Cluster analysis for the derivation of agents for ABMs in the context of an ageing, super-diverse population: a mixed-methods approach

Hannah Haacke
Applied Geoinformatics
Humboldt Universität Berlin
Rudower Chaussee 16
12489 Berlin
hannah.haacke@geo.hu-berlin.de

Friederike Enssle
Cultural- and Socialgeography
Humboldt Universität Berlin
Rudower Chaussee 16
12489 Berlin
friederike.enssle@geo.hu-berlin.de

Ilse Helbrecht
Cultural- and Socialgeography
Humboldt Universität Berlin
Rudower Chaussee 16
12489 Berlin
ilse.helbrecht@geo.hu-berlin.de

Tobia Lakes
Applied Geoinformatics
Humboldt Universität Berlin
Rudower Chaussee 16
12489 Berlin
tobia.lakes@geo.hu-berlin.de

Blake B. Walker
Applied Geoinformatics
Humboldt Universität Berlin
Rudower Chaussee 16
12489 Berlin
blake.byron.walker@geo.hu-berlin.de

Abstract

Across Europe, the population is simultaneously ageing and becoming more socially diverse. However, the intersection between super-diversity and ageing has been largely absent from the literature, despite its importance for urban and social planning. Agent-Based Modelling (ABM) comprises a commonly used method to simulate and analyse urban development in the context of demographic change, but suffers from high sensitivity to parameterisation. We therefore propose a mixed-methods approach to developing agents for the purpose of ABM, using hierarchical cluster analysis, expert interviews, and focus groups. The quantitative and qualitative results are compared and contrasted to derive empirical agents. A geospatial dataset is then used to explore the feasibility of assigning multimethod-derived agents to a 'home location'. Our statistical results exhibited a high level of agreement with the qualitative analysis, while the derivation of agent locations presented some unique methodological challenges meriting further research. We conclude by asserting the utility of mixed-methods for deriving agents to be used in population modelling.
Keywords: cluster analysis, data mining, Agent-Based Modelling, mixed-methods, super-diversity

1 Introduction

The population in Europe is simultaneously ageing and diversifying (Lutz et al., 2008). In major urban centres throughout the Global North, immigration heterogeneity is accelerating, a process and state now referred to as *super-diversity* (Vertovec, 2007). Despite the acknowledged importance of super-diversity in understanding and adapting to an ageing population, only minimal research has yet explored the intersection of these social phenomena (Angel and Angel, 2006).

In Berlin, for example, nearly 25 % of the population is above the age of 65, a figure that is expected to rise along with a simultaneous rapid increase in the number of residents with a migration background (see table 1). The term *migration background* is used to refer to persons who were born outside of Germany or have more than one nationality (SBB, 2017).

The spatial distribution of an ageing, super-diverse population is of particular importance, as these populations have unique needs and are more dependent on their local social structures and networks (Mahne et al., 2017). While some studies have recently emphasised the need for governments to make cities more age-friendly (Steels, 2015; Ruza et al., 2014; OECD, 2015), it remains necessary to identify, locate, and study sub-populations with unique needs in *super-diverse* settings. This poses a significant challenge of growing importance.

Agent-Based Modelling (ABM) constitutes a commonly used toolset for simulating and analysing spatial behaviours and population change. The resulting models can be used for planning urban policy, social and economic services, infrastructure, etc.

Agents form the fundamental basis of an ABM. They are virtual objects that simulate real-world individuals' actions and interactions (Rounsevell et al., 2012). Agents react according to rules (parameters), can move through space, and interact with their environment and one another (Crooks and Heppenstall, 2012).

Agents are parameterised using one of two methods: inductive analysis (e.g. clustering) or deductive reasoning (based on theory or expert knowledge) (Rounsevell et al., 2012). Most commonly, agents' rules are selected based on

heuristic decisions made by the researcher or model developer (Macal and North, 2010). These decisions are informed through the interpretation of thematic literature reviews, interviews, or other qualitative data analysis techniques (Rounsevell et al., 2012). However, Fontaine et al. (2014) demonstrate that statistical cluster analysis may serve as a purely data-driven approach for creating agent rules. By using clusters to select and parameterise agents, the agent building process is less biased by the results of previous studies and researchers' assumptions and preferences. This also provides a potentially valuable approach to developing agent rules in subjects where little or no previous research was done, as it is the case with super-diversity and ageing.

Table 1: Population distribution in Berlin in 2016
Source: Amt für Statistik Berlin Brandenburg

Variable	People above 65	People below 65
Germans without migration BG	621 755	1 897 135
Germans with migration BG	25 489	449 502
Immigrants	52 688	624 053

In cluster analysis, quantitative data are divided into subsets (clusters), in which every datum within a cluster has attributes or characteristics that are more similar to the data in its own cluster, compared to the data in another cluster. The most common clustering algorithms are partitional and hierarchical clustering, the former of which uses a preselected number of clusters while the latter of which derives the number of clusters from the data structure itself (Hastie et al., 2017). Therefore, hierarchical clustering is more robust when seeking to reduce researcher bias in exploratory analysis. Importantly, cluster analyses are easily reproducible and can be quantitatively validated and deployed for other ABMs with a similar purpose.

This study seeks to further explore, develop, and assess the efficacy of using cluster analysis to specify agents for the purpose of modelling population change in a super-diverse setting. While Fontaine et al. (2014) used cluster analysis in combination with PCA as a tool for programming agents, we seek herein to further advance the development of empirical agent derivation and parameterisation, exploring whether the results of a multidimensional cluster analysis can produce qualitatively valid results when compared with information derived from expert interviews and focus groups.

2 Data and methods

To qualitatively identify population sub-groups with common characteristics in a super-diverse setting, and to complement and contextualise the results of the statistical cluster analysis, we conducted interviews with selected experts in Berlin. We identified and interviewed 18 experts from facilities and services that support the elderly, particularly those that cater to seniors with a migration background. These qualitative data collection activities had an explicit focus on ethnic diversity in an aging population. We also conducted expert interviews at senior citizen information centres and housing projects specifically serving homosexual seniors.

The interview transcripts were analysed in the context of preselected literatures on super-diversity and ageing. This phase comprised the extraction of key themes, round-table discussions among project participants, and triangulation with researcher observations and the results of the statistical analysis.

For this study, two datasets from the year 2014 were selected. The DEAS (Deutscher Alterssurvey) dataset provided by the DZA (Deutsches Zentrum für Altersfragen: German Centre for Aging Research) contains aspatial tabulations from a detailed sociodemographic survey with numeric and categorical data comprising work and retirement, economic situation, social relationships, leisure time, lifestyle, health, attitudes and perception of ageing, and life satisfaction. The data were collected via personal interview and written questionnaire (Klaus and Engstler, 2017). A total of 153 persons over 64 years participated in the survey (SUF DEAS, 2014).

The Einwohner Register (ER: resident registration) dataset provided by the Statistisches Bundesamt Berlin Brandenburg (National Statistics Institute of Berlin and Brandenburg) has fewer variables but contains spatial data (SSB, 2017). The attributes were cross-tabulated to produce a multidimensional table containing multivariable subgroups of interest. Gender, age, migration background, marital status, and LOR (Lebenweltlich Orientierte Raum: living-space-orientated geographical units) were provided for all 447 LOR units in Berlin. LORs are government-defined districts that form the fundamental spatial unit for population observation and planning (SSW, 2016).

The population of Berlin was divided into groups with similar attributes due to data protection policies regarding personal information, the absence of a comprehensive population census in Germany, limited computational capacity, and the need to preserve attribute consistency between datasets. The absence of spatial data and an underrepresentation of persons with a migration background in the DEAS dataset, and an insufficient level of sociodemographic detail in the ER dataset, both limited our ability to conduct a comprehensive quantitative analysis. We therefore elected to use the qualitative data to derive additional agents, thus supplementing the statistical analysis with deductive reasoning.

The cluster analysis variable selection and parameterisation was supported by the qualitative analysis results. A semi-structured content analysis of interview transcripts and interview observations was conducted together by the authors of this study in multiple round-table meetings. Key themes were extracted from the qualitative data and iteratively compared to the list of variables and cluster results. Selecting input parameters was of particular importance for the DEAS dataset, which contained over 200 attributes. The restriction of attributes was necessary because an excessive number of attribute or the inclusion of irrelevant attributes can distort the cluster analysis results (Bortz, 2005).

For the spatial cluster analysis, numeric data were rescaled as recommended by Kabacoff (2015). The distance matrix for the ER dataset was calculated with the Euclidean distance. The aspatial distance matrix for the DEAS dataset was

calculated using Gower distance, which standardises each variable. The distance between two units is then calculated as the sum of all variable-specific distances (Maechler et al., 2017).

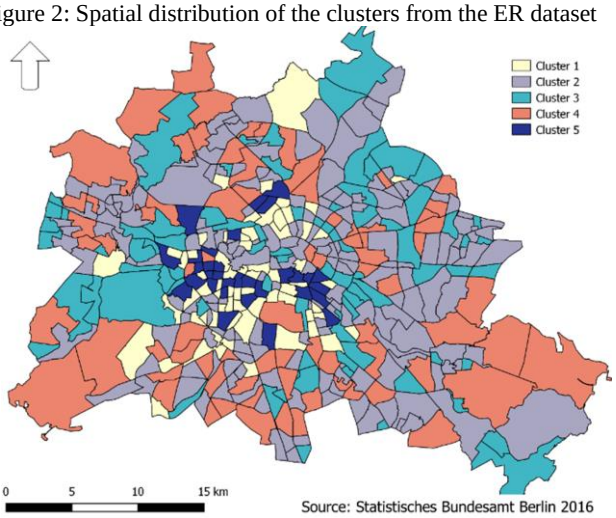
We conducted a preliminary analysis using various methods for cluster definition and validation, and selected the Ward method, as it provided the most reasonable and rigorous results. The Ward method tends to fit clusters as homogeneously as possible (Murtagh and Legendre, 2014), and was therefore deemed suitable for our study. This technique tends to create clusters with small and relatively similar numbers of observations (Kabacoff, 2015). It starts with *n* clusters, each containing a single observation, then iteratively grouping them together in such that every step produces the smallest possible increase in the within-cluster variance until all observations are finally grouped into one cluster. The total variance for each iteration is plotted on the scree plot, which is used in combination with the elbow criterion to select the optimal number of clusters. The scree plot shows the proportion of the variance within each cluster and the number of clusters after which the variance drops off (see figure 1). As the goal is to find the cluster with the smallest variance while retaining as few clusters as deemed valid, this 'elbow' point at which the variance in the scree plot drops is chosen by the researchers (James et al., 2017). All calculations were completed using R v.3.3.3 (R, 2017).

3 Results

For the ER dataset, 5 spatial clusters were derived based on the scree plot shown in figure 1. The mapped clusters are shown in figure 2. Aside from clusters 1 and 5 the spatial distribution of these groups is dispersed, indicating that while they share similar population characteristics, but do not spatially cluster.

Figure 3 indicates that between-cluster differences in population structure were relatively minor. There is minimal discernible heterogeneity between, for example, married people, people with migration background, or older people. However, some trends are visible. Cluster 1 and 5 appear to have a younger age structure with more individuals with a migration background and more single/divorced persons. Cluster 3 contains more married people without a migration background. Through the expert interviews and focus groups we identified spatial clusters of persons with specific migration backgrounds, supporting this finding.

Figure 2: Spatial distribution of the clusters from the ER dataset



From the clusters, it can be inferred that the gender has no influence on the agents' location in space. The most dominant variables appear to be marital status and migration background (see table 4 in the appendix). However, it is necessary to note that this dataset has a limited number of variables upon which to discern clusters, such as educational background or social networks. However, it is reasonable to infer that marital status and the migration background has a greater influence on place of residence than gender, the spatial distribution of which tends to be relatively homogeneous.

Qualitative analysis, in combination with the cluster analysis of the aspatial DEAS dataset, identified 8 different agents (see figure 4), which have been derived with different attributes (see scree plot in figure 1). The resulting agents can be seen in table 2. In this table the agents are described according to common characteristics and named for convenience. The attributes were derived from table 3 in the appendix.

Figure 3: Resulting graphs of the ER dataset

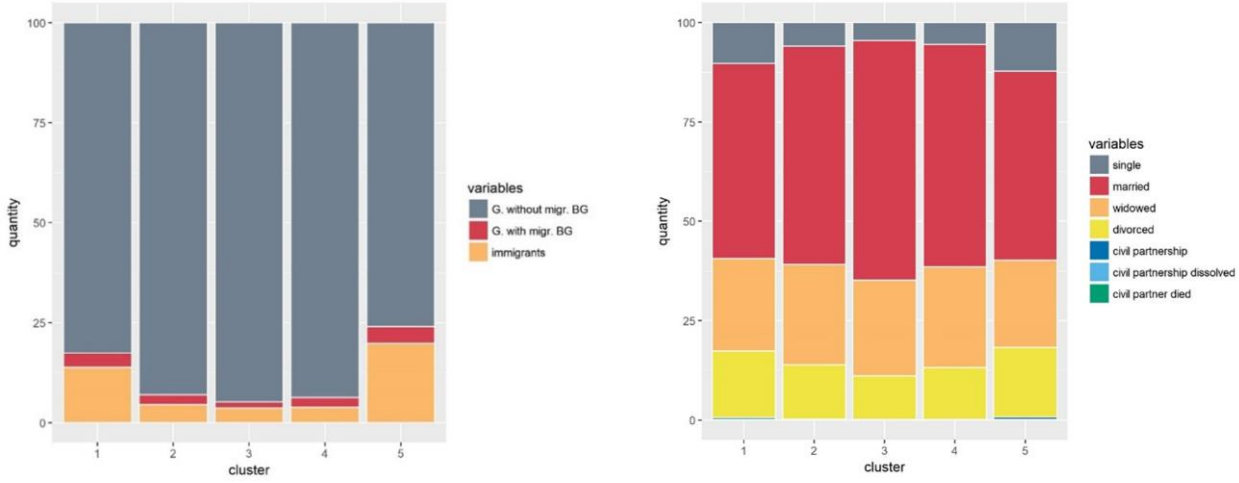
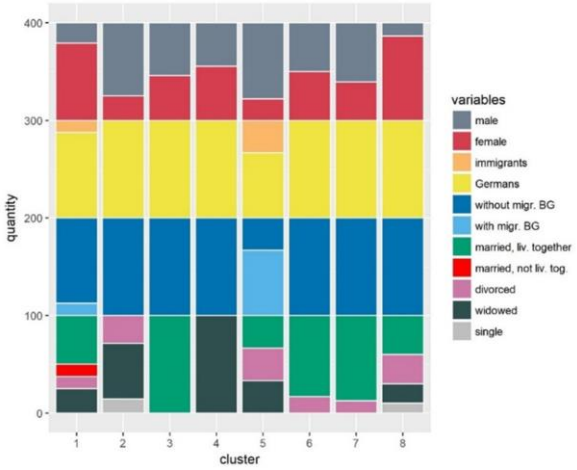


Figure 4: Example of a resulting graph from the DEAS dataset



The comparison of spatial clusters and aspatial cluster results, in combination with the qualitative analysis results, provided adequate corroborating evidence for locating Herr/Frau Meyer, Schmidt and Becker (spatial cluster 3). Miriam and Alexis are more likely to live in spatial cluster 1 or 5.

4 Discussion

Consulting our results of expert interviews, the results of the cluster analysis seem to be reasonable even though there are some differences. This were to be expected as it is not possible to cover the entire range of the population with interviews

It can be seen, that some variables give a clear impression for the resulting agents, while other variables' influence remain unclear. In other cases, a declaration about a clusters' important attributes can only be made in comparison to another cluster.

Form the results it can be derived that, for example, agent Jakob is more likely move in the near future, than Herr/Frau Becker because Jakob lives alone and is older. This means that he is more likely not to be able to care for himself, get health issues in the near future, be affected by rising rents or to die. Herr/Frau Becker, on the other hand, own their flat. Therefore, they are unaffected by rising rents. Furthermore, they are in twos which means that it is more likely that one of them is able to take care of the other. Also, it is possible that they have a higher income, and can thus afford to pay a nurse. However, we also know from expert interviews, that people move easier when they are in pairs, whereas people who live alone pay disproportionately higher taxes and and are thus not able to move at all. From the example it can be seen that the problem is very complex and several layers need to be taken into account.

The results of the cluster analysis in the DEAS dataset can help to derive behaviour rules. As we found out that the elderly population prefers to stay in their homes, independent on their background, the most important task will be to find thresholds which cause agents to move. As we know that people nearly never move voluntarily, it is necessary to find out under what circumstances this happens. Furthermore, involuntary aspects for moving need to be figured out (rent, health issues, etc.).

It is difficult to tell whether the migration background and thus the super-diverse aspect, has a big influence on the place of residence. In the ER dataset, there is not enough information about the social background, while in the DEAS dataset only 3 persons have a migration background. Therefore, it is difficult to come to a reliable conclusion. In the future it has to be established if super-diversity needs to be analysed or if diversity is enough to cover the background.

That not all possible aspects are covered is also noticeable in some clusters. For instance, there is the cluster where the marital status is 100% widowed, but still 10% are living together with their partner. Which leads to two possibilities: a) widowed but in a relationship or b) the person is married but did not answer the question to his/her relationship. The clusters are not false but still need some further interpretation.

There are other, deterministic, methods to validate the cluster. However, if the results are validated with a qualitative approach, the clusters can be used, independent of their statistical validity. Another aspect that needs to be mentioned is that the data is not covering all attributes necessary to

Table 2: Resulting clusters of the DEAS dataset and qualitative analysis, named and summarized

Name	Description
Miriam	Female, married, eventually with migration background, children, only 1/3 worked before being a pensioner, mostly living in twos, moved recently into their flat, tenant, live in a good living situation, best relationship with their family
Jakob	Male, not married, oldest cluster, no migration background, born in former eastern regions of Germany, lower education, children, highest amount of formerly unemployed, live alone, moved just recently into the flat, tenant, worst relationship to their family, sees friends scarcest
Herr/Frau Schmidt	Married, no migration background, children, no one formerly unemployed, live in twos, half own their homes, live together with their partner, health status of partner is good, participates scarcest in group meetings
Herr/Frau Meyer	Widowed, no migration background, no children, highly educated, live alone, homeowner, live the longest on their flats, 4 rooms
Alexis	Male, 2 children, migration background, no one formerly unemployed (but biggest cluster with early retirement), live alone, 1 – 2 rooms
Frau/Herr Becker	Married, children, no migration background, highly educated, does not live alone, homeowner, 5 – 6 rooms, living situation good or very good, health status of partner is good, meet friends often
Frau/Herr Hoffmann	Married, children, no migration background, did not work before, live in twos, very good relationship to the family, participates most often in groups
Anna	Female, children, no migration background, lower education, tenant, some live with their children, worst health status partner

derive agents. Through the interviews, we know that there must be, for example, an agent living from basic income. It is possible to add this agent, define its behaviour and include it with the other agents into the ABM. Furthermore, it might be necessary to split the agents with (im)migration background into several subcategories like guest workers or people who came as student to Germany.

5 Conclusion

Cluster analysis for finding agents in population data provides an independent approach with minimal bias. As the input variables need to be estimated and the resulting clusters have to be validated, expert interviews were conducted. The results of the cluster analysis were 8 clusters for the DEAS dataset without spatial distribution and 5 clusters for the ER dataset with spatial distribution but not many variables that provide background information about the population.

The mixed-methods approach using cluster analysis and expert interviews enabled us to identify consistent and defined population sub-groups, which are then suitable for future deployment in an Agent-Based Model. Cluster analysis enables to build reproducible agents and the approach can be used for other datasets with similar aim as well. However, this is only the first step in the project. Behaviors and location need to be derived to use the agents in an ABM, therefore the resulting clusters need to be merged. That means, that it needs to be estimated where agents are in the ER clusters.

Beyond the scope of this study, the future modelling and validation of an ABM for super-diverse population will enable researchers to further evaluate the efficacy of a mixed-methods approach to ABM parameterization and development.

Acknowledgements

This work was carried out within DFG grant HA 3484/8-1.

References

Anderberg, M. (1973) *Cluster analysis for applications*. New York [u.a.], Academic Press.

Angel, R. & Angel, J. (2006) Diversity and ageing in the United States. In: Binstock, R. & George, L. (eds.) *Handbook of aging and the social sciences*, 6th ed. London, Academic Press, pp. 4 – 110.

Bortz, J. (2005) *Statistik für Human- und Sozialwissenschaftler*. 6th ed. Heidelberg, Springer.

Crooks, A. & Heppenstall, A. (2012) Introduction to agent-based modelling. In: Heppenstall, A., Crooks, A., See, L. & Batty, M. (eds.) *Agent-based models of geographical systems*, Dordrecht, Springer, pp. 85 – 108.

Fontaine, C. M., Rounsevell, M. D. A. & Barbette, A.-C. (2014) From actors to agents in socioecological systems models. *Environment and Planning B: Planning and Design*, 1, 163 – 184.

Hastie, T., Tibshirani, R. & Friedman, J. (2017) *The elements of statistical learning – data mining, inference, and prediction*. 2nd ed. New York, Springer.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2017) *An introduction to statistical learning with applications in R*. 8th ed. New York, Springer.

Kabacoff, R. I. (2015) *Data analysis and graphics with R*. 2nd ed. Shelter Island, Manning.

Klaus, D. & Engstler, H. (2017) Daten und Methoden des deutschen Alterssurvey. In: Mahne, K., Wolff, J. K., Simonson, J. & Tesch-Römer, C. (eds.) *Alter im Wandel: Zwei Jahrzehnte Deutscher Alterssurvey (DEAS)*. Springer: Open-access, pp. 29– 46.

Lutz, W., Sanderson, W. & Scherbov, S. (2008) The coming acceleration of global population ageing. *Nature*, 451, 716 –719.

Macal, C. M. & North, M. J. (2010) Tutorial on agent-based modelling and simulation. *Journal of simulation*, 4, 151 – 162.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2017) *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.6.

Mahne, K., Wolff, J. K., Simonson, J. & Tesch-Römer, C. (2017) Altern im Wandel: Zwei Jahrzente Deutscher Alterssurvey. In: Mahne, K., Wolff, J. K., Simonson, J. & Tesch-Römer, C. (eds.) *Alter im Wandel: Zwei Jahrzehnte Deutscher Alterssurvey (DEAS)*. Springer: Open-access, pp. 11- 28.

Murtagh, F. & Legendre, P. (2014) Ward's hierarchical clustering method: which algorithms implement Ward's criterion? *Journal of classification*, 31, 274 – 295.

OECD (2015) *Ageing in cities, OECD Publishing. Paris*. [Online] Available from: http://dx.doi.org/10.1787/9789264231160-en [Accessed 13th February 2018].

R Core Team (2017) *R: A language and environment for statistical computing*. [Online] Available from: https://www.R-project.org [Accessed 5th February 2018].

Rounsevell, M. D. A., Robinson, D. T. & Murray-Rust, D. (2012) From actors to agents in socio-ecological system models. *Philosophical transactions of the royal society B*, 367, 259 – 269.

Ruza, J., Kim, J. I., Leung, I., Kam, C. & Man Ng, S. Y. (2014) Sustainable, age-friendly cities: An evaluation

framework and case study application on Palo Alto, California. *Sustainable Cities and Society*, 14, 390 – 396.

Amt für Statistik Berlin Brandenburg (SBB) (2017) *Einwohnerregister*. [Online] Available from: https://www.statistik-berlin-brandenburg.de/Statistiken/statistik_met.asp?Ptyp=650&Sageb=12041&creg=BBB&anzwer=10 [Accessed 5th February 2018].

Senatsverwaltung für Stadtentwicklung und Wohnen (SSW) (2016) *Lebensweltlich orientierte Räume (LOR) in Berlin - Planungsgrundlagen*. [Online] Available from: http://www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/lor/ [Accessed 5th February 2018].

Steels, S. (2015) Key characteristics of age-friendly cities and communities: A review. *Cities*, 47, 45 – 52.

SUF DEAS (2014) *Version 1.0*. DOI: 10.5156/DEAS.2014.M.001.

Vertovec, S. (2007) Super-diversity and its implications. *Ethnic and Racial Studies*, 0 (6), 1024 – 1054.

Appendix

Table 3: Results of cluster analysis of the DEAS dataset (selection)

Variables		Cluster 1 [%]	2 [%]	3 [%]	4 [%]	5 [%]	6 [%]	7 [%]	8 [%]
Gender	Male	21	75	54	44	78	50	61	14
	Female	79	25	46	56	22	50	39	86
Nationality	Foreign	13	100	100	100	33	100	100	100
	German	87	100	100	100	67	100	100	100
Migration BG	Without	87	100	100	100	33	100	100	100
	With	13	100	100	100	67	100	100	100
Marital status	Married	63	0	100	0	33	83	88	40
	Divorced	13	29	0	0	33	17	13	30
	Widowed	25	57	0	100	33	0	0	20
	Single	0	14	0	0	0	0	0	10
High school diploma	Lowest secondary school	29	50	11	0	0	0	38	70
	Middle secondary school	14	17	33	0	50	0	12	10
	Highest secondary school	57	17	33	100	50	100	50	20
	No diploma	0	17	0	0	0	0	0	0
Children	0	8	11	8	56	11	9	26	18
	1	29	16	29	0	11	23	39	32
	2	33	42	38	22	56	41	22	41
	3	13	16	17	22	11	18	4	9
	> 3	17	16	21	0	11	10	8	0
Persons in household	1	25	80	4	89	63	9	30	45
	2	71	20	91	11	37	77	70	45
	> 2	4	0	4	0	0	14	0	10
Amount of rooms	1	0	25	0	0	25	0	0	8
	2	25	50	11	0	75	0	38	33
	3	33	13	33	0	0	60	0	0
	4	25	13	22	100	0	0	50	33
	5	17	0	11	0	0	60	0	0
	6	0	0	22	0	0	20	0	0
Flat	Owner	21	5	48	67	25	95	22	14
	Tenant	75	90	43	11	75	0	78	86
	Other	4	5	9	22	0	5	0	0
Meeting acquaintances	Daily	0	5	0	11	0	0	0	5
	Several times per week	13	10	0	33	22	9	22	9
	Once per week	8	15	4	11	11	14	22	9
	1 - 3 times per month	71	10	63	22	33	77	43	55
	Rarer	8	45	33	11	33	0	13	23
	Never	0	15	0	11	0	0	0	0

Table 4: Results of cluster analysis of the ER dataset

Variables		Cluster 1 [%]	2 [%]	3 [%]	4 [%]	5 [%]
Age	65 -70	27	23	24	22	30
	70 – 75	29	29	29	29	29
	75 - 80	22	24	23	25	21
	80 +	22	24	24	25	21
Gender	Male	44	43	45	42	45
	Female	57	57	55	58	55
Migration Background	Germans without migr. BG	83	93	95	94	76
	Germans with migr. BG	4	2	2	3	4
	Immigrants	14	5	4	4	20
Marital status	Single	10	6	5	5	12
	Married	49	55	60	56	48
	Divorced	17	14	11	13	17
	Civil Union	0.48	0.19	0.14	0.15	0.63
	Civil Union lifted	0.02	0.01	0	0	0.05
	Civil Union died	0.07	0	0.01	0	0.06