual instances do not really fit to a certain cluster but also match to another cluster based on their attributes by a slightly lesser degree. To plan effective marketing campaigns in our case the issue of identifying and handling a huge number of intermediate objects has to be solved. For the purpose, we will transfer a Fuzzy C-Means clustering algorithm to the use case and search for suitable configurations as a second goal. Finally, as a third goal a recommendation which approach is most appropriate in the presented case will be given. The approaches will be evaluated on a real-world example of a service provider with more than 80,000 customers.

At first, in Section 2 we will introduce the related use case and provide some background information to get a better understanding of the issue we intend to solve. Afterwards, in Section 3, we provide an overview of the current state of the art of approaches in this field. Based on the results of the industry partner's data, which is described in Section 4, different approaches and configurations are analyzed in Section 5. For this purpose, we start with a brief overview of deterministic clustering and the transfer to our use case. Then, fuzzy approaches are tested in different configurations to get improved results. Subsequently, the benefits and open issues of the approaches are discussed and compared in Section 6. Finally, in Section 7 the results are summarized and further enhancements will be discussed. It is important to note that despite the regarded industry partner addresses a special market, the analyzed problems can be found in many other areas and the proposed solutions can be transferred.

2. CASE DESCRIPTION

Before we go into details some background information on the industry, the industry partner and the related use case is needed to understand the issue and to evaluate the proposed solution later on in Section 6. As already mentioned, the number of officially registered general powers of attorney strongly increased. A main reason for this is the demographic change in Germany which will intensify according to forecasts ([25]). In general, everybody can quickly get into a situation in which a power of attorney is needed even at a young age. So this issue basically affects everyone. In order to set up valid powers of attorney correctly, a number of legal requirements must be regarded (see for instance, §1829 especially sub-paragraph 5 German BGB ([4]). Furthermore, experts advise updating powers of attorney at least every three to five years ([27]). Accordingly, a market has emerged with a variety of service providers. Services offered include consulting services on powers of attorney, text blocks and templates, annual update services, and emergency call services. Providers range from notaries, publishers, insurers or financial services brokers, automobile clubs as intermediaries, to platforms supporting the easy creation, distribution and sale of legal documents. The products and services of the regarded company include services to create general, durable as well as medical powers of attorney and a hotline service for emergency cases. Since these services are useful for almost all parts of the population, the structure of the customers in terms of their demographic features should correspond to the structure of the population. By comparing the current customer structure with the demographic structure, mismatches can be used to identify potential starting points for future marketing campaigns. For instance, if younger people are underrepresented in the customer structure a targeted campaign addressing this customer group can be launched. If we consider clusters of customers, the number of clusters is unknown, in advance. Furthermore, it is likely that many customers do not fit perfectly to a certain cluster. Since, we may have to deal with many intermediate customer instances a dichotomic assignment to clusters would distort the results and would not be a good representation of the reality. Based on such a segmentation positive impacts of campaigns could not be fully exploited, or in borderline cases a decision maker might choose the wrong option when faced with several alternative courses of action. For this reason we will consider fuzzy approaches in this research. Before more information on the data set is given, we will provide a detailed review on the subject and approaches applied in the field of customer segmentation.

3. RELATED WORK

In literature, the term market segmentation describes the idea of grouping similar consumers in terms of specific characteristics [23]. Based on this segmentation, attractive segments can be identified and targeted, and marketing actions for each segment can be customized. According to [7], different categories of segmentation, for instance, geographic, demographic, psychological, psychographic, sociocultural, user-related, user-situation and benefit segmentation, are distinguished. Also mixtures of segmentation approaches, so-called hybrid segmentation methods, are possible [7]. In our paper, based on the regarded input data, see Section 4, demographic and geographic segmentation is addressed. Thus, we will focus on a hybrid segmentation approach. Since the data set is unlabeled, see Section 4, unsupervised learning approaches have to be considered. In particular, many successful applications of cluster approaches can be found. In some cases, segmentation is used to address the more or less volatile behavior of customers, see for instance, [6]. The idea is to develop relationships with customers based on accurate information which is optimized for the customers' taste and preferences [6]. Other research focuses on customer segmentation for an online seller of automobile accessories and fittings based on a k-means approach. The idea is to recommend methods to increase customer influx and give suggestions for better performance and sales [21]. A disadvantage of this approach is the a priori selection of the number of clusters which has an impact on the quality of the algorithm's outcome [13]. Additionally, this approach tends to have issues in cases when clusters are of different sizes, densities or when their shapes are not globular [8]. Also, outliers can cause issues because the approach is based on the arithmetic mean of points in an n-dimensional space [13]. Besides, in our case categorical data has to be processed which is not supported by the k-means approach [8]. In [11] the case of demographic customer segmentation of banking users is discussed, but as well as [21] they do not address the aspect of fuzzy memberships to clusters.

[28] applies fuzzy clustering to customers' segmentation in securities industry. The idea is to identify the customers with similar characteristics and value by means of a Fuzzy C-Means (FCM) approach which will be explained in Section 5.3. Many other successful applications of customer segmentation by means of fuzzy clustering can be found. For instance, fuzzy clustering is applied in the fields of energy consumers [22], telecom [1], life insurance [14] and B2B for food and beverages merchants [16]. [19] analyze the case of implementing marketing strategies by observing dynamic changes in the customer segments over time. Like in the other cases, they use a FCM approach. These many successful practical applications indicate that fuzzy clustering may also lead to promising results in the present case. Nevertheless, it must be analyzed whether the FCM approaches' results are of such a good quality that a recommendation can be made in terms of goal 3, see Section 1. Besides the fact that we consider a different line of business, the FCM algorithm which is used in most of the mentioned use cases has some limitations. The approach has issues with outlier data and is sensitive to initialization. Hence, a general optima can not be guaranteed [19]. This problem needs to be addressed. Before we go into details, the data set and the preprocessing step will be explained.

4. DATA SET AND PREPROCESSING

As already mentioned the basic idea is to cluster customers in terms of their demographic and geographic features and use the results for the preparation of marketing campaigns. The data set provided by the industry partner is unlabeled and contains mixed data types. The feature "year of birth" was converted to a 0-1-normalized attribute. For reporting purposes is can be reconverted or transformed to the customers' age, depending on the purpose of the analysis. In addition to this attribute, the features "gender" and "place of residence" were included. These features are used, because social media campaigns can be targeted by these attributes and this data could be provided by the industry partner. The feature "gender" is categorical data, but we used it as binary variable because the data from the officially registered German population based on the last census was just available for two genders, see [24]. When the data from the 2022 census currently conducted in Germany will be published, certainly data on all genders will be available. For our case, we just considered the genders "male" and "female". However, it is recommended to add other genders as soon as the new related official statistics is available.

The "place of residence" feature is present in multiple different granularity levels. For our study, we had to restrict our comparisons of the customer base to the demographic structure of the population on the level federal states. Although the customer data was available in the form of postal codes, data for population statistics was available only in a separate geographical region structure which can not be matched exactly to postal codes. In general, a more fine-grained analysis would be possible, if more detailed data from the government would be accessible. Because of this, the lowest common denominator where regions between the industry data set and the population statistics match is the federal state level. So, in the industry partner's input data set as well as in the data set from the official German population statistics preprocessing was executed to unify "place of residence" to the federal state level. The information about the postal code for each instance of the industry partner's data set is still available. This can be used to display instances on a geographic map with more detail than the federal state level. Additionally, the feature "marital status" (yes/no) was considered as a binary value for the first set of experiments.

The above mentioned features can be used to target customers, for instance, in the preparation process of social media campaigns on popular platforms, see for instance [18]. To choose an appropriate approach for the segmentation of the customers a pre-analysis was executed. For this purpose, the data was analyzed, and accumulations of characteristics could be identified. For instance, in Germany, about 41.8% of the total population are married [26]. The data set of the regarded company with about 80,000 instances included about 69.4% with this marital status. It is already apparent at this point that there are major discrepancies between the customer structure and the population structure. Of course, the results of this pre-analysis will be refined, elaborated and enhanced in the next section.

5. APPROACH

5.1 OVERVIEW

As already outlined in Section 1 the basic idea of this research is to identify underrepresented population groups in the customer base in order to use this information for targeting when future marketing campaigns are prepared. To compare the actual demographic structure of the German population with the structure of the customer base the data which was described in Section 4 will be used. Due to a mix of demographic and geographic characteristics, a hybrid segmentation approach, see Section 3, is chosen.

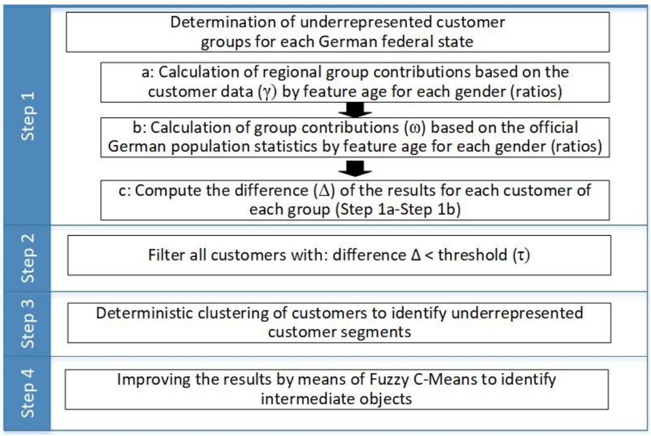

Step 1
Determination of underrepresented customer groups for each German federal state
a: Calculation of regional group contributions based on the customer data (γ) by feature age for each gender (ratios)
b: Calculation of group contributions (ω) based on the official German population statistics by feature age for each gender (ratios)
c: Compute the difference (Δ) of the results for each customer of each group (Step 1a-Step 1b)
Step 2
Filter all customers with: difference Δ < threshold (τ)
Step 3
Deterministic clustering of customers to identify underrepresented customer segments
Step 4
Improving the results by means of Fuzzy C-Means to identify intermediate objects

Figure 1: Approach: overview

As shown in Figure 1, we determine underrepresented customer groups for each German federal state, in Steps 1a-c. In Step 1a, the group contributions γ_ϕ by the features age a and gender g based on the customer data for each federal state ϕ are calculated as the intersection of the related feature set divided by the total number of customers Π in federal state ϕ:

$$\gamma_{(a \cap g, \phi)} = \frac{\mid \pi_{a,\phi} \cap \pi_{g,\phi} \mid}{\Pi} \quad \forall \phi = 1, \ldots, \Phi \tag{1}$$

$\pi_{a,\phi}$ and $\pi_{g,\phi}$ denote the set of customers who have a certain characteristic profile in the customer base.

Analogously, in Step 1b the group contributions of the officially registered population ω by features age a and gender g based on the official German population statistics are determined for each federal state ϕ divided by the total number of citizens ϱ in federal state ϕ:

$$\omega_{(a \cap g, \phi)} = \frac{\mid \rho_{a,\phi} \cap \rho_{g,\phi} \mid}{\varrho} \quad \forall \phi = 1, \ldots, \Phi \tag{2}$$

$\rho_{a,\phi}$ and $\rho_{g,\phi}$ denote the set of officially registered citizens who have a certain characteristic profile in the official German population statistics. These ratios will be used as references and will be compared to the results of Step 1a in the following Step 1c

This means that each individual group of Step 1a is compared to its corresponding group of Step 1b, like shown in Equation 3:

$$\Delta_{(a \cap g, \phi)} = \gamma_{(a \cap g, \phi)} - \omega_{(a \cap g, \phi)} \quad \forall \phi = 1, \ldots, \Phi \tag{3}$$

By computing the differences $\Delta_{(a \cap g, \phi)}$, the contributions of each age and gender group in the dataset is compared to its corresponding average group contribution in the German population for each state. So, instances with positive $\Delta_{(a \cap g, \phi)}$ denote that this instance is part of an overrepresented group in the customer data set. Whereas negative values of $\Delta_{(a \cap g, \phi)}$ indicate that the instance is part of an underrepresented group.

If a difference is smaller than a user defined threshold, all objects of the related group are filtered in Step 2. Afterwards, in Step 3 all filtered objects are clustered to identify the underrepresented customer segments. Finally, in Step 4, the results are improved by means of a Fuzzy C-Means approach.

5.2 DETERMINISTIC CLUSTERING

5.2.1 APPROACH AND CONFIGURATION

According to Figure 1 (Steps 3 and 4), the filtered objects which include all underrepresented customers are clustered. Since, cluster analysis has proven its efficiency in the area of customer segmentation, see Section 3, we apply this method to our case. Deterministic cluster approaches are based on the assumption that an object o (of all O filtered object instances I) is always member of only one cluster c. Thus, the following condition has to be satisfied:

$$\mu_{oc} \in \{0, 1\}; \forall o = 1, \ldots, O \tag{4}$$

In our case, we choose a hierarchical approach for Step 3, see Figure 1. This category of clustering allows to visualize the results as dendrograms. By means of a dendrogram, the analyst can determine a threshold for the proximity measure and fix it in a way that the clusters have a desired homogeneity. Thus, in advance, there is no knowledge about the number of clusters necessary. Within the hierarchical approaches, we have selected an agglomerative complete linkage clustering. It tends to form small compact groups and avoids the chaining effect which might be an issue related to other hierarchical approaches [9]. In detail, the steps shown in Figure 2 are executed.

Like other agglomerative approaches, each object initially forms a cluster. Subsequently, this initial partition is modified by successively merging the clusters into larger aggregates. For this reason, the distances or the similarities between the clusters have to be calculated. In our case, we use the Gower distance. In general, it is used to compute the distances between objects having numerical and categorical features [10] as in our case. As Equation 5 shows, the Gower distance allows to weight the impact of each of the N features individually [10] - which is another requirement in our case.

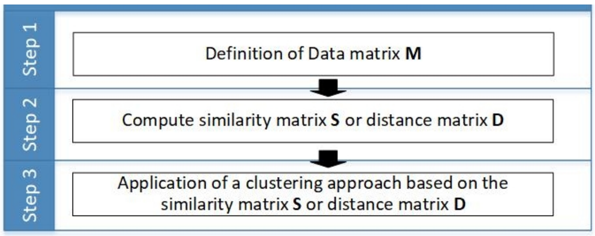

Step 1
Definition of Data matrix **M**
Step 2
Compute similarity matrix **S** or distance matrix **D**
Step 3
Application of a clustering approach based on the similarity matrix **S** or distance matrix **D**

Figure 2: Hierarchical-agglomerative clustering methods

$$d(o, c) = \frac{\sum_{k=1}^{N} \delta_{oc}^{(k)} d_{oc}^{(k)}}{\sum_{k=1}^{N} \delta_{oc}^{(k)}} \tag{5}$$

For binary features, the distances are computed as shown in Equation 6, and metric features are included according to Equation 7.

$$d_{oc}^{(k)} = \begin{cases} 1 & \text{if} \quad x_{ok} \neq x_{ck} \\ 0 & \text{if} \quad x_{ok} = x_{ck} \end{cases} \tag{6}$$

$$d_{oc}^{(k)} = \frac{|x_{ok} - x_{ck}|}{R(k)} \quad \text{with} \quad R(k) = \max_{o} x_{ok} - \min_{o} x_{ok} \tag{7}$$

In our case, we use one binary variable for each federal region of Germany. Thus, for Germany we need 16 binary variables to indicate if a customer belongs to a federal region or not. By means of the weight $\delta_{oc}^{(k)}$ the influence of each geographic binary variable is reduced to $\frac{1}{16}$ to avoid an inappropriate large influence of the “place of residence” feature on the distance. Due to the reasons mentioned in Section 4 the “gender” feature was also regarded as a binary variable with a weight of one. With the same weight the age was considered as a numerical value. Furthermore, the feature “marital status” was initially included in the Gower distance. However, as the analyses showed, this feature had no significant influence on the cluster structures and is therefore no longer considered in the further explanations. To compute the distances or the similarities between the clusters, the maximum distance or the minimum similarity of a cluster pairs’ elements are determined, see Equation 8.

$$d_{lm} = \underset{(I_o, I_c) \in P_l \times P_m}{\text{Max}} \{d_{oc}\} \quad \text{or} \quad s_{lm} = \underset{(I_o, I_c) \in P_l \times P_m}{\text{Min}} \{s_{oc}\} \tag{8}$$

When the distances, like in our case, or the similarities between all clusters have been determined, the clusters which have the minimum distance or the maximum similarity to each other are merged.

Afterwards, the proximities between the new agglomerate P_q resulting from the fusion of the clusters P_l and P_m and the remaining classes have to be calculated. This procedure is repeated until a certain number of clusters is achieved or all objects are in one cluster. In the latter case, the cluster structure can be visualized by means of a dendrogram. The dendrogram helps to define a threshold for the maximum tolerated distance, and the cluster can be split. Finally, a desired number of more or less compact clusters is resulted. A disadvantage of this approach are the high computational costs for analyses [17], which our experiments confirm. Even with the filtered subset of customers and a powerful computer, $\frac{O(O-1)}{2}$ comparisons must be made to perform the algorithm. In our case, it was possible to test several configurations of the algorithm within a day. But with the complete number of customers, the performance of the available hardware would have been exceeded with many million comparisons. Another problem with this cluster procedure is that the selection of the clusters to be merged is determined only by two extremely positioned objects of the cluster pairs. In addition, this method does not detect outliers. The extent to which these problems are relevant in our case is explained below in combination with the presentation of the results of the deterministic part of the experiments.

5.2.2 DETERMINISTIC RESULTS

As Figure 3 shows, 10 clusters ($C_{\mathrm{HI}}1$,...,$C_{\mathrm{HI}}10$), which instances are denoted by their numbers, are the result of the deterministic cluster approach. Due to the hierarchical approach the clusters varied widely in size. Besides 3 large clusters also 3 medium and 4 small clusters were identified.

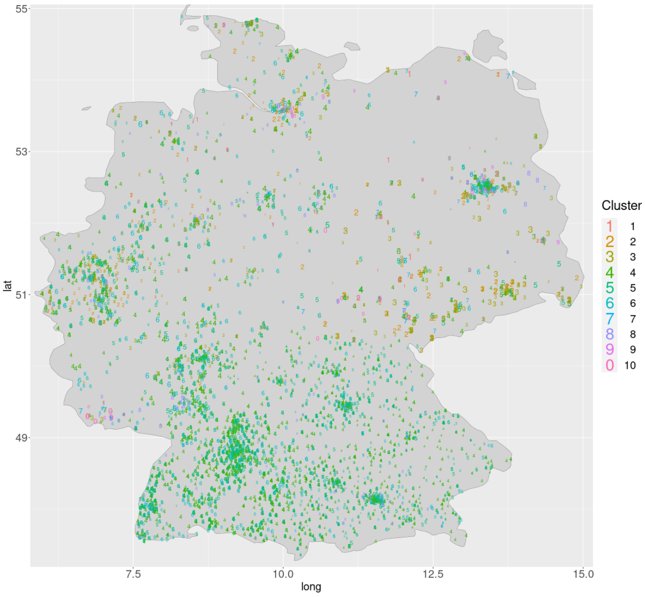
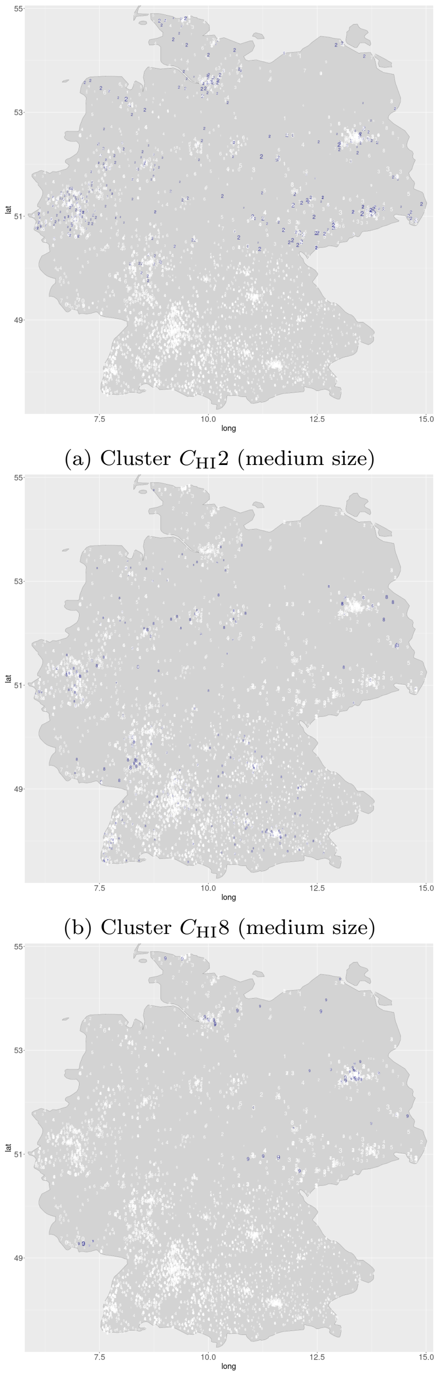


Figure 3: Visualization of the deterministic cluster results (The larger the symbol size the higher the underrepresentation.)

The large clusters ($C_{\mathrm{HI}}4$, $C_{\mathrm{HI}}5$, and $C_{\mathrm{HI}}6$) are not geographically grouped but are relatively compact in the age dimension. If the range between first and third quartile around the median is considered the age dimension of cluster $C_{\mathrm{HI}}6$ has a range of ±2 and cluster $C_{\mathrm{HI}}4$ and $C_{\mathrm{HI}}5$ are almost as compact with a range of ±3, see Table 1.

For smaller clusters this is not always the case, but for instance, cluster $C_{\mathrm{HI}}9$, which represented roughly 1% of the filtered data set, had a range of ±3 around the birth year of 1976 and its instances were all male and geographically located in different compact areas in the northern half of Germany, see Fig. 4c. Yet, there are also medium clusters with a larger scatter overlapping others, like, for instance, clusters $C_{\mathrm{HI}}2$ and $C_{\mathrm{HI}}8$, see Fig. 4a and 4b.



(a) Cluster $C_{\mathrm{HI}}2$ (medium size)

(b) Cluster $C_{\mathrm{HI}}8$ (medium size)

(c) Cluster $C_{\mathrm{HI}}9$ (small size)

Figure 4: Visualization of selected individual clusters in hierarchical results (blue symbols represent the corresponding cluster membership)

Both of these clusters consist again of males, around a certain age ($C_{\mathrm{HI}}2$: 1934 ± 4, $C_{\mathrm{HI}}8$: 1940 ± 3) but overlap geographically.

			Birth Year (19xx)				
Cluster-No.	**Percentage**	**Gender**	**Min**	**1.Q**	**Median**	**3.Q**	**Max**
$C_{\mathrm{HI}}2$	6.0 %	male	35	40	43	45	51
$C_{\mathrm{HI}}4$	38.0 %	male	80	87	90	93	96
$C_{\mathrm{HI}}5$	15.3 %	female	21	29	32	35	40
$C_{\mathrm{HI}}6$	27.8 %	female	73	90	92	94	96
$C_{\mathrm{HI}}8$	5.1 %	male	27	31	34	37	43
$C_{\mathrm{HI}}9$	0.9 %	male	64	73	76	79	85

Table 1: Selected statistical data for hierarchical cluster results

However, they are not only intersecting by location, but if minimum and maximum of the birth year is considered they also overlap in the age dimension. The minimum and maximum values of year of birth are for cluster $C_{\mathrm{HI}}2$ in the range 1935-1951 and for cluster $C_{\mathrm{HI}}8$ in the range 1927-1943, see Table 1. So, some instances of cluster $C_{\mathrm{HI}}8$ are well within the first quartile (1940) and the median (1943) year of birth of cluster $C_{\mathrm{HI}}2$. It is to be doubted that the instances possessing features of different clusters should exclusively have the degree of belonging 0 or 1 to exactly one cluster. In addition, a non-dichotomous mapping would be helpful to better estimate the potential of marketing campaigns because there might be cases where additional customers can be reached by a single campaign. Therefore, in the following will be investigated to what extent the results can be improved by fuzzy approaches, see Figure 1, Step 4.

5.3 FUZZY PART

In contrast to deterministic cluster approaches, condition 4 is relaxed in fuzzy cluster analyses, so that an object can be member of multiple clusters:

$$\mu_{oc} \in [0,1] \tag{9}$$

This relaxed condition ensures that the above mentioned problems will be solved. In order to be able to evaluate the quality of the results of the fuzzy algorithm, the approach should be examined more closely beforehand.

5.3.1 APPROACH AND CONFIGURATION

As mentioned in Section 3, the Fuzzy C-Means (FCM) approach, which is one of the most known and most commonly used variant of fuzzy cluster analysis algorithms [20], is used in many cases in the field of customer segmentation. It is based on the classical ISODATA algorithm and minimizes the generalized variance criterion σ_m [2]:

$$\begin{aligned} \sigma_m &= \sum_{c=1}^{C}\sum_{o=1}^{O} \mu_{oc}^{m} \cdot |\mathbf{f}_o - \overline{\mathbf{u}}_c|^2 \\ \text{with} \quad \overline{\mathbf{u}}_c &= \frac{\sum_{o=1}^{O} \mu_{oc}^{m}\mathbf{f}_o}{\sum_{o=1}^{O} \mu_{oc}^{m}}, \quad \forall\, c = 1,\dots,C \end{aligned} \tag{10}$$

	1.Q - Median		3.Q - Median		Max - Min	
Experiment	**Median**	**Min**	**Median**	**Max**	**Median**	**Max**
Hierarchical	-3.0	-3.0	+3.0	+4.0	17.5	23.0
FCM	-2.5	-4.0	+2.0	+6.0	12.5	27.0
FCM with Start U	-2.0	-4.0	+2.0	+6.0	12.5	27.0

Table 2: Selected statistical properties of age dimension over all clusters between experiments

$\mathbf{f}_o$ represents the feature vector of object o and $\mathbf{u}_c$ the centroid of a cluster c. The power $m \in [1, \infty[$ is to be determined by the user. The use of the objective function 10 identifies relatively homogeneous partitions. Besides, the elements of the membership matrix U must satisfy the following conditions:

$$\mu_{oc} \geq 0 \quad \forall\, c = 1,\dots,C; \quad \forall\, o = 1,\dots,O \tag{11}$$

$$\sum_{c=1}^{C} \mu_{oz} = 1 \quad \forall\, o = 1,\dots,O \tag{12}$$

It should be noted that the fulfillment of condition 12 can distort the results in case of outliers which would be best represented by very small membership values to all clusters. Considering this aspect, it has to be analyzed whether this approach can improve the results of the deterministic method in Section 5.2.2. Since we need a start partition to solve the problem described above, it appeared reasonable to use the results of Section 5.2.2. Alternative experiments with random start partitions and different seeds confirmed that this approach is superior and achieves the best results for our use case. Details will be given later in Section 5.3.2. The use of the deterministic start partition obviously seems to help solving the problem that FCM is sensitive to the start partition. To determine the FCM results, first, the class centroids of each cluster $\overline{\mathbf{u}}_c$ must be determined for all clusters $c = 1,\dots,C$. In the next step, the new membership values of the objects are calculated by Equation 13:

$$\begin{gathered} \mu_{oc}^{new} = \left(\sum_{h=1}^{C} \left(\frac{|\mathbf{m}_o - \overline{\mathbf{u}}_c|}{|\mathbf{m}_o - \overline{\mathbf{u}}_h|} \right)^{\frac{2}{m-1}} \right)^{-1} \\ \forall c = 1,\dots,C; \quad \forall o = 1,\dots,O \end{gathered} \tag{13}$$

According to (13), the membership value of an object o to the cluster c is determined by the distance between the object and the cluster centroids. In case the features of object O_o are identical to those of a centroid, the membership value is set to 1. For all other μ_{ih}^{new}–values with $h \neq c$, zero is set [3]. The approach ends after the q–th iteration if

$$\max_{((O_o, C_h) \in E \times C)} |\mu_{oh}^{(q)} - \mu_{oh}^{(q-1)}| \leq \xi, \tag{14}$$

where ξ is specified by the user. If $m \to 1$, the result tends to the result of a deterministic ISODATA–method. For $m \to \infty$, the membership values tend to the reciprocal of the number of classes $\frac{1}{C}$.

5.3.2 FUZZY RESULTS

At first, an FCM configuration was used to determine clusters without any previous knowledge about the cluster structure. The fuzzy clustering yields, as configured, 10 clusters ($C_{\mathrm{F0}}1$,...,$C_{\mathrm{F0}}10$). The instances of the filtered data set are grouped into 4 large, 5 medium-sized and 1 small cluster. As it can be seen in Table 3 the amount of small clusters is reduced.

	Cluster		
Experiment	**Small $< 3\%$**	**Medium $< 9\%$**	**Large $\geq 9\%$**
Hierarchical	4	3	3
FCM	1	5	4
FCM with Start U	0	6	4

Table 3: Cluster sizes of different experiments

Another finding about the statistical properties of the corresponding cluster instances concerning the dimension "age" is shown in Table 2. The year differences between first quartile and median are slightly reduced over all clusters when considering the median. However, the maximum difference over all clusters is slightly increased. The same effect occurs with the difference between the median and third quartile of the year difference. It is also slightly reduced, but the maximum difference is also slightly increased. Furthermore, the year range between minimum and maximum over all clusters is also decreased over all clusters. However, as in the previous two considerations, the maximum value is also increased. This means that most of the clusters have a slightly smaller age range between first quartile and median, as well as between median and third quartile. Considering the maximum and minimum values over all clusters the median of the age range is reduced. So, the majority of the clusters is more compact concerning the age dimension. However, there are also clusters which are broader than in the hierarchical case.

As it can be seen in Figure 5, the geographic location had more influence on cluster formation.

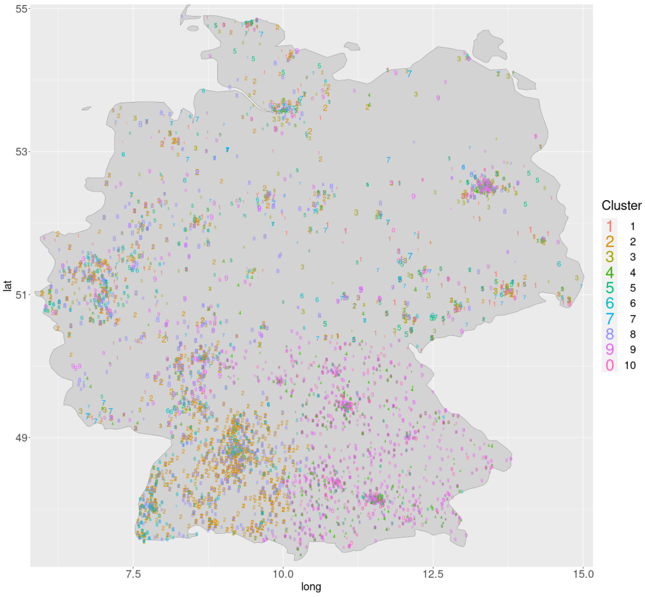


Figure 5: Visualization of the FCM cluster results (without knowledge)

In the second set of experiments, the hierarchical clusters $C_{\mathrm{HI}}1,...,C_{\mathrm{HI}}10$ were used as starting partition for the cluster membership values U. The computation yielded 10 clusters $C_{\mathrm{FH}}1,...,C_{\mathrm{FH}}10$. As shown in Table 3 no small clusters were formed.

Considering the age dimension, the results were marginally improved by a minimal decrease in the age range of first quartile to median, see Table 2). However, the results in terms of the geographic properties of clusters improved significantly, as it can be seen in Figure 6, in comparison to the hierarchical cluster result in Figure 3.

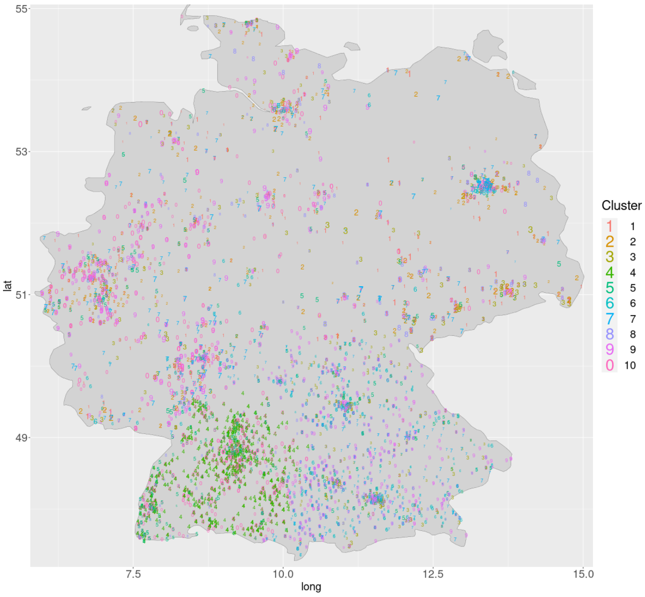


Figure 6: Visualization of the FCM cluster results with a hierarchical start partition U

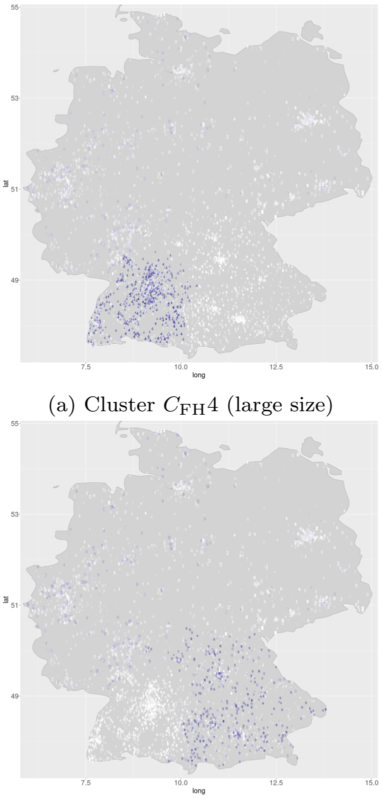
More specifically, for instance, clusters $C_{\mathrm{FH}}4$ and $C_{\mathrm{FH}}9$ have roughly the same properties as $C_{\mathrm{HI}}4$, $C_{\mathrm{F0}}2$ and $C_{\mathrm{F0}}9$, see Table 4 - being male and roughly around year of birth 1990. It seems that the fuzzy approaches found a reasonable geographic split for $C_{\mathrm{HI}}4$ into two or more separate clusters $C_{\mathrm{F0}}2$ and $C_{\mathrm{F0}}9$ for the first iteration without knowledge, whereby the combined sum of the cluster sizes of 39.0% is slightly larger than the original cluster with 38.0% of all objects. In the second iteration with a start partition generated by a hierarchical approach the results are improved even more, since $C_{\mathrm{FH}}4$ and $C_{\mathrm{FH}}9$ combined represent only 30.2% instead of 38.0%. So, the remaining instances were allocated to other, more suitable clusters.

Taking up on the geographic distinction the clusters $C_{\mathrm{F0}}2$ and $C_{\mathrm{FH}}4$ have a rather dominant region BW (short for Baden-Württemberg) and clusters $C_{\mathrm{F0}}9$ and $C_{\mathrm{FH}}9$ have a rather dominant region BY (short for Bavaria). The fainter the blue in Figure 7, the lower the membership degree to the assigned cluster. So, the fuzzy approaches managed better to associate clusters to a geographic region.

Besides the size of the clusters between first and second iteration, also the age dimension is more compact in the second iteration. Comparing the minimum and maximum values in the age dimension, the clusters found by the second iteration of the fuzzy approach are much more compact, see Table 4.

Cluster-No.	Percentage	Gender	Birth Year (19xx)					Region
			Min	1.Q	Median	3.Q	Max	
$C_{\mathrm{HI}}4$	38.0 %	male	80	87	90	93	96	all
$C_{\mathrm{F0}}2$	18.9 %	male	86	89	91	94	96	BW+
$C_{\mathrm{F0}}9$	20.1 %	male	69	85	89	91	96	BY+
$C_{\mathrm{FH}}4$	14.4 %	male	86	88	91	93	96	BW+
$C_{\mathrm{FH}}9$	15.8 %	male	88	90	91	93	96	BY+

Table 4: Selected statistical data of clusters in all three experiments (Hierarchical, FCM, FCM with hierarchical start partition U)



(a) Cluster $C_{\mathrm{FH}}4$ (large size)

(b) Cluster $C_{\mathrm{FH}}9$ (large size)

Figure 7: Visualization of selected individual clusters in FCM with hierarchical start partition (The fainter the blue symbol, the lower the membership for the corresponding cluster. The larger the symbol size the higher the underrepresentation.)

6. EVALUATION

After testing many configurations of deterministic cluster algorithms, a configuration was found which delivered a useful segmentation of the underrepresented customer groups. Yet, to prepare a concrete marketing campaign, it was an issue that the deterministic approach does not consider that some clusters overlapped more or less with others. Thus, a crisp assignment of the respective objects is not an ideal representation. It causes problems when social media campaigns are planned, since transparency is not given. The presented fuzzy approach solves these issues. The tendency towards larger clusters and smaller "residual clusters" in the results of the fuzzy approach indicates that the potential of marketing campaigns for the bigger clusters is much larger than it would be expected regarding the corresponding results of the deterministic clusters. On the other hand, for the smaller clusters which were detected by the deterministic approach the effect of targeted campaigns would have been overestimated. It is obvious that in the first case many more underrepresented customers could be targeted by means of a single campaign. In the second case, a waste of resources for campaigns which might have a lower impact due to an overestimation of the positive impact can be avoided. Thus, the fuzzy approach provides a better quality of information and a more realistic representation of the underestimated customer segments is achieved. This information is also important for the evaluation of campaigns.

As outlined in Section 5.3.2, the better quality of the fuzzy results is especially evident for the dimensions "age" and "place of residence". Since the filtered data set consisted roughly of 49 % female and 51 % male instances all clustering attempts resulted in a strict distinction between genders. This could have been mitigated by reducing the weight of the gender attribute. However, since gender is a possible configuration parameter for marketing campaigns we decided not to do so.

Additionally, it should be mentioned that the first set of the fuzzy experiments (without a starting partition of a hierarchical cluster approach) could be executed more swiftly with less computational resources. It generated reasonable results and outperformed the deterministic outcome. In this study, computational effort was not be a big issue, but when analyzing customer groups in other domains this aspect should be considered.

7. CONCLUSION AND FUTURE WORK

Due to many reasons which were mentioned at the beginning, the power of attorney business is a growing market. Since potential customers can basically come from all population groups, marketing campaigns that leverage the greatest potential on a target group-specific basis are of great importance. In order to identify major underrepresented target groups customer segmentation plays an important role. Unfortunately, this issue has been relatively unexplored in this business. By means of this research, we narrowed the gap by analyzing approaches to support targeting for marketing campaigns by means of cluster analyses and developed a course of action. All research goals could successfully be achieved.

In detail, we compared the actual regional demographic structures of the German population with the regional structures of the regarded service provider's customer base in order to identify underrepresented population groups (goals 1 and 2). In the analyzed case, mixed categorical and continuous data had to be considered. The clusters were identified by means of deterministic and fuzzy clustering approaches combined with a Gower distance to address the issue of mixed data types. As we compared the results, it became obvious that the FCM approach comes along with several benefits compared to the hierarchical cluster approach and its results were superior to those of the deterministic variants used in the present case. The FCM approach identified more or less overlapping clusters in a way that the quality of the customer segmentation was improved. The fuzzy analyses tended towards larger clusters and smaller "residual clusters". Thus, valuable information for targeting was generated and can be used for the preparation of future marketing campaigns to focus on the bigger groups and to avoid an unnecessary waste of resources for inefficient campaigns. By means of more realistic results, the post evaluation of marketing campaigns can also be carried out at a better level of information.

In any case, we recommend the FCM approach for the analyzed use case (goal 3). Even for the worst FCM result in a setting when no start partition of an deterministic approach is available, the fuzzy results outperformed the deterministic results. Obviously, the issues of FCM mentioned in Section 3 and Section 5.3.1 are not so severe in the present use case as to lead to worse results. Hence, in cases with larger data volumes and increasingly higher efforts to generate a deterministic start partition, we still recommend the FCM method.

Furthermore, it is recommended to repeat the analysis which was presented here from time to time because the structures of the underrepresented population groups in the customer base may change over time.

Nevertheless, there are still ideas for further improvements. It might be possible to differentiate the identified cluster structures by adding additional features, such as "education". The question is, however, whether this data could be obtained from governmental sources with the appropriate mapping to other demographic and geographic characteristics. Apart from that, the geographic feature had to be mapped to the level of regions due to the reasons explained in Section 4. Possibly, finer granular data could be available after the 2022 census currently conducted in Germany. Moreover, the geographic data would also be more up-to-date and other genders as already mentioned in Section 4 could be considered.

Yet, an important step has been taken toward a solution in the field of customer segmentation for the power of attorney business. With achieving all objectives which were set, marketing campaigns can now be carried out more effectively by means of the results obtained.

Acknowledgements

This research is supported by JURA DIREKT GmbH, Nuremberg/Germany. We thank Mr. Domenico Anic to make this research possible. Furthermore, we are grateful to him and his team for their time and their great support.

REFERENCES

[1] G. Asokan and S. Mohanavalli. Fuzzy Clustering for Effective Customer Relationship Management in Telecom Industry. In D. Nagamalai, E. Renault, and M. Dhanuskodi, editors, *Trends in Computer Science, Engineering and Information Technology*, pages 571–580, Berlin, Heidelberg, 2011. Springer.

[2] J. C. Bezdek. A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms. *IEEE Trans. on PAMI*, 2(1):1–8, 1980.

[3] J. C. Bezdek and J. C. Dunn. Optimal Fuzzy Partitions: A Heuristic for Estimating the Parameters in a Mixture of Normal Distributions. *IEEE Transactions on Computers*, 24(8):835–838, 1975.

[4] Bundesministerium der Justiz. Bürgerliches Gesetzbuch (BGB): § 1829 Genehmigung des Betreuungsgerichts bei ärztlichen Maßnahmen, 2023. `https://www.gesetze-im-internet.de/bgb/__1829.html` (Online; Last accessed: 2023-05-15).

[5] Bundesnotarkammer. Jahresbericht und Statistik: Das Zentrale Vorsorgeregister in Zahlen, 2023. `https://www.vorsorgeregister.de/footer/jahresbericht-und-statistik` (Online; Last accessed: 2023-05-15).

[6] S. Das and J. Nayak. Customer Segmentation via Data Mining Techniques: State-of-the-Art Review. In J. Nayak, H. Behera, B. Naik, S. Vimal, and D. Pelusi, editors, *Computational Intelligence in Data Mining*, pages 489–507. Springer Nature Singapore, 2022.

[7] N. de Vries and P. Moscato. Consumer Behaviour and Marketing Fundamentals for Business Data Analytics. In P. Moscato and N. de Vries, editors, *Business and Consumer Analytics: New Ideas*, pages 119–164. Springer Nature Switzerland, 2019.

[8] N. de Vries, L. Olech, and P. Moscato. Introducing Clustering with a Focus in Marketing and Consumer Analysis. In P. Moscato and N. de Vries, editors, *Business and Consumer Analytics: New Ideas*, pages 165–212. Springer Nature Switzerland, 2019.

[9] T. Eckes and H. Roßbach. *Clusteranalysen*. Kohlhammer, Stuttgart et al., 1980.

[10] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–872, 1971.

[11] R. Gupta, H. Kumar, T. Jain, A. Shrotriya, and A. Sinha. Demographic customer segmentation of banking users based on k-prototype methodology. *12th International Conference on Cloud Computing, Data Science & Engineering*, pages 578–584, 2022.

[12] P. M. Hoffmann, U. Hütter, M. T. Korte, and C. von Ferber. Die Lebenslage älterer Menschen mit rechtlicher Betreuung: Abschlussbericht zum Forschungs- und Praxisprojekt, 2005. `https://www.bmfsfj.de/resource/blob/78932/459d4a01148316eba579d64cae9e1604/abschlussbericht-rechtliche-betreuung-data.pdf` (Online; Last accessed: 2023-05-15).

[13] A. K. Jain. Data clustering: 50 years beyond k-means. Pattern Recognition Letters. *Award winning papers from the 19th International Conference on Pattern Recognition (ICPR)*, pages 651–666, 2010.

[14] G. Jandaghi, H. Moazzez, and Z. Moradpour. Life insurance customers segmentation using fuzzy clustering. *World Scientific News*, pages 24–35, 2015.

[15] Kester-Haeusler-Forschungsinstitut für Betreuungsrecht. Entwicklung des Betreuungsrechts / Betreuungszahlen, 2022. `http://www.betreuungsrecht.de/betreuung/entwicklung-des-betreuungsrechts-betreuungszahlen/` (Online; Last accessed: 2023-05-15).

[16] N. R. Maulina, I. Surjandari, and A. M. M. Rus. Data Mining Approach for Customer Segmentation in B2B Settings using Centroid-Based Clustering. In *16th International Conference on Service Systems and Service Management (ICSSSM), 3-15 July 2019*. IEEE, 2019.

[17] S. Mehrotra and S. Kohli. Data Clustering and Various Clustering Approaches. In S. Bhattacharyya, S. De, I. Pan, and P. Dutta, editors, *Intelligent Multidimensional Data Clustering and Analysis*, pages 90–108. IGI Global, 2016.

[18] Meta Platforms. Ad targeting: Help your ads find the people who will love your business., 2023. `https://www.facebook.com/business/ads/ad-targeting`, (Online; Last accessed: 2023-05-15).

[19] S. Munusamy and P. Murugesan. Modified dynamic fuzzy c-means clustering algorithm - application in dynamic customer segmentation. *Applied Intelligence*, 50:1922–1942, 2020.

[20] J. Nayak, B. Naik, and H. Behera. Fuzzy C-Means (FCM) Clustering Algorithm: A Decade Review from 2000 to 2014. *Computational Intelligence in Data Mining*, 2:133–149, 2015.

[21] R. Punhani, V. Arora, A. Sai Sabitha, and V. Shukla. Segmenting e-Commerce Customer through Data Mining Techniques. *Journal of Physics: Conference Series*, 1714(1):012026, 2021.

[22] H. Schäfer, J. L. Viegas, M. C. Ferreira, S. M. Vieira, and J. M. C. Sousa. Analysing the segmentation of energy consumers using mixed fuzzy clustering. In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–7, 2015.

[23] T. Schlager and M. Christen. Market Segmentation. In C. Homburg, M. Klarmann, and A. Vomberg, editors, *Handbook of Market Research*, pages 939–967. Springer Nature Switzerland, 2022.

[24] Statistische Ämter des Bundes und der Länder. Regionaldatenbank Deutschland, 2023. `https://www.regionalstatistik.de/genesis/online?operation=table&code=12111-04-01-4-B&bypass=true&levelindex=0&levelid=1659423506709#abreadcrumb` (Online; Last accessed: 2023-05-15).

[25] Statistisches Bundesamt. Bevölkerung: Mitten im demografischen Wandel, 2023. `https://www.destatis.de/DE/Themen/Querschnitt/Demografischer-Wandel/demografie-mitten-im-wandel.html` (Online; Last accessed: 2023-05-15).

[26] Statistisches Bundesamt. Bevölkerungsstand: Bevölkerung nach Familienstand, 2023. `https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/familienstand-jahre-5.html` (Online; Last accessed: 2023-05-15).

[27] Stiftung Warentest. So funktioniert die Patientenverfügung, 2022. `https://www.test.de/Vorsorgevollmacht-und-Patientenverfuegung-Wie-Sie-rechtzeitig-Klarheit-schaffen-4641470-5384641/` (Online; Last accessed: 2023-05-15).

[28] D. Zheng. Application of Silence Customer Segmentation in Securities Industry Based on Fuzzy Cluster Algorithm. *Journal of Information & Computational Science*, (10):4337–4347, 9 2013.