Bauhaus-Universität Weimar
Informatik in der Architektur | InfAR

Arbeitspapiere
Working Papers

Nr. 11, Februar 2012

Wissen wer wo wohnt

Dominik Kalisch

ISSN 2191-2416

Bauhaus-Universität Weimar, Professur Informatik in der Architektur, Belvederer Allee 1, 99421 Weimar
Fon: +49/3643/584201, caad@architektur.uni-weimar.de, http://infar.architektur.uni-weimar.de

Dominik Kalisch
Wissen wer wo wohnt

Weimar 2012
Arbeitspapiere (Working Papers) Informatik in der Architektur, Nr. 11
Herausgegeben von Prof. Dr. Dirk Donath und Dr. Reinhard König
ISSN 2191-2416

Bauhaus-Universität Weimar, Professur Informatik in der Architektur
Belvederer Allee 1, 99425 Weimar
http://infar.architektur.uni-weimar.de

Titelbild: Jugendstil-Wendeltreppe im Hauptgebäude © Bauhaus-Universität Weimar

Redaktionelle Anmerkung:

Dominik Kalisch ist Wissenschaftlicher Mitarbeiter an der Professur Informatik in der Architektur an der Bauhaus-Universität Weimar.

Der Text ist im Rahmen des von der DFG geförderten Forschungsprojekts „CoMStaR: Computerbasierte Methoden für eine sozial nachhaltige Stadt- und Raumplanung“ (DO 551/18-2) entstanden.
http://infar.architektur.uni-weimar.de/service/drupal-cms/comstar

Wissen wer wo wohnt

Dominik Kalisch
dominik.kalisch@uni-weimar.de
Professur Informatik in der Architektur
Fakultät Architektur, Bauhaus-Universität Weimar, Belvederer Allee 1, 99421 Weimar, Germany

Abstract

In cities people live together in neighbourhoods. Here they can find the infrastructure they need, starting with shops for the daily purpose to the life-cycle based infrastructures like kindergartens or nursing homes. But not all neighbourhoods are identical. The infrastructure mixture varies from neighbourhood to neighbourhood, but different people have different needs which can change e.g. based on the life cycle situation (Duranton and Puga, 2001) or their affiliation to a specific milieu (Scheiner, 2005 or Bourdieu, 1982). We can assume that a person or family tries to settle in a specific neighbourhood that satisfies their needs. So, if the residents are happy with a neighbourhood, we can further assume that this neighbourhood satisfies their needs. The socio-oeconomic panel (SOEP) of the German Institute for Economy (DIW) is a survey that investigates the economic structure of the German population. Every four years one part of this survey includes questions about what infrastructures can be found in the respondents neighbourhood and the satisfaction of the respondent with their neighbourhood. Further, it is possible to add a milieu estimation for each respondent or household. This gives us the possibility to analyse the typical neighbourhoods in German cities as well as the infrastructure profiles of the different milieus. Therefore, we take the environment variables from the dataset and recode them into a binary variable – whether an infrastructure is available or not. According to Faust (2005), these sets can also be understood, as a network of actors in a neighbourhood, which share two, three or more infrastructures. Like these networks, this neighbourhood network can also be visualized as a bipartite affiliation network and therefore analysed using correspondence analysis. We will show how a neighbourhood analysis will benefit from an upstream correspondence analysis and how this could be done. We will also present and discuss the results of such an analysis.

Keywords: Urban planning, cluster analysis, network analysis, urban research-quantitative, urban research-milieu, neighbourhood analysis, complex data analysis, singular value decomposition, Germany.

1. Einleitung

Es ist evident, dass für eine umfassende Multi-Agenten-Simulation zur residentiellen Segregation entsprechende (Übergangs-)Regeln für den Standortwechsel eines Agenten zu formulieren sind. Diese müssen, wenn es sich um den Versuch der Abbildung der Wirklichkeit handelt, auf empirischen Daten beruhen, die repräsentativ für den Untersuchungsgegenstand sind. Eine Studie zur Herleitung solcher Regeln muss zum einen den wissenschaftlichen Anforderungen genügen und zum anderen eine Aussage über die zu simulierenden Parameter und die formale Gestaltung des Regelwerks treffen. Bei den Untersuchungen im Rahmen des vorliegenden Projekts interessieren wir uns insbesondere für die Verteilung von Versorgungsstrukturmerkmalen im Raum sowie deren Nähe zu spezifischen sozialen Milieus. Trotz der vielen vorhandenen Untersuchungen zum menschlichen Mobilitätsverhalten genügte keine der Studien unseren Anforderungen. Keine der Studien macht eine repräsentative Aussage über die Wohnstandortbedürfnisse von Bevölkerungsgruppen, die als Indikator für die Wohnzufriedenheit dienen können und als Grundlage zur Erstellung von Übergangsregeln geeignet gewesen wären. Daher musste ein Analyseverfahren entwickelt werden, mit dem aus einer bereits bestehenden Studie die Ableitung von Wohnstandortbedürfnissen der jeweiligen Milieus möglich wird.

Bereits Keim (1979) und Mathiesen (1998) zeigten, dass Milieudifferenzierungen, ergänzend zu tradierten sozialstrukturellen Unterscheidungen der Bevölkerung nach Klassen, Schichten, Ethnien und Generationen, in der Stadtforschung ein tieferes Verständnis sozialer Strukturen und Prozesse in urbanen Räumen ermöglicht (vgl. Breckner 2003, 2004). In der Vergangenheit wurden jedoch bislang eher hierarchische Modelle in den Mittelpunkt der Untersuchungen gerückt, obgleich in der Stadtforschung auch zunehmend auch auf die Notwendigkeit der Analyse der vertikalen Ebene verwiesen wird. So kommt Scheiner & Holz-Rau zu dem Ergebniss, dass „travel mode is more affected by life situation then by life style. However, life style plays an important role by affecting location attitudes and location decisions." (Scheiner and Holz-Rau, 2007, S. 508). Was die Mobilität der Menschen anbelangt, so stellt Becker (2003) in seiner Dissertation fest: „Für die Wohnstandortwahl sind die jeweiligen Lebensbedingungen von Bedeutung. So sind für Wanderungen zwischen Regionen Arbeit, Ausbildung und Freizeit ausschlaggebend, während für die Nahwanderung die Wohnverhältnisse eine größere Rolle spielen." (Becker, 200, S. 105). Für das DFG-Forschungsprojekt „CoMStaR" wurde daher versucht den Milieuansatz in die Untersuchung mit einzubeziehen.

Als einzige repräsentative Studie zur Untersuchung der Wohnstandortbedürfnisse kam das Sozioökonomische Panel (SOEP) als Längsschnittstudie des Deutschen Instituts für Wirtschaftsforschung (DIW) - hier die Welle von 2004 - in Betracht, da in diesem Jahr neben den üblichen ökonomischen Fragen der Schwerpunkt auf das Thema Wohnen und Wohnumfeld gelegt wurde und für diese Erhebungswelle auch die SINUS-Milieus des SINUS Sociovision zugespielt werden konnten. Da jedoch eine Auswertung der Wohnstandortbedürfnisse in der oben beschriebenen Form nicht vorlag und auch mit den klassischen Verfahren nicht möglich war, wurde ein der Korrespondenzanalyse ähnliches Verfahren adaptiert und weiterentwickelt.

2. Verfahren

Das Ziel der vorliegenden Analyse war die Identifizierung von typischen Nachbarschaften für beliebige Bevölkerungsgruppen anhand von Infrastrukturmerkmalen in der Wohnumgebung. Um die befragten Personen zu Gruppen zusammenschließen zu können muss eine Klassifizierungsstrategie herangezogen werden, die es ermöglicht die Gruppen zu finden, die innerhalb der Gruppe eine maximale Homogenität und zwischen den Gruppen eine maximale Heterogenität aufweisen. Da sich die vorliegenden Daten auf Grund Ihrer Komplexität nicht in ihrer ursprünglichen Struktur analysieren lassen, wurden sie erst mit einem korrespondenzanalytischen Verfahren für die Clusteranalyse vorbereitet. Hierzu werden in einem ersten Schritt die anteiligen Infrastrukturmerkmalskombinationen der im SOEP erhobenen Infrastrukturen (siehe Tabelle ??) mit Hilfe der Singulärwertzerlegung nach einem Ansatz von Faust (2005) berechnet und analysiert. In einem zweiten Schritt wurden, mit Hilfe von Distanzmaßen, die Infrastrukturmerkmale bestimmt, die eine besondere Bedeutung in der jeweiligen Merkmalskombination haben. Abschließend wurden mit Hilfe eines Clustering-Verfahrens, typische Bewohnermilieus eines Nachbarschaftstypen lokalisiert.

2.1. Singular Value Decomposition

Bei der Singulärwertzerlegung einer komplexen Matrix $D \in \mathbb{C}^{m*n}$ vom Rang r , ist D das Produkt (1.1) aus der unitären[1] Matrix $U \in \mathbb{C}^{m*m}$, der adjungierten[2] V' der unitären Matrix $V \in \mathbb{C}^{n*n}$ sowie der reellen diagonal Matrix Σ^{n*n}.[3]

$$D = U\Sigma V' \tag{1.1}$$

1 Eine Matrix ist unitär, wenn die Spalten einer komplexen quadratischen Matrix orthonormal zueinander sind.

2 Eine adjungierte Matrix ist eine Matrix, deren Spalten vertauscht und deren Einträge konjugiert (gespiegelt) sind.

3 Die ursprüngliche Matrix Σ hat die gleiche Größe wie D. Allerdings sind alle Einträge der außerhalb der Diagonalen 0.

U und V sind dabei die rechten und linken Singulärvektoren von D. Der Rang einer Matrix stellt die Anzahl der linear unabhängigen Zeilen dar, die entsprechend identisch ist mit der Anzahl linear unabhängiger Spalten. Orthogonale Transformationen haben auf die Anzahl linear unabhängiger Zeilen und Spalten keinen Einfluss. Daraus ergibt sich, dass der Rang einer Matrix gleich der Anzahl der Singulärwerte ist, die ungleich Null sind. Werden nun alle r Dimensionen verwendet, so reproduzieren Σ U und V die Matrix D vollständig. Werden hingegen weniger als r Dimensionen verwendet, stellt das Ergebnis eine Approximation an D dar. Eine graphische Darstellung der Zerlegung ist in Figure 1 dargestellt.

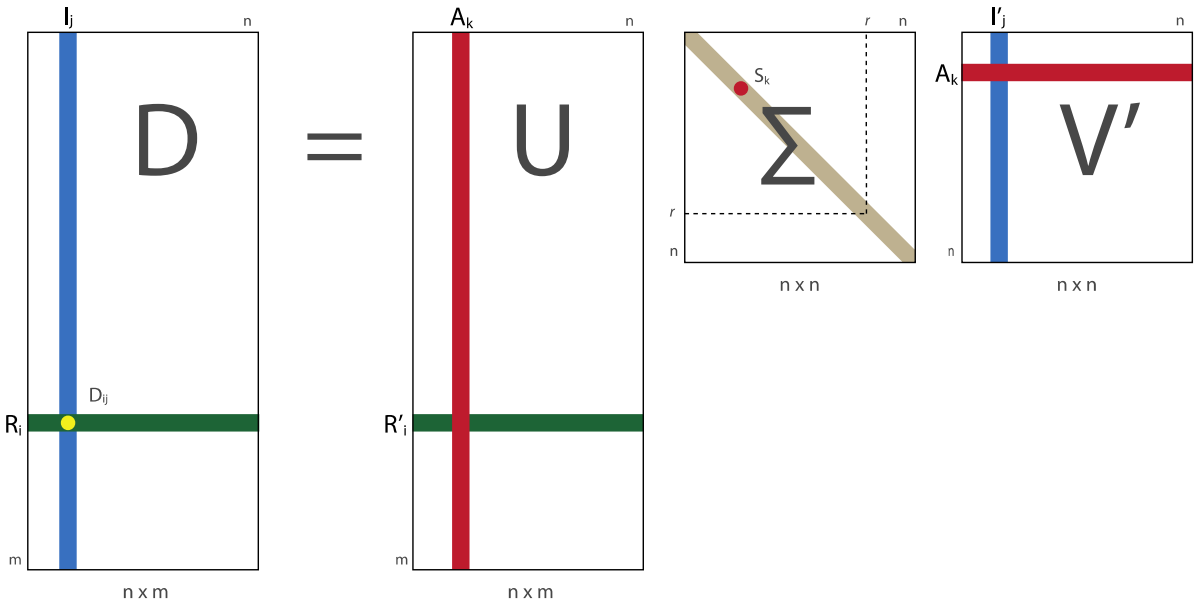

Figure 1: Bedeutung der Matrizen, Eigene Darstellung nach Faust (2005)

Gegeben sei eine $4*5$ Matrix mit den folgenden Werten[4]:

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 2 \\ 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \end{bmatrix} \tag{1.2}$$

Wird diese Matrix D entsprechend (1.1) in die Matrizen $U\Sigma V^{'}$ zerlegt, so erhält man folgendes Ergebnis:

4 Das Beispiel entstammt aus dem Wikipedia-Artikel zur SVD http://en.wikipedia.org/wiki/Singular_value_decomposition (Stand 21.03.2011)

$$U = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{1.3}$$

$$\Sigma = \begin{bmatrix} 4 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{5} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{1.4}$$

$$V' = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix} \tag{1.5}$$

Dabei kann gezeigt werden, dass UU' (1.6) und VV' (1.7) jeweils unitäre Matrizen (1.8) (1.9) sind.

$$UU' = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{1.6}$$

$$VV' = \begin{bmatrix} 0 & 0 & \sqrt{0.2} & 0 & -\sqrt{0.8} \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & \sqrt{0.8} & 0 & \sqrt{0.2} \end{bmatrix} * \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \sqrt{0.2} & 0 & 0 & 0 & \sqrt{0.8} \\ 0 & 0 & 0 & 1 & 0 \\ -\sqrt{0.8} & 0 & 0 & 0 & \sqrt{0.2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{1.7}$$

$$U*U'=I \tag{1.8}$$

$$V*V'=I \tag{1.9}$$

Die Elemente σ_i der Diagonalen von Σ (1.4) sind die Singulärwerte von D. Sie entsprechen der Quadratwurzel aus den Eigenwerten von $D'D$. Die Singulärwerte werden dabei absteigend sortiert (1.10).

$$\sigma_1 \geq \ldots \geq \sigma > 0. \tag{1.10}$$

Zusammenfassend kann man sagen, dass das Ziel einer SVD die Projektion eine Menge von n Punkten eines p-dimensionalen Raums $\mathbb{R}^p$ in einen q-dimensionalen Unterraum $\mathbb{R}^q$ ist, in dem die enthaltenen Informationen maximal und die redundanten Informationen minimal sind.

Die SVD bildet das Fundament vieler Methoden, die versuchen ein linear inverses Problem[5] zu lösen. In der Statistik bildet sie die Grundlage für die Korrespondenz- und Hauptkomponentenanalyse. Die Korrespondenzanalyse wird häufig benutzt um kategoriale Variablen in Kontingenztabellen und Adjazenzmatritzen zu analysieren. Sie kann aber auch eingesetzt werden um soziale Netzwerke, insbesondere sog. „afiliation networks" zu analysieren (vgl. Faust, 2005, S. 123).

2.2. Korrspondenzanalyse

Wenn eine Matrix D in Form von (1.11) gegeben ist, dann werden für eine Korrespondenzanalyse die Zeilen und Spalten der Matrix normalisiert (1.12) und (1.13). Das heißt, die Korrespondenzanalyse ergibt sich aus der Gleichung (1.14).

$$D = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & \cdots & a_{j1} \\ 0 & 0 & 0 & 1 & 1 & \cdots & a_{j2} \\ 1 & 1 & 0 & 1 & 0 & \cdots & a_{j3} \\ 1 & 0 & 0 & 0 & 1 & \cdots & a_{j4} \\ 0 & 1 & 1 & 0 & 0 & \cdots & a_{j5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ a_{i1} & a_{i2} & a_{i3} & a_{i4} & a_{i5} & & a_{ij} \end{bmatrix} \tag{1.11}$$

$$C^{-\frac{1}{2}} = diag\left(\frac{1}{\sqrt{\sum_{i=1}^{g} a_{ih}}} \right) \tag{1.12}$$

5 Eine Fragestellung wird in der Mathematik als inverses Problem beschrieben, wenn von den beobachteten Wirkungen eines Systems auf die der Wirkung zugrunde liegenden Ursachen geschlossen werden soll.

$$R^{-\frac{1}{2}} = diag\left(\frac{1}{\sqrt{\sum_{j=1}^{h} a_{gj}}}\right) \tag{1.13}$$

$$R^{-\frac{1}{2}} D C^{-\frac{1}{2}} = U \Sigma V' \tag{1.14}$$

U und V sind hier die jeweiligen linken und rechten Singulärvektoren (siehe Figure 1) und Σ beinhaltet die Singulärwerte σ_k. Entsprechend produziert (1.15) wieder die Ursprungsmatrix D (vgl Faust, 2005, S. 125).

$$D = R^{-\frac{1}{2}} U \Sigma V' C^{-\frac{1}{2}} \tag{1.15}$$

Nimmt man eine Matrix an, wie sie in Table 1 dar gestellt ist, und wendet das Verfahren der SVD auf diese an, so erhält man die in Table 2 dargestellten drei Matrizen.

Table 1: Beispielverteilung der Infrastrukturmerkmale

	Merkmal 1	Merkmal 2	Merkmal 3
Haushalt 1	1	0	1
Haushalt 2	0	0	1
Haushalt 3	0	1	1
Haushalt 4	1	1	0
Normalisierte Matrix $R^{-\frac{1}{2}} \Sigma C^{-\frac{1}{2}}$			
Haushalt 1	0,500	0,000	0,408
Haushalt 2	0,000	0,000	0,577
Haushalt 3	0,000	0,500	0,408
Haushalt 4	0,500	0,500	0,000

Table 2: SVD der "normalisierten" Beispielmatrix

	Dimensionen		
Linke Singulärvektoren U	1	2	3
Haushalt 1	-0,535	-0,120	0,707
Haushalt 2	-0,378	-0,676	0,000
Haushalt 3	-0,535	-0,120	-0,707
Haushalt 4	-0,535	0,717	0,000
Rechte Singulärvektoren V	1	2	3
Merkmal 1	-0,535	0,463	0,707
Merkmal 2	-0,535	0,463	-0,707
Merkmal 3	-0,655	-0,756	0,000
Singulärwerte Σ	1,000	0,645	0,500

Werden nur der erste Singulärwert und die jeweils ersten Singulärvektoren benutzt, so würden diese, unter der Annahme eines statistisch Unabhängigen Modells, die erwartbaren Häufigkeiten der Matrix D reproduzieren (1.16).

$$\frac{\sum_{j=1}^{h} a_{gj} \sum_{i=1}^{g} a_{ih}}{\sum_{j=1}^{h} \sum_{i=1}^{g} a_{ij}} = \sqrt{\sum_{j=1}^{h} a_{gj} \sum_{i=1}^{g} a_{ih}} U_{i1} \sigma_1 V_{j1} \tag{1.16}$$

Daher wird die erste Dimension in der weiteren Analyse nicht berücksichtigt, da diese dementsprechend lediglich eine Funktion der Randverteilung ist und keine Struktur der Zeilen und Spalten beinhaltet (Faust, 2005, S. 126).

Dabei stellen die Dimensionen jedoch nicht mehr einzelne Infrastrukturmerkmale dar, sondern jede Dimension stellt eine im Sinne der SVD bedeutende Infrastrukturmerkmalskombinationen dar (siehe auch Figure 1). Um nun sowohl die Haushalte als auch die Infrastrukturmerkmale gemeinsam graphisch darstellen und auswerten zu können, müssen die Werte nach der SVD in ein einheitliches Referenzsystem überfürt werden. Dazu gibt es grundsätzlich zwei mögliche Herangehensweisen, die im Folgenden kurz skizziert werden sollen.

2.3. Punktwerte für die Korrespondenzanalyse

Asymmetrische Darstellung

In der Korrespondenzanalyse können verschiedene Methoden zur Reskalierung der rechten und linken Singulärvektoren U und V verwendet werden. Ein häufig verwendetes Verfahren wird als „optimal scores", „standard scores" oder „standard coordinates" bezeichnet. Dieses Verfahren multipliziert die Werte der linken Singulärvektoren in U mit der Quadratwurzel des Kehrwertes seiner Zeilenanteile und entsprechend die rechten Singulärvektoren in V mit der Quadratwurzel des Kehrwertes seiner Spaltenanteile. Bezeichnet man die neuen Werte als $\tilde{u}_{ik}$ und $\tilde{v}_{jk}$, dann gilt entsprechend (1.17) für die Zeilenwerte und (1.18) für die Spaltenwerte.

$$\tilde{u}_{ik} = x_{ik} \sqrt{\frac{\sum_{j=1}^{h} \sum_{i=1}^{g} a_{ij}}{\sum_{j=1}^{h} a_{gi}}} \tag{1.17}$$

$$\tilde{v}_{jk} = y_{jk} \sqrt{\frac{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}}{\sum_{i=1}^{g} a_{ih}}} \tag{1.18}$$

Das Ergebnis sind zwei Matrizen, deren jeweilige Dimensionen einen gewichteten Mittelwert von 0 (1.19) und eine gewichtete Varianz von 1 (1.20) aufweisen.

$$\sum_{i=1}^{g} \tilde{u}_{ik} \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \sum_{j=1}^{h} \tilde{v}_{jk} \frac{\sum_{i=1}^{g} a_{ij}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = 0 \tag{1.19}$$

$$\sum_{i=1}^{g} \tilde{u}_{ik}^2 \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \sum_{j=1}^{h} \tilde{v}_{jk}^2 \frac{\sum_{i=1}^{g} a_{ij}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = 1 \tag{1.20}$$

Da bei dieser Methode die Varianz jeder Dimension gleich 1 ist, drücken die Dimensionen in den „standard coordinates", nicht die relative Bedeutung der Dimensionen aus (vgl. Weller and Romney, 1990).

Eine Alternative stellen die „principal scores" bzw „principal coordinates" dar. Bezeichnet man diese Werte als u_{ik} und v_{ik} , dann werden die Zeilen- und Spaltenkoordinaten entsprechend (1.21) für die Zeilenkoordinaten und (1.22) für die Spaltenkoordinaten berechnet.

$$u_{ik} = \sigma_k x_{ik} \sqrt{\frac{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}}{\sum_{j=1}^{h} a_{gj}}} \tag{1.21}$$

$$v_{jk} = \sigma_k y_{jk} \sqrt{\frac{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}}{\sum_{i=1}^{g} a_{ih}}} \tag{1.22}$$

Diese Werte haben nun in jeder Dimension einen gewichteten Mittelwert von 0 (1.23) und eine gewichtete Varianz der entsprechenden quadrierten Singulärwerte (1.24).

$$\sum_{i=1}^{g} u_{ik} \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \sum_{j=1}^{h} v_{jk} \frac{\sum_{i=1}^{g} a_{ij}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = 0 \tag{1.23}$$

$$\sum_{i=1}^{g} u_{ik}^2 \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \sum_{j=1}^{h} v_{jk}^2 \frac{\sum_{i=1}^{g} a_{ij}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \sigma_k^2 \tag{1.24}$$

Die Dimensionen in den „principal coordinates" spiegeln also ihre jeweilige Bedeutung entsprechend der in Σ angegebenen Singulärwerte wieder.

Die Distanzen in der graphischen Darstellung von Korrespondenzanalysen entsprechen den χ^2-Distanzen zwischen den Spalten- und Zeilenprofilen. Das Zeilenprofil für einen Wert ergibt sich aus der Division des Werte durch die entsprechende Zeilenrandhäufigkeit ($\frac{a_{ij}}{\sum_{j=1}^{h} a_{gj}}$). Entsprechend ergibt sich das Spaltenprofil durch die Division der Wertes durch die entsprechende Spaltenrandhäufigkeit ($\frac{a_{ij}}{\sum_{i=1}^{g} a_{ih}}$). Bezogen auf die Beispielmatrix ergeben sich damit die in Table 3 angegebenen Werte.

Table 3: Zeilen- und Spaltenprofile der Beispielmatrix

Zeilenprofile

	Merkmal 1	Merkmal 2	Merkmal 3	Summe
Haushalt 1	0,500	0,000	0,500	1,000
Haushalt 2	0,000	0,000	1,000	1,000
Haushalt 3	0,000	0,500	0,500	1,000
Haushalt 4	0,500	0,500	0,000	1,000
Masse	0,286	0,286	0,429	

Spaltenprofile

	Merkmal 1	Merkmal 2	Merkmal 3	Masse
Haushalt 1	0,500	0,000	0,333	0,286
Haushalt 2	0,000	0,000	0,333	0,143
Haushalt 3	0,000	0,500	0,333	0,286
Haushalt 4	0,500	0,500	0,000	0,286
Summe	1,000	1,000	1,000	

Die Profile der unterschiedlichen Zeilen und Spalten können dann miteinander verglichen werden um die Distanz zwischen verschiedenen Zeilen und Spalten zu ermitteln. Die gewichtete χ^2-Distanz zwischen den jeweiligen Zeilenprofilen i und j ergibt sich dabei aus (1.25) und entsprechend für die Spaltenprofile i und j aus (1.26).

$$\Delta(i,i') = \sqrt{\sum_{j=1}^{h} \frac{\left(\frac{a_{ij}}{\sum_{j=1}^{h} a_{gj}} - \frac{a_{i'j}}{\sum_{j=1}^{h} a_{g'j}} \right)}{\frac{\sum_{i=1}^{g} a_{ih}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}}}} \tag{1.25}$$

$$\Delta(j,j') = \sqrt{\sum_{i=1}^{g} \frac{\left(\frac{a_{ij}}{\sum_{i=1}^{g} a_{ih}} - \frac{a_{ij'}}{\sum_{i=1}^{g} a_{ih'}} \right)}{\frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}}}} \tag{1.26}$$

Ebenso können die χ^2-Distanzen der durchschnittlichen Zeilen- und Spaltenprofile analysiert werden. Die durchschnittlichen Zeilenprofile, oder auch (Zeilen-)Masse genannt, ergeben sich dabei aus (1.28) und entsprechend ergeben sich die durchschnittlichen Spaltenprofile aus (1.27).

$$i^{+} = \left(\frac{\sum_{i=1}^{g} a_{ih}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} \right) \tag{1.27}$$

$$j^{+} = \left(\frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} \right) \tag{1.28}$$

Die χ^2-Distanz einer Zeile bzw. Spalte und dem durchschnittlichem Zeilen- bzw. Spaltenprofil ergibt sich dann aus

$$\Delta(i,i^+) = \sqrt{\sum_{j=1}^{h} \frac{\left(\frac{a_{ij}}{\sum_{j=1}^{h} a_{gj}} - \frac{\sum_{i=1}^{g} a_{ih}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} \right)^2}{\frac{\sum_{i=1}^{g} a_{ih}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}}}} \tag{1.29}$$

und

$$\Delta(j,j^+) = \sqrt{\sum_{i=1}^{g} \frac{\left(\frac{a_{ij}}{\sum_{i=1}^{g} a_{ih}} - \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} \right)^2}{\frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}}}} \tag{1.30}$$

Für eine ausführlichere Darstellung der Distanzberechnungen und der daraus folgenden sei an dieser Stelle auf Faust (2005), Carroll et al. (1986) und Greenacre (1984) verwiesen.

Für eine graphische Darstellung der Ergebnisse und deren Auswertung bestehen grundsätzlich zwei Möglichkeiten. Zum einen kann man die Zeilenkoordinaten u_{ik} zusammen mit den Spaltenkoordinaten $\tilde{v}_{ik}$ darstellen oder die Spaltenkoordinaten v_{ik} mit den Zeilenkoordinaten $\tilde{u}_{ik}$. Beide Ansätze haben jedoch den Nachteil, dass sie nur eine asymmetrische Beschreibung der Matrix sind. Welche der beiden Kombinationsmöglichkeiten gewählt wird, hängt entscheidend von der Fragestellung und dem damit verbundenen Fokus auf eine entsprechende Repräsentation ab. Dabei muss unbedingt berücksichtigt werden, dass die jeweiligen Punkte nur innerhalb der gleichen Dimension verglichen werden können, nicht aber die Position zweier Punkte unterschiedlicher Dimensionen (vgl. Faust, 2005, S. 128ff). Sollen jedoch die berechneten Koordinaten die Grundlage für eine Clusteranalyse bilden, kann keine der beiden Skalierungen verwendet werden, da bei einer Clusteranalyse

in einem >1-Dimensionalen Raum die interdimensionale Lage der Werte für die Berechnung der Distanzen verwendet wird.

Symmetrische Darstellung

Um das Problem der Interkoordinateninterpretation zu lösen, entwickelten Carroll et al. (1986) eine symmetrische Projektion der Matrix D. Dabei betrachten sie das Problem als eine „multiple correspondence analysis"(vgl. Carroll et al., 1986, S. 274ff). Für diesen Ansatz wird die ursprüngliche Kontingentstabelle in eine „Pseudo-Kontingenztabelle" transformiert. Hierzu wird die ursprüngliche Kontingenztabelle der Matrix D mit h Spalten, gZeilen und N Beobachtungen so transformiert, dass jede Zeile der neuen Tabelle eine Beobachtung der originalen Tabelle entspricht und die Spalten die jeweiligen Variablen. Diese neue „Pseudo-Kontingenztabelle" hat N Zeilen und $g+h$ Spalten (sieh Table 4).

Table 4: Pseudo-Kontingenztabelle der Beispielmatrix

HH 1	HH 2	HH 3	HH 4	I 1	I 2	I3
1	0	0	0	1	0	0
1	0	0	0	0	0	1
0	1	0	0	0	0	1
0	0	1	0	0	1	0
0	0	1	0	0	0	1
0	0	0	1	1	0	0
0	0	0	1	0	1	0

Zur besseren Unterscheidung wird im Folgenden die Pseudo-Kontingenztabelle als F bezeichnet. Diese hat $g+h$ Spalten und so viele Zeilen wie Einträge in D gleich 1 sind, also $\Sigma_{j=1}^{h}\Sigma_{i=1}^{g}a_{ij}$. Die Spaltensummen entsprechen den ursprünglichen Zeilen- und Spalten-Randhäufigkeiten. Eine Korrespondenzanalyse von F ergibt wiederum Werte sowohl für die Spalten, als auch für die Zeilen. Dabei haben die Zeilenwerte jedoch keine Bedeutung mehr. Wie auch bei den asymmetrischen Darstellungen, entsprechen die Werte der ursprünglichen Spaltenvariablen (h) der χ^2-Distanz zwischen der Spaltenprofilen. Da die Randhäufigkeiten der „Pseudo-Kontingenztabelle" der Anzahl der Variablen entsprechen, in der Beispielmatrix zwei, und die Anzahl der Zeilen von F gleich $\Sigma_{j=1}^{h}\Sigma_{i=1}^{g}a_{ij}$ sind, ist die Anzahl der 1sen in der Beispielmatrix entsprechend $2*\Sigma_{j=1}^{h}\Sigma_{i=1}^{g}a_{ij}$ und die Zeilenrandhäufigkeiten entsprechend (1.31).

$$\frac{\sum_{j=1}^{h} f_{gi}}{\sum_{j=1}^{h}\sum_{i=1}^{g} f_{ij}} = \frac{1}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} \tag{1.31}$$

Daraus resultiert, dass die χ^2-Distanz zwischen zwei Spaltenprofilen (1.32) entspricht.

$$\Delta(j,j') = \sqrt{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij} \sum_{i=1}^{N} \left(\frac{f_{ij}}{\sum_{i=1}^{g} f_{ih}} - \frac{f_{ij'}}{\sum_{i=1}^{g} f_{ih'}} \right)} \tag{1.32}$$

Der Ansatz von Carroll, Green und Schaffer (1986, S. 275f) wenden die Korrespondenzanalyse auf die Matrix F an (1.33).

$$R^{-\frac{1}{2}} F C^{-\frac{1}{2}} = U \Sigma \mathrm{V} \tag{1.33}$$

Die Werte der Carroll, Green und Schaffer Reskalierung, die im Folgenden als $\ddot{u}_{ik}$ und $\ddot{v}_{jk}$ bezeichnet werden, sind als „principal coordinates" skaliert. Da in der „Pseudo-Matrix" sowohl die Haushalte als auch die Infrastrukturmerkmale die Spalten bilden, sind die Distanzen zwischen Koordinaten unterschiedlicher Spalten als χ^2-Distanz interpretierbar (1.32).

Die Carroll, Green und Schaffer Koordinaten lassen sich mit (1.34) und (1.35) mit den „principal coordinates" und den „standard coordinates" von D in Beziehung setzen.

$$\ddot{u}_{ik} = \tilde{u}_{ji} \sqrt{\frac{(1+\sigma_k)}{2}} = \frac{u_{ik}}{\sigma_k} \sqrt{\frac{(1+\sigma_k)}{2}} \tag{1.34}$$

$$\ddot{v}_{jk} = \tilde{v}_{ji} \sqrt{\frac{(1+\sigma_k)}{2}} = \frac{v_{ik}}{\sigma_k} \sqrt{\frac{(1+\sigma_k)}{2}} \tag{1.35}$$

Die Singulärwerte $\ddot{\sigma}_k$ der Carrol, Green, Schaffer Skalierung stehen dann entsprechend (1.36) mit den Singulärwerten der ursprünglichen Kontingenztabelle in Beziehung.

$$\sigma_k^2 = \sqrt{\frac{\sigma_k + 1}{2}} \tag{1.36}$$

Diese Koordinaten haben in den jeweiligen Dimensionen ebenfalls einen gewichteten Mittelwert von 0 (1.37).

$$\sum_{i=1}^{g} \ddot{u}_{ik} \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \sum_{i=1}^{h} \ddot{v}_{jk} \frac{\sum_{i=1}^{g} a_{ih}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = 0 \tag{1.37}$$

Die gewichtete Varianz der jeweiligen Dimension entspricht der Quadratwurzel der Singulärwerte der normalisierten Kontingenztabelle. Für ein Array mit zwei Variablen, zeigen Carroll et al. (1986), dass dies $\frac{\sigma_k + 1}{2}$ entspricht (1.38).

$$\sum_{i=1}^{g} \ddot{u}_{ik}^2 \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \sum_{i=1}^{h} \ddot{v}_{jk}^2 \frac{\sum_{i=1}^{g} a_{ih}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} = \frac{\sigma_k + 1}{2} \tag{1.38}$$

Die Distanzen der Elemente zueinander kann nun symmetrisch interpretiert werden. Diese Symmetrie hat jedoch den Nachteil, dass nun die χ^2-Distanzen durch die Randverteilung der Spalten- und Zeilensummen der ursprünglichen Matrix D determiniert werden (Carroll et al., 1986). Greenacre (1989) machte darauf aufmerksam, dass der Umstand der absoluten Abhängigkeit der Distanzen von den Randhäufigkeiten diese Skalierung in der Auswertung sehr limitiert ist. Trotzdem ist dieser Ansatz, wenn die Möglichkeit der Interpretation zwischen den Dimensionen, insbesondere bei Joint Displays, gegeben sein soll, eine mögliche Herangehensweise (Carroll et al.,1989 und Faust, 2005).

Skaliert man die Werte aus der SVD entsprechend der Carrol, Green, Schaffer Skalierung, so lassen sich die Haushalte und die Infrastrukturmerkmale in einer gemeinsamen Graphik (Figure 2) darstellen.

Dabei repräsentieren die gelben Punkte die befragten Haushalte und die roten Punkte die Infrastrukturmerkmale. Die Nummer bezeichnet jeweils das entsprechende Infrastrukturmerkmal. Es ist deutlich zu erkennen, dass sich die befragten Haushalte in der Figure 1 insbesondere durch das Merkmal Ausländeranteil in der Nachbarschaft (11) ausdifferenzieren.

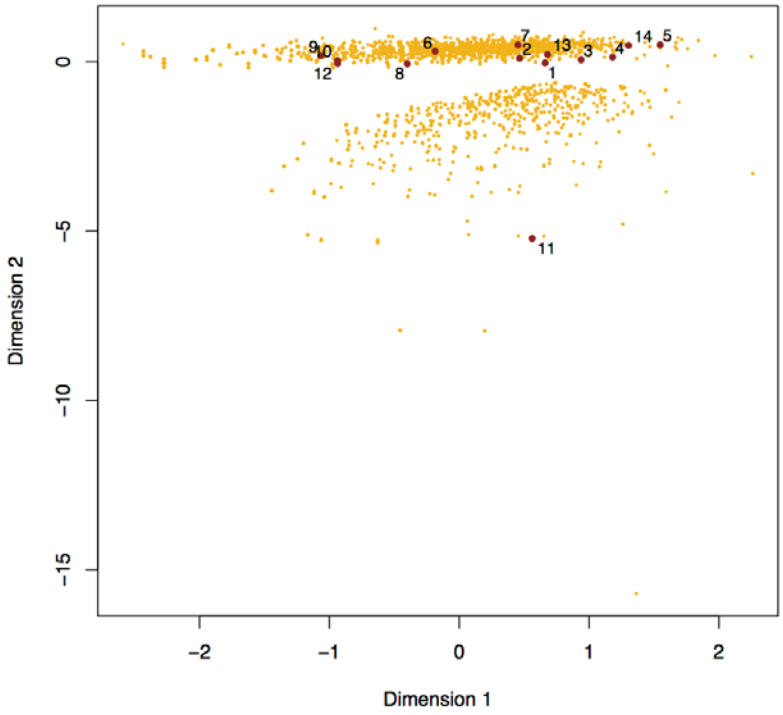

Figure 2: Joint Display von Haushalten und Infrastrukturmerkmalen von D, Eigene Berechnung

Inertia

Die Inertia einer Matrix, auch Trägheit genannt, bezeichnet in der Korresondenzanalyse den Anteil der Variation in den Daten(vgl. Greenacre, 1984). Für die Profile der beschriebenen Matrix D errechnet sich die Gesamtinertia aus der χ^2-Distanz eines Zeilen- oder Spaltenprofils zum durchschnittlichen Zeilen- oder Spaltenprofil (1.39) und (1.40). Diese wird mit der relativen Häufigkeit mit der ein Profil auftritt, gewichtet.

$$\sum_{i=1}^{g} \frac{\sum_{j=1}^{h} a_{gj}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} \delta(i,i^+)^2 \tag{1.39}$$

$$\sum_{j=1}^{h} \frac{\sum_{i=1}^{g} a_{ij}}{\sum_{j=1}^{h}\sum_{i=1}^{g} a_{ij}} \delta(j,j^+)^2 \tag{1.40}$$

Die Gesamtinertia einer Matrix kann auch aus der Summe der quadrierten Singulärwerte berechnet werden.

$$\sum_{k=1}^{r} \sigma_k^2 \tag{1.41}$$

3. Datenvorbereitung

Die verwendeten Daten stammen aus dem Sozio-oekonomischen Panel (SOEP) aus der Erhebungswelle von 2004. Das SOEP ist eine wissenschaftliche Längsschnittstudie, die seit 1984 repräsentativ ausgewählte Haushalte in Deutschland und darin lebende Personen einmal im Jahr statistisch erfasst und alle Personen ab einem Alter von 17 Jahren persönlich befragt. Die Fragebatterien werden dabei nicht von amtlichen und politischen Fragestellungen bestimmt, sondern von theoriegeleiteten wissenschaftlichen Fragen der Politikberatung (vgl. Wagner et al., 2007, S. 302ff). Das SOEP ist auch Teil einer weltweiten Forschungs-Infrastruktur, die für empirische Analysen von Personen und Haushalten Längsschnittsdaten zur Verfügung stellt (vgl. Butz and Torrey, 2006, Frick et al., 2008).

Für die Analyse wurden alle Datensätze ausgefiltert, die keine Haushaltsnummer besaßen (HHNR > 0) da diese nicht eindeutig einem Milieu zuordenbar waren. Weiterhin wurden alle Personen ausgefiltert, die zum Erhebungszeitpunkt (2004) jünger als 18 Jahre alt waren, da diese in der Regel über keinen eigenständigen Haushalt verfügen (GEBJAHR < 1987). Da die zugespielten Milieuinformationen auf deutschen SINUS-Milieus beruhen, konnten nur die Befragten/Haushalte berücksichtigt werden, die deutsche Staatsbürger sind (UPNAT == 1). Auf Grund des Untersuchungsgegenstandes, wurden nur die Haushalte ausgewählt, die in Städten mit mindestens 5 000 Einwohnern leben (gk_reg_a < 6). Um die Wohnstandortbedürfnisse zu analysieren, wurden darüber hinaus in dieser Analyse nur die Personen berücksichtigt, die mit ihrer Wohnumgebung sehr zufrieden sind (UP0107 > 7).

Table 5: Reduktion des Datensatzes durch die Regeln

Regel	Reduktion	Kummuliert	Anzahl
			28329
$\mathrm{HHNR} > 0$		0	28329
$\mathrm{GEBJAHR} < 1987$	-5697	-5697	22632
$\mathrm{UPNAT} == 1$	-4576	-11955	16374
$gk_reg_a < 6$	-5187	-17142	11187

Dies ergibt sich aus der Annahme, dass Personen nach Möglichkeit in die Nachbarschaften ziehen, die ihren Ansprüchen am ehesten entsprechen. Sind diese nun sehr zufrieden mit ihrer Umgebung, kann davon ausgegangen werden, dass ihre Ansprüche erfüllt sind. Das Ergebnis ist eine Stichprobe D aus der Originalstichprobe (n=21.890) von n=11.187 dies sind 51,11% aller Befragten. Die Befragten Personen lebten dabei in 9.227 Haushalten.

Dies sind im Durchschnitt 2,37 Personen pro Haushalt.[6] Für die Nachbarschaftsanalyse wurden die in Table 6 genannten Variablen herangezogen.

Table 6: Variablenübersicht

Variable	Nummer	Beschreibung
UH6401	1	Entfernung zu Fuß zu Geschäften
UH6402	2	Entfernung zu Fuß zu Gaststätten
UH6403	3	Entfernung zu Fuß zur Bank
UH6404	4	Entfernung zu Fuß zum Hausarzt
UH6405	13	Entfernung zu Fuß zum Kindergarten
UH6406	13	Entfernung zu Fuß zur Grundschule
UH6407	14	Entfernung zu Fuß zum Gymnasium
UH6408	14	Entfernung zu Fuß zum Treffpunkt für Jugendliche
UH6409	5	Entfernung zu Fuß zur Einrichtung für Ältere
UH6410	6	Entfernung zu Fuß zu Grünanlagen
UH6411	7	Entfernung zu Fuß zu Sportstätten
UH6412	8	Entfernung zu Fuß zu öffentlichen Verkehrsmitteln
UH6501	12	Beeinträchtigung durch Lärmbelastung
UH6502	12	Beeinträchtigung durch Luftverschmutzung
UH6503	9	Beeinträchtigung durch Mangel an Grünflächen
UH66	10	Kriminalität im Wohngebiet
UH68	11	Leben ausländische Familien im Wohngebiet

Bei den Daten des SOEP handelt es sich um ordinalskalierte Variablen. Da die SVD mit diesem Datentyp nicht sinnvoll operieren kann, wurden die vorliegenden Daten in eine Binärmatrix umgewandelt, so das ein Vektor der Form

$$R_i = (I_1, \rightleftharpoons, I_n) \tag{1.42}$$

für jeden Befragten vorliegt. Dabei wurde angenommen, dass eine Nähe zu einem Infrastrukturmerkmal I (UH64xx) dann vorhanden ist, wenn dies in weniger als 10 Minuten zu Fuß erreichbar ist und entsprechend mit 1 kodiert. Weiterhin wurden die Variablen UH6405 und UH6406 sowie UH6407 und UH6408 zusammengefasst, da es sich hierbei um Infrastruktureinheiten handelt die für die gleiche Zielgruppe (Kinder bzw. Jugendliche) kodieren. Darüber hinaus wurden die Umweltvariablen UH6501 und UH6502 - die eine hohe Lärmbelastung und Luftverschmutzung kodieren - als „Umgebungsverschmutzung" zusammengefasst und deren Negation („gar nicht") mit 1 kodiert. Bei der Variable UH67 wurden

6 Nach Angaben des Statistischen Bundesamtes von 2002 belief sich die durchschnittliche Haushaltsgröße in Osten und Westen im Jahr 200 auf 2,2 Personen pro Privathaushalt.

die Befragten mit 1 kodiert, die angaben sich mindestens „ziemlich sicher" in ihrem Quartier zu fühlen. Die Variable UH68 wurde mit 1 kodiert, wenn es sich um eine multikulturelle Umgebung handelt. Als Ergebnis erhalten wir eine Binärmatrix der Form

$$D = \begin{bmatrix} 1 & 0 & 1 & 1 & 0 & \cdots & a_{j1} \\ 0 & 0 & 0 & 1 & 1 & \cdots & a_{j2} \\ 1 & 1 & 0 & 1 & 0 & \cdots & a_{j3} \\ 1 & 0 & 0 & 0 & 1 & \cdots & a_{j4} \\ 0 & 1 & 1 & 0 & 0 & \cdots & a_{j5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ a_{i1} & a_{i2} & a_{i3} & a_{i4} & a_{i5} & & a_{ij} \end{bmatrix} \tag{1.43}$$

Diese Matrix kann auch als eine zweidimensionale Soziomatrix verstanden werden. Dabei werden die Haushalte als H mit g Haushalten und die Infrastrukturmerkmale als I mit h Infrastrukturmerkmalen bezeichnet. Im Folgenden wird diese Matrix als D bezeichnet mit den Einträgen a_{ij} . Dabei gilt $a_{ij} = 1$ wenn ein Haushalt i in der Nähe des Infrastrukturmerkmals j zu finden ist und $a_{ij} = 0$ wenn nicht. Jede Zeile H_i repräsentiert also einen Befragten B und jede Spalte I_j repräsentiert ein Infrastrukturmerkmal I.

Definition Infrastrukturmerkmal. Unter Infrastrukturmerkmalen werden die Variablen aus Table 6 verstanden.

Definition Umgebung. Eine Umgebung E ist definiert als eine spezifische Kombination von I. $E = \{I_1, \ldots, I_j\}$

3.1. Beschreibung der Datengrundlage

Wie oben dargestellt, beträgt der Gesamtstichprobenumfang 21.890 Befragte in 9.227 Haushalten. Von dieser Stichprobe wurden, wie dargestellt, die Befragten ausgewählt, die zum Zeitpunkt der Erhebung mindestens 18 Jahre, deutsche Staatsangehörige sind und in Städten mit mindestens 5.000 Einwohnern wohnten. Der Umfang dieser Auswahl beträgt 11.187 Personen in 5.982 Haushalten und damit 51,11% der insgesamt Befragten bzw. 64,83% der befragten Haushalte. In den Haushalten der Auswahl lebten im Durchschnitt 1,87 Personen.

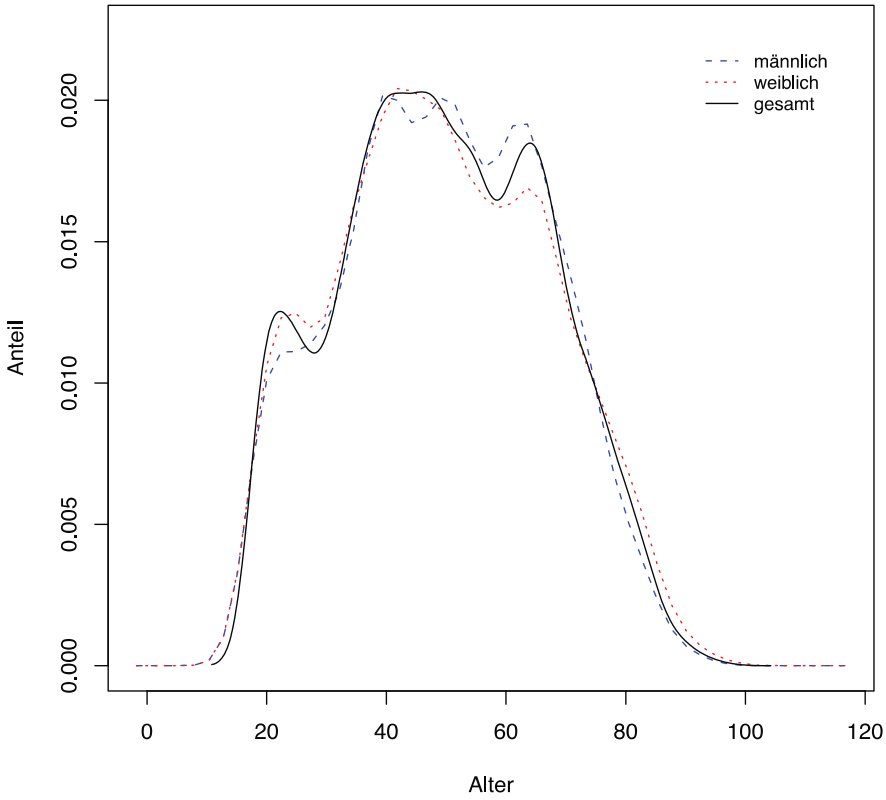

Figure 3: Altersverteilung der Befragten, SOEP, eigene Berechnung

Die Auswahl (n=11.187) besteht zu 47% aus männlichen und 53% weiblichen Befragten und entspricht damit der Verteilung aller Befragten sowie der gesamtdeutschen Bevölkerung. Das Alter der Befragten in der Auswahl reicht von 18 bis 97 Jahren. Dabei liegt das mittlere Alter 49,42 Jahren bzw. 49 Jahre bei einer Standardabweichung von 17,18. In der Altersverteilung sind beide Geschlechter nahe zu gleich repräsentiert (siehe Figure 3). Vergleicht man die Altersverteilung mit der aller Befragten, so kann gezeigt werden, dass die Verteilung in der Auswahl mit der aller Befragten weitestgehend übereinstimmt. So weist die Gesamterhebung eine Spannweite von 16 bis 99 Jahren bei einem durchschnittlichen Alter von 47,15 Jahren bzw. 46 Jahre auf.

Das durchschnittliche monatliche Nettoeinkommen der Befragten[7] (n=5.648 Netto und n=5 472 Brutto) liegt bei 1.844,87 EUR (Brutto 2.887,38 EUR) bzw. 1.500 EUR (2 400 EUR) bei einer Standardabweichung von 1.484,04 (2.575,74). Lässt man die 10 % an den äußeren Seiten unberücksichtigt, ergibt sich ein durchschnittliches monatliches Nettoeinkommen von 1.631,74 EUR (2.534,63 EUR). Dabei liegt das niedrigste monatliche Netto-Einkommen bei 295 EUR[8] und einem maximalen monatlichem Netto-Einkommen von 30.000 EUR (60.000 EUR).

7 Es wird stets das ungewichtete Haushaltseinkommen angegeben.

8 Das entspricht dem Mindestregelsatz der Sozialhilfe im Westen 2004.

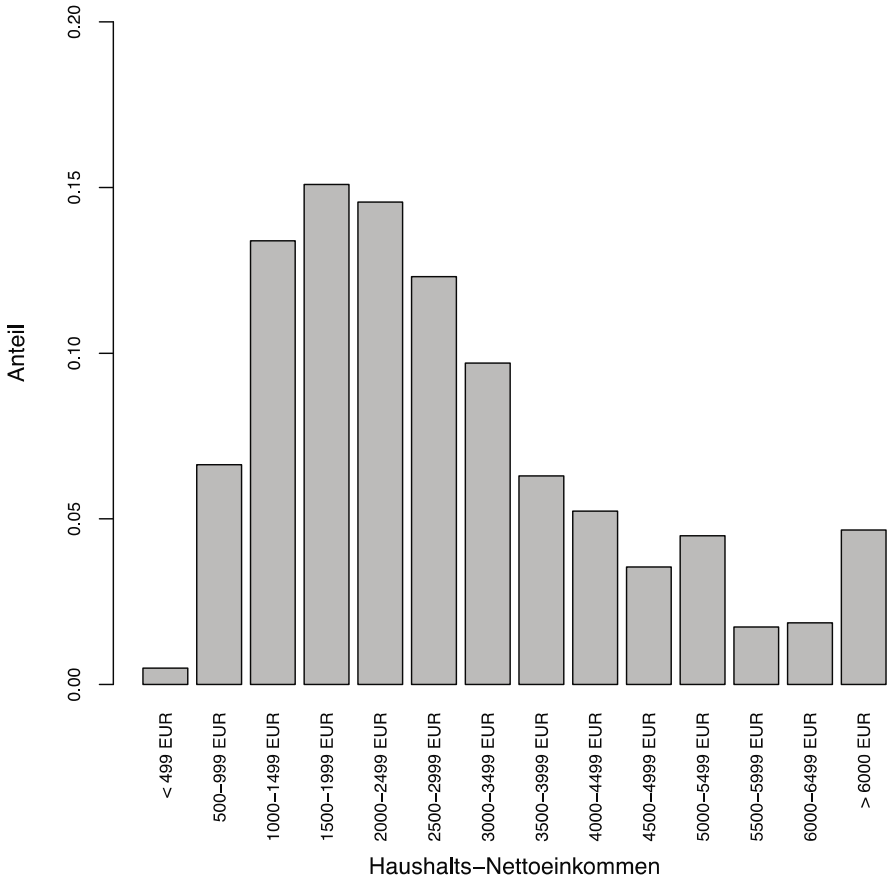

Figure 4: Haushaltseinkommen der Auswahl, SOEP, eigene Berechnung

In der gesamten Erhebung (n=10.866 Netto und n=10.491 Brutto) lag das durchschnittliche Nettoeinkommen der Befragten bei 1.678,15 EUR (Brutto 2.627,10 EUR) bzw. 1.400 EUR (2.200 EUR) bei einer Standardabweichung von 1.326,31 (2.285,5). Lässt man auch hier die 10% an den äußeren Seiten unberücksichtigt, ergibt sich ein durchschnittliches monatliches Nettoeinkommen von 1.489,2 EUR (2.317,79 EUR). Dabei liegt das niedrigste monatliche Netto-Einkommen bei 291 EUR und einem maximalen monatlichem Netto-Einkommen von 30.000 EUR (60.000 EUR).

Das durchschnittliche Haushalts-Nettoeinkommen der Befragten in der Auswahl (n=5.639) lag bei 2.908,39 EUR bzw. 2.475 EUR bei einer Standardabweichung von 2.317,89. Werden die äußeren 10% nicht berücksichtigt, ergibt sich ein durchschnittliches Haushalts-Nettoeinkommen von 2606,37 EUR. Das durchschnittliche Haushalts-Nettoeinkommen in Deutschland lag 2003 gemäß dem Statistischen Bundesamt bei 2.770 EUR.[9]

42% der Haushalte (n=5.919) wohnen in Mehrfamilienhäusern mit 3 oder mehr Wohnungen. Die meisten befragten Haushalte bewohnten dabei ein Wohnhaus mit 5-8 Mietpartei-

9 Für eine detaillierte Beschreibung der Einkommensstruktur siehe Sachverständigenrat (2006), Statistisches Bundesamt (2006, 2007).

en (17%) und nur 1% der Befragten in der Auswahl wohnen Hochhäusern. 36% der Haushalte in der Auswahl leben in einem freistehenden 1-2 Familien Haus und 21% in einem Reihenhaus.

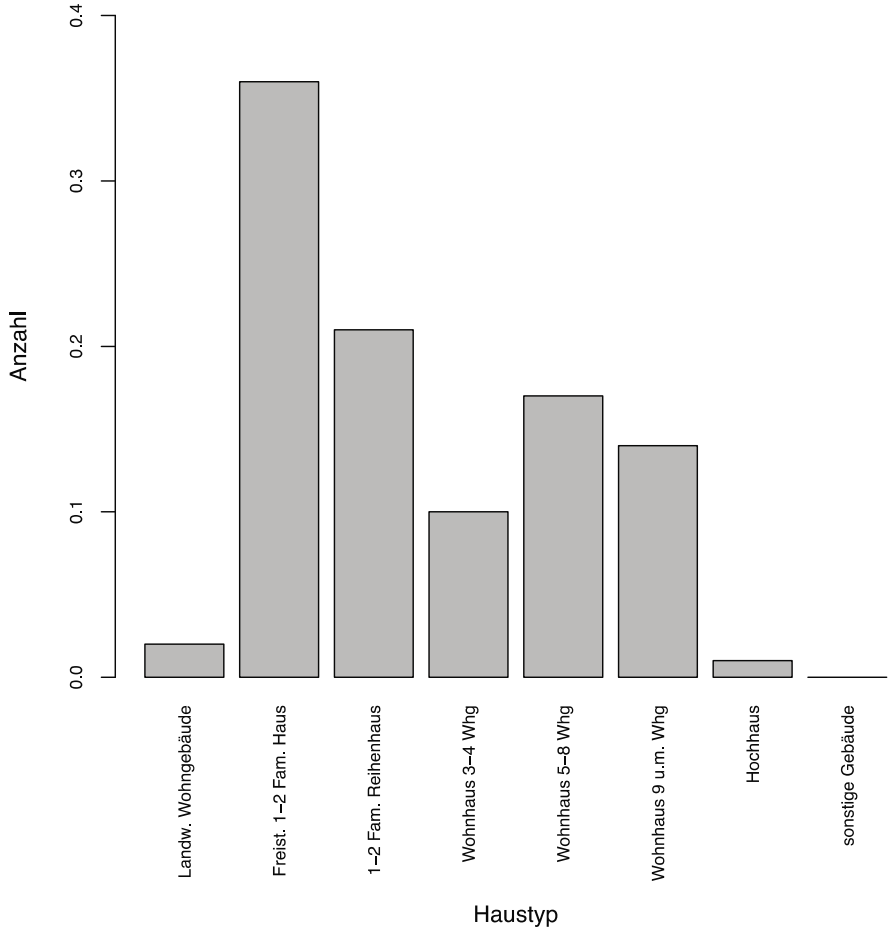

Figure 5: Wohnverhältnisse der befragten Haushalte, SOEP, eigene Berechnung

Dies entspricht auch der Gesamterhebung des SOEP. Hier wohnen 43% in Mehrfamilienhäusern und 53% in 1-2 Familienhäusern. Heimbewohner kommen in der Auswahl nicht vor. In 43% der Fälle waren die befragten (n=5.982) Mieter oder Untermieter der bewohnten Wohnung. 57% der Befragten gaben an, dass die bewohnte Immobilie ihr Eigentum ist. In der Gesamterhebung lag die Quote der Mieter bei 48% und die der Eigentümer bei 53%. Damit sind die Eigentümer im Vergleich zur Gesamterhebung leicht überpräsentiert. Nach Angaben des GdW (2004) lag die Mieterquote 2001 in der Bundesrepublik bei 57,4%. Damit sind sowohl im SOEP als auch in der Auswahl die Eigentümer deutlich überrepräsentiert. Interessant ist, dass es keinen Zusammenhang zwischen dem Nettoeinkommen und dem bewohnten Gebäudetyp (Cramer V: 0,125) als auch dem Eigentum von Wohnraum (Cramer V: 0,225) gibt. Fast die Hälfte der Bewohner eines 1-2 Familienhauses verfügen über ein Haushaltsnettoeinkommen zwischen 1.500 EUR 3.500 EUR. Gleiches gilt

auch für Eigentümer eines Wohnraums. Die meisten Eigentümer sind dabei Paare ohne Kinder (40 %) oder mit einem Kind (35%).

39% der Haushalte (n=5.982) in der Auswahl haben Kinder. In 4% der Fälle wachsen die Kinder in alleinerziehenden Strukturen auf. 21% der befragten Haushalte sind 1 Personenhaushalte. Dies sind deutlich weniger, als im bundesdeutschen Durchschnitt (37%). 38% der Haushalte sind Paarhaushalte ohne Kinder. Damit unterscheidet sich diese Variable nur geringfügig bei den 1 Personenhaushalten (24%) und den Paarhaushalten (35%) von der Verteilung der Gesamtstichprobe.

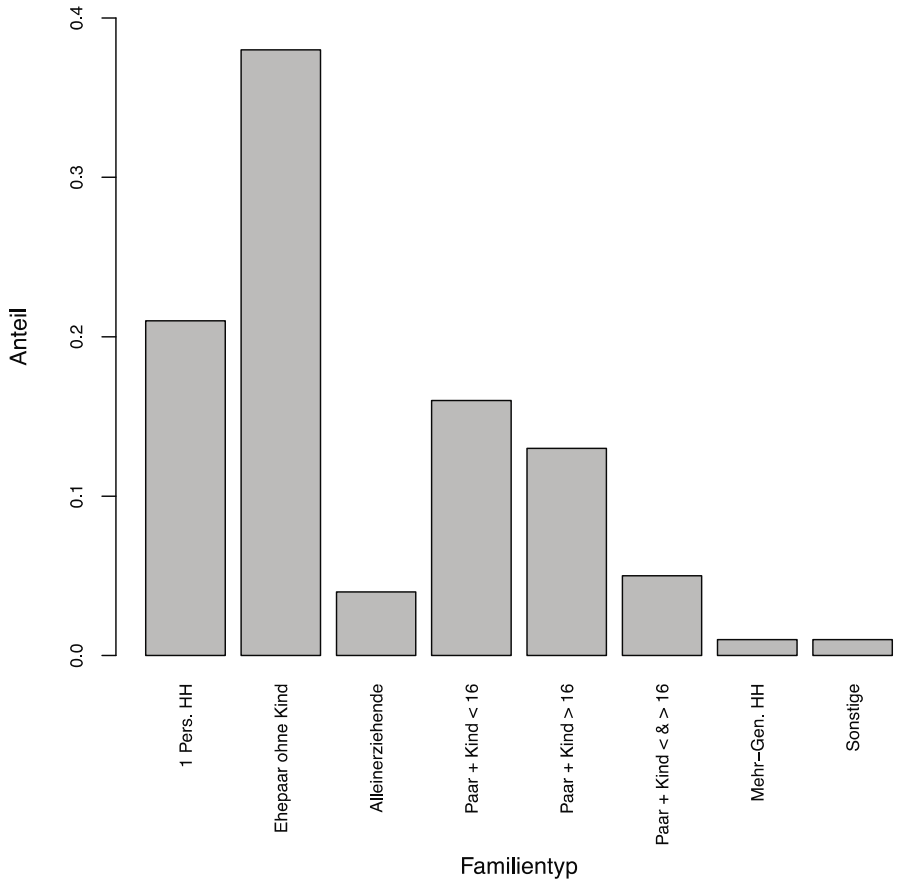

Figure 6: Familienstrukturen bei den befragten Haushalten, SOEP, eigene Berechnung

14% der Befragten, verfügen weder über eine Ausbildung, noch sind sind sie zur Zeit in Ausbildung. 4% der Befragten befinden sich derzeit noch in einer Ausbildung. Die Hälfte der in Ausbildung begriffenen Personen streben die Gesellenprüfung an, während die andere ein Studium absolviert. Bei 48% der Befragten ist eine abgeschlossene Ausbildung der höchste Bildungsstand. Während 23% der Befragten über eine akademische Ausbildung verfügen. Mit 16% haben dabei mehr Personen einen Universitätsabschluß als einen Fachhochschulabschluß (7%). In 2% der Fälle sind die Befragten verbeamtet und in 5% der Fälle wurde eine Meisterprüfung abgelegt.

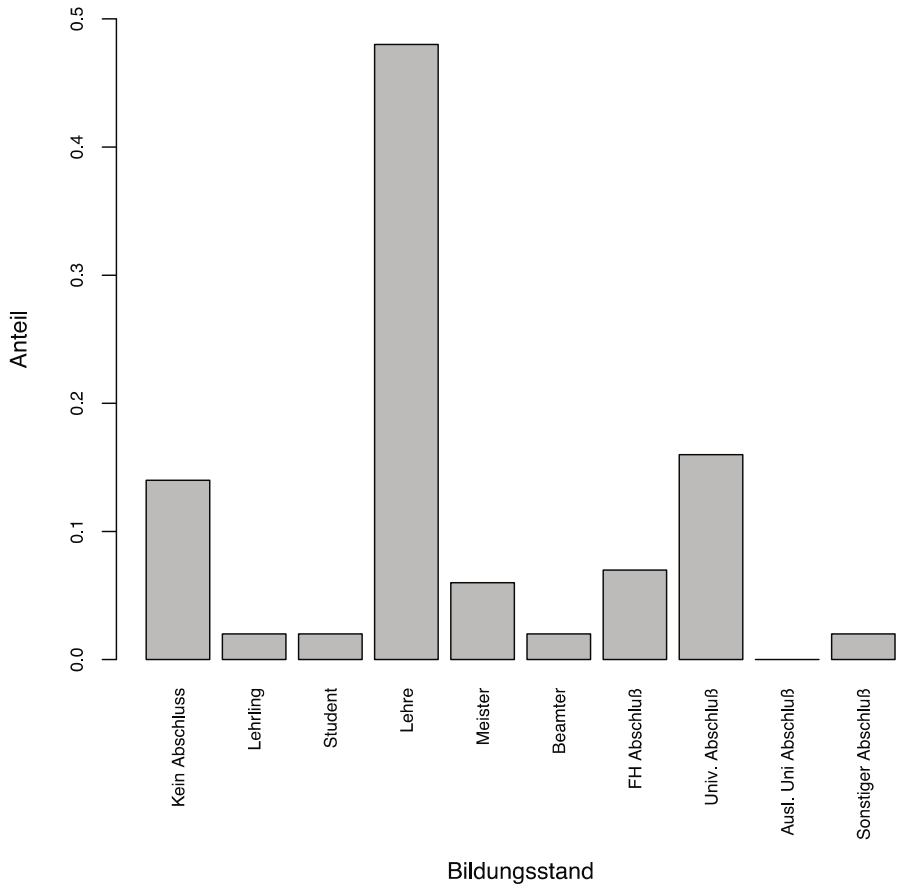

Figure 7: Bildungsstand bei den Befragten, SOEP, eigene Berechnung

3.2. Milieuverteilungen

Die Miliverteilung ist seit der letzten großen Anpassung der Milieus im Jahr 2000 weitestgehend konstant (siehe Figure 8). Die Schwankungen in den einzelnen Jahren lässt sich auf Stichprobenschwankungen zurückführen. Lediglich in den Jahren 2002 und 2003 gab es größere Abweichungen im Bereich der gesellschaftlichen Leitmilieus. Dabei ist jedoch zu beachten, dass die einzelnen Jahre nicht direkt miteinander verglichen werden können, da sich die Algorithmen für die Zuordnung zu einem spezifischen Milieu stetig ändern[10].

Die Milieuverteilung im SOEP weicht im Jahr 2004 bei den meisten Milieus um maximal 0 bis (+/-) 2 % von der vom SINUS-Institut angegebenen Milieuverteilung in Deutschland ab. Lediglich das experimentalistische Milieu ist mit -6 % im Gegensatz zur gesamtdeutschen Milieuverteilung Unterrepräsentiert (siehe Figure 10).

10 Erst ab 2010 soll es für die Zukunft eine vergleichbare Milieuverteilung über die Jahre geben.

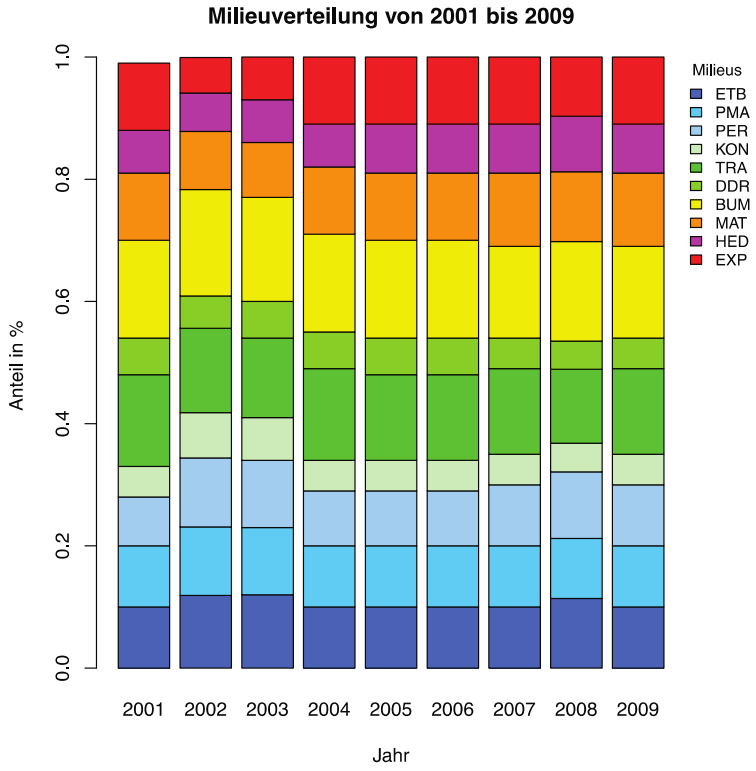

Figure 8: Milieuverteilung in Deutschland von 2001 bis 2009 (In 2001 ist die Summe nur 99%!), eigene Berechnungen auf Basis der SINUS Reports

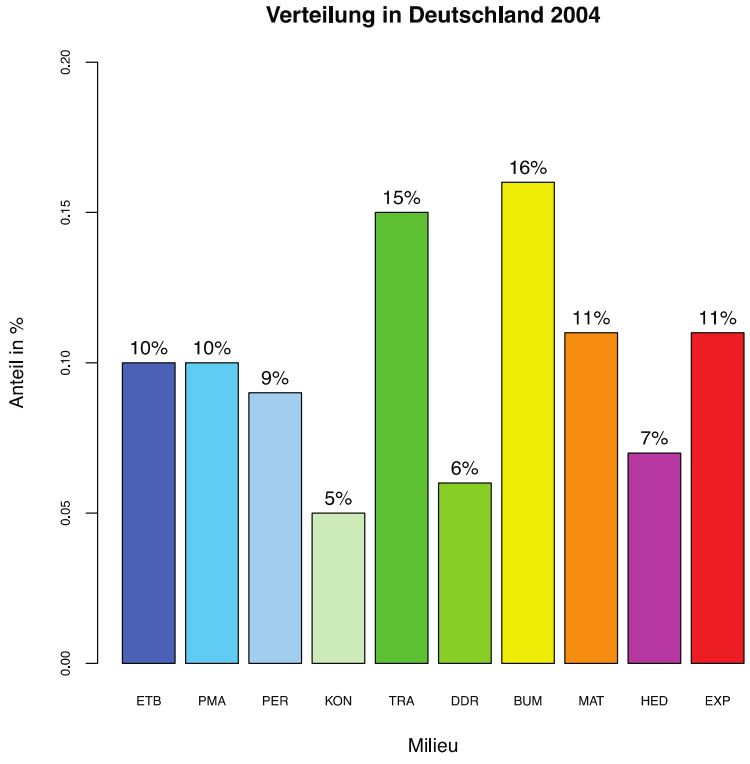

Figure 9: Milieuverteilung in Deutschland 2004, SINUS Institut, eigene Berechnung

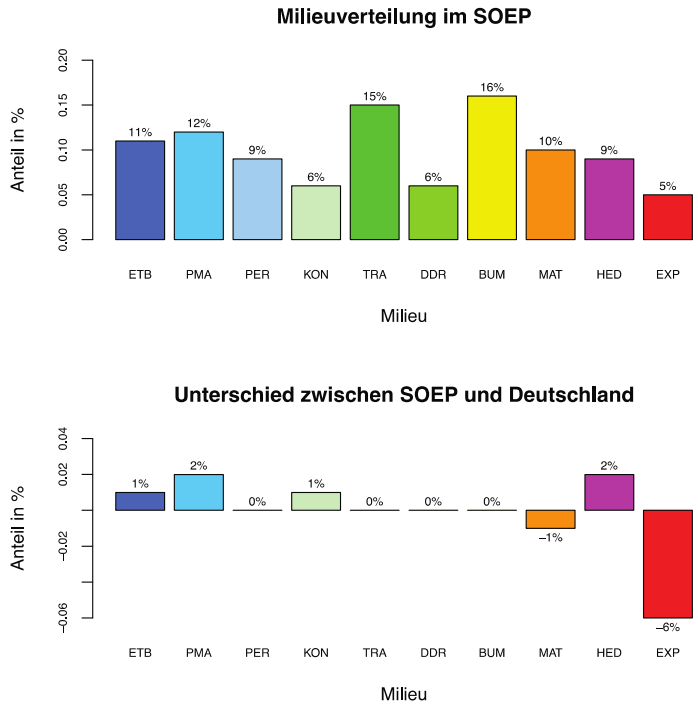

Figure 10: Milieuverteilungen im Vergleich Deutschland und dem SOEP, SINUS Institut Report, SOEP, eigene Berechnung

Auch die Milieuverteilung der Auswahl weicht von der Deutschlands ab (siehe Figure 11). Hier liegt die Abweichung bei den meisten Milieus bei maximal (+/-) 1%. Lediglich die Postmaterialisten (PMA) (+4%) weichen leicht und die Experimentellen (EXP) (-7%) deutlich hiervon ab. Vergleicht man die Verteilungen der Auswahl und der des SOEP, so beträgt die Abweichung der meisten Milieus maximal (+/-) 2%.

Diese leichten Abweichungen waren zu erwarten, da das DIW keine Milieu repräsentative Studie erhebt. Insgesamt entsprechen die dargestellten Verteilungen der Verteilung der deutschen Bevölkerung wie sie vom SINUS-Institut für 2004 angegeben wurde. Die Auswahl aus der Studie kann daher als hinreichend repräsentativ für die nachfolgenden Analysen angesehen werden.

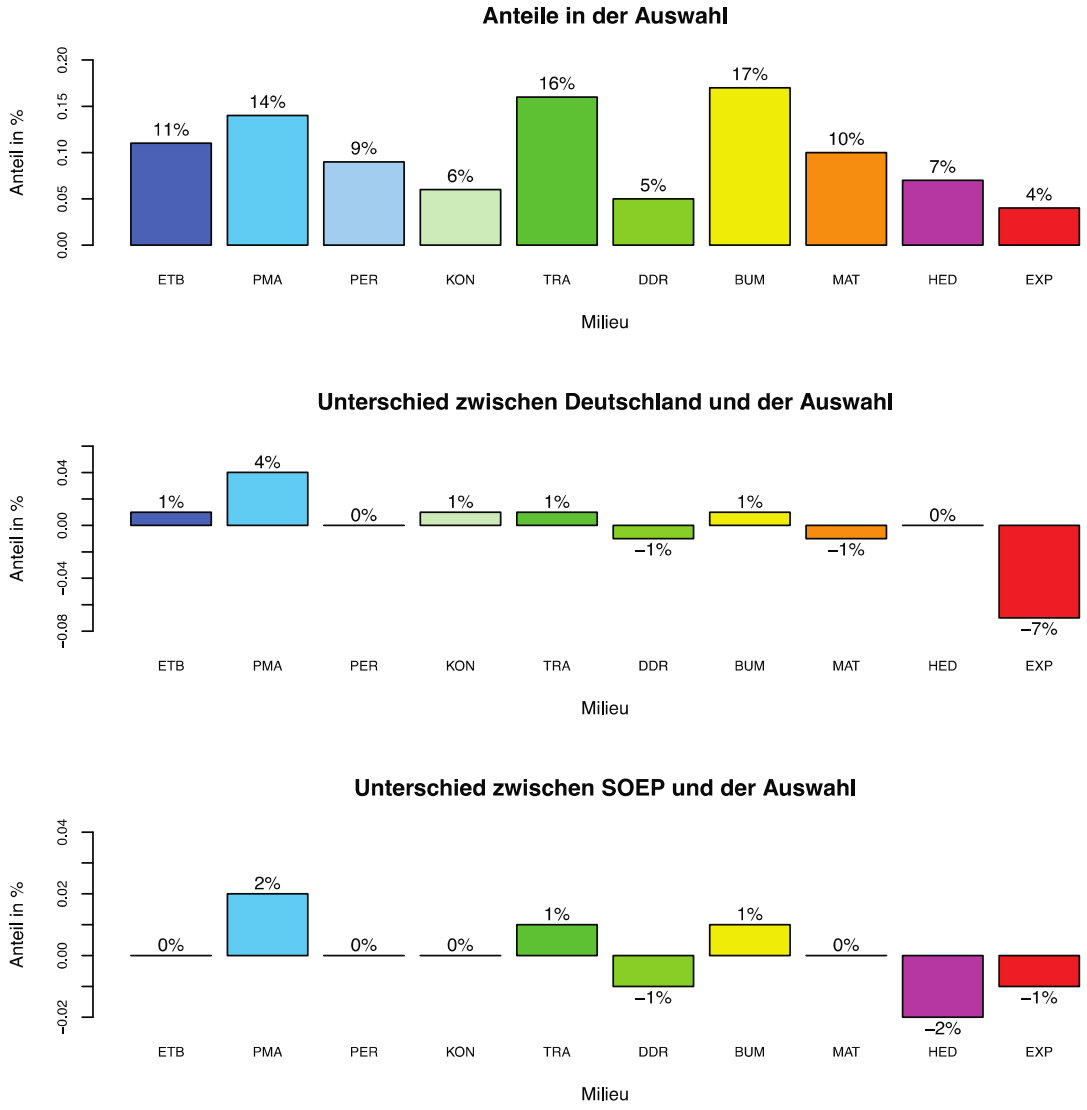

Figure 11: Milieuverteilung der Auswahl im Vergleich mit Deutschland und der Stichprobe, SINUS-Institut, SOEP, eigene Berechnung

4. Vorgehensweise

Die Datenanalyse basiert auf einem dreistufigen Verfahren. Im ersten Schritt wurde mit Hilfe der Singulärwertzerlegung (siehe Kap. 2.1) eine Orthogonaltrasformation der Orginalmatrix gesucht. Das Ergebnis dieses Verfahrens wurde dann in einem zweiten Schritt nach dem Verfahren von Carrol, Green, Schaffer skaliert und so eine interdimensionale Interpretation der einzelnen Variablen ermöglicht, die eine besonders gute Vorbereitung auf eine im dritten Schritt erfolgte Clusteranalyse war. Dieses Verfahren ermöglicht auch die Haushalte und Infrastrukturmerkmale in einem Joint Display darzustellen und grafisch zu analysieren.

Zunächst wurden, mit Hilfe des SOEPinfo Dienstes[11] des DIW, die Variablen aus dem Datenpool ausgewählt, die für die Analyse wichtig sind und aus dem Erhebungsjahr 2004 stammen. SOEPinfo generiert dabei wahlweise SPSS, SAS oder STATA kompatiblen Code für die Selektion. In der vorliegenden Auswertung wurde für diesen Aspekt das Statistikprogramm SPSS verwendet. Im weiteren Verlauf der Analyse wurde das freie Softwarepaket R in der Version 2.13.1 (2011-07-08) auf der Plattform x86_64-apple-darwin9.8.0/x86_64 (64-bit) (R Development Core Team, 2011) mit Paketen von Bowman and Azzalini (2010), Graffelman (2010), Harrell and with contributions from many other users. (2010), Lemon (2006), Neuwirth (2011), Neuwirth (2011), Revelle (2011), Wickham (2009), Wickham (2007) verwendet. Die mit SPSS erstellte Matrix wurde daher für die weitere Analyse in R eingelesen.

Nach dem Import wurden als nächstes die Haushalte ausgefiltert, die nicht den oben beschriebenen Kriterien entsprachen. Darüber hinaus wurden alle doppelten Einträge ausgefiltert[12]. Die Variablen der Infrastrukturmerkmale des so gefilterten Datensatzes wurden dann in eine Binärmatrix übertragen und die Variablen UH6405 und UH6406, UH6407 und UH6408 sowie UH6501 und UH6502 jeweils zusammengefasst. Das heißt, dass für jedes Merkmal I die Merkmalsausprägung 0 (ist nicht zu Fuß von 10 Minuten erreichbar) und 1 (ist in 10 Minuten zu Fuß erreichbar) pro Befragtem R gegeben ist. Hieraus resultiert bei 13 Infrastrukturmerkmalen ein $\mathbb{R}^{13}$ dimensionaler Würfel, auf dessen äußere Kanten die Merkmalsausprägungen aufgetragen sind (die Figure 12 zeigt dies am Beispiel der Variablen UH6401 und UH 6402). Normiert man die Matrix D entsprechend den Gleichungen (1.12) und (1.13), so erhält man einen gleich dimensionierten Würfel, auf dessen Flächen und Diagonalen die Merkmalsausprägungen aufgetragen sind (die Figure 13 zeigt dies am gleichen Beispiel).

11 http://panel.gsoep.de/soepinfo2009/

12 Die doppelten Einträge können durch Mehrfachbefragung eines Haushaltes entstehen.

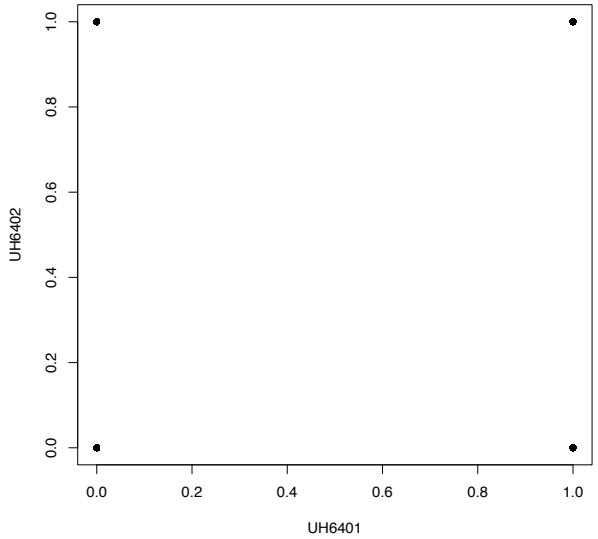

Figure 12: Sample D

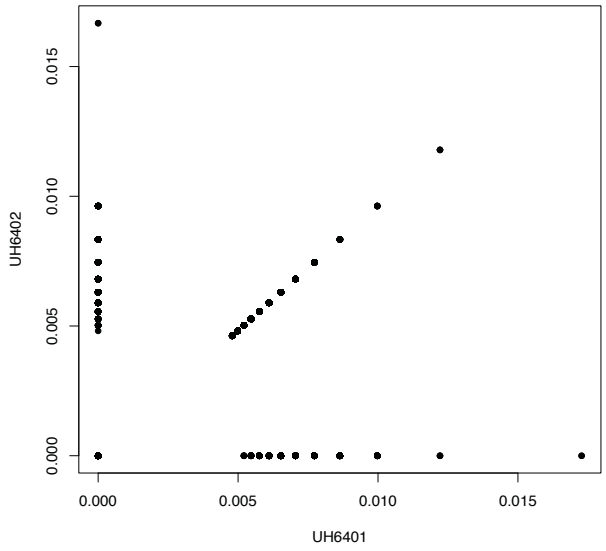

Figure 13: Sample M
Beispielhafte Darstellung der ursprünglichen Daten

Auf die so „normalisierte" Matrix, wurde dann die SVD angewandt und die Ergebnisse jeweils in die Matrizen U, S und V' geschrieben. Zur Reskalierung der Werte in die Carrol, Green, Schaffer Koordinaten, wurden die Matrizen zunächst in die „standard coordinates" ($\tilde{u} \hat{=} Util$ und $\tilde{v} \hat{=} Vtil$) und in einem zweiten Schritt in die Carrol, Green, Schaffer Skalierung ($\ddot{u} \hat{=} Urd$ und $\ddot{v} \hat{=} Vrd$) transformiert. Die Ergebnisse wurden getestet und weisen einen gewichteten Mittelwert von 1 und eine gewichtete Varianz von 1 bzw. $\frac{\sigma_k + 1}{2}$ auf.

Im Folgenden wurde nun versucht eine Dimensionsreduktion vorzunehmen. Hierzu wurde die Inertia für jede Dimension berechnet und ausgewertet. Betrachtet man Table 7: Inertia von D, so erkennt man, dass eine noch graphisch darstellbare Projektion vom Rang 3 nur

50 % der Varianz erklären könnte. Weiterhin kann man erkennen, dass es jeweils bei der Dimension 5 und 12 einen Knick in der Zunahme des Varianzanteils pro Dimension gibt (siehe Figure 14). Die Dimensionen 1 bis 5 erklären jedoch auch nur einen Varianzanteil von 65,12 %. Eine Dimensionsreduktion mit dem Verlust von fast 35% erscheint hier jedoch nicht vertretbar, insbesondere, da es keinen Vorteil in der graphischen Darstellung ermöglicht. Daraus kann gefolgert werden, dass die Infrastrukturmerkmale nicht wie erhofft in wichtige und unwichtige getrennt werden können, sondern dass alle Infrastrukturmerkmale für die Beschreibung einer Nachbarschaft herangezogen werden müssen. Daher wurden in der anschließenden Analyse alle 13 Dimensionen berücksichtigt.

Table 7: Inertia von D, eigene Berechnung

Dimension	σ_k	σ_k^2	Anteil In(D)	Kummulierter In(D) Anteil
1	0.412	0.170	23.545	23.545
2	0.332	0.110	15.320	38.865
3	0.284	0.081	11.217	50.082
4	0.249	0.062	8.564	58.645
5	0.216	0.047	6.476	65.121
6	0.210	0.044	6.112	71.233
7	0.201	0.040	5.607	76.840
8	0.193	0.037	5.156	81.996
9	0.183	0.034	4.657	86.653
10	0.172	0.030	4.120	90.772
11	0.163	0.027	3.685	94.457
12	0.157	0.025	3.420	97.878
13	0.124	0.015	2.122	100.000
Summe		0.721		

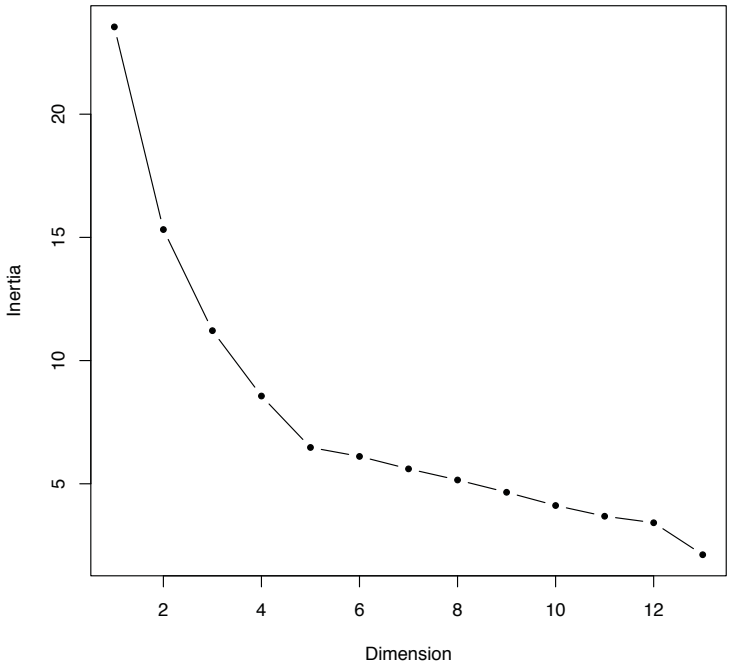

Figure 14: Inertia pro Dimension von D, eigene Berechnung

Im Anschluss an die Singulärwertzerlegung und der Reskalierung der Daten wurde eine Clusteranalyse auf Basis der Daten aus $\ddot{U}$ durchgeführt. Die Klassifizierungsanalyse wurde in einem zweistufigen Verfahren durchgeführt. In einem ersten Schritt wurde eine hierarchische Clusteranalyse vorgenommen und die Ergebnisse dieser Klassifizierung in einem zweiten Schritt mit dem K-means Verfahren optimiert. Da es sich bei den Koordinaten in $\ddot{U}$ um metrisch skalierte Daten handelt, wurde als Distanzmaß für die hierarchische Clusteranalyse die euklidische Distanz gewählt. Da sich die Verteilung der Befragten, auf Grund der SVD, um ein Infrastrukturmerkmal orientieren, wurde der Ward-Algorithmus als Fusionierungsmethode gewählt, da dieser besonders zum Auffeinden sphärischer Cluster geeignet ist. In Versuchen mit andern Fusionierungsstrategien zeigte der Ward-Algorithmus darüber hinaus auch die besten Separationsergebisse.

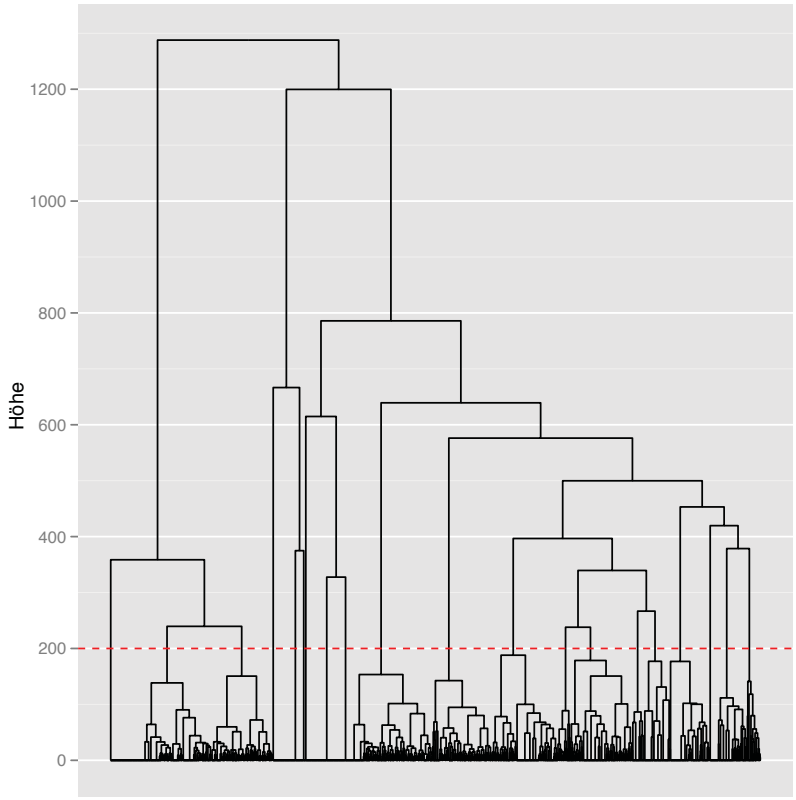

Figure 15: Dendrogram der Clusteranalyse von $\ddot{U}$, eigene Berechnung

Das Ergebnis der hierarchischen Clusteranalyse über den 13-dimensionalen Raum ist in Figure 15 dargestellt. Diese Abbildung legt eine Einteilung in 20 Clustern dar (rote gestrichelte Linie). In dieser Höhe differenzieren sich die verschiedenen Gruppen noch sehr gut, bei einer noch geringen Heterogenität innerhalb der Cluster[13]. In einem zweiten Schritt wurde nun versucht die gefundene Konfiguration mit Hilfe des k-means Verfahrens zu optimieren.

13 Je länger die vertikalen Linien, desto größer ist der Abstand der zu fusionierenden Cluster und damit die Hetrogenität.

Table 8: Typische Nachbarschaften, eigene Berechnung

Cluster	Geschäfte	Gaststätten	Banken	Hausärzte	Alteneinrichtungen	Park	Sportstätten	ÖPNV	Mangel an Grünflächen	Keine Kriminalität	Multikulturelle Umgebung	Keine Umweltbeeinträchtigungen	Kindereinrichtungen	Einrichtungen für Jugendliche	Anzahl Befragter	Anteil
1	0.103	0.097	0.098	0.115	0.000	0.070	0.043	0.106	0.084	0.108	0.000	0.106	0.070	0.000	572	9,56 %
2	0.074	0.075	0.069	0.039	0.088	0.073	0.062	0.086	0.065	0.072	0.050	0.077	0.081	0.088	471	7,87 %
3	0.000	0.037	0.005	0.032	0.005	0.069	0.005	0.291	0.159	0.328	0.053	0.000	0.016	0.000	62	1,04 %
4	0.185	0.094	0.000	0.000	0.000	0.062	0.001	0.151	0.120	0.170	0.000	0.167	0.050	0.000	284	4,75 %
5	0.077	0.076	0.062	0.044	0.000	0.069	0.040	0.116	0.091	0.111	0.137	0.120	0.055	0.000	526	8,79 %
6	0.039	0.064	0.000	0.006	0.000	0.120	0.153	0.140	0.135	0.147	0.000	0.148	0.049	0.000	436	7,29 %
7	0.000	0.000	0.003	0.003	0.000	0.000	0.000	0.000	0.332	0.331	0.000	0.331	0.000	0.000	209	3,49 %
8	0.080	0.086	0.075	0.055	0.107	0.080	0.045	0.102	0.078	0.094	0.023	0.097	0.077	0.000	420	7,02 %
9	0.074	0.082	0.067	0.051	0.000	0.070	0.055	0.100	0.083	0.100	0.021	0.101	0.087	0.110	743	12,42 %
10	0.104	0.114	0.132	0.000	0.000	0.073	0.048	0.120	0.100	0.125	0.000	0.125	0.060	0.000	474	7,92 %
11	0.000	0.183	0.000	0.016	0.000	0.081	0.000	0.165	0.151	0.169	0.000	0.175	0.059	0.000	245	4,10 %
12	0.000	0.000	0.001	0.018	0.000	0.000	0.000	0.248	0.248	0.236	0.000	0.248	0.000	0.000	220	3,68 %
13	0.000	0.000	0.000	0.015	0.000	0.029	0.000	0.000	0.279	0.000	0.147	0.529	0.000	0.000	36	0,60 %
14	0.078	0.079	0.078	0.082	0.082	0.075	0.067	0.081	0.065	0.077	0.000	0.076	0.079	0.082	607	10,15 %
15	0.000	0.000	0.017	0.024	0.000	0.083	0.000	0.133	0.159	0.190	0.000	0.193	0.201	0.000	152	2,54 %
16	0.000	0.000	0.007	0.007	0.000	0.203	0.000	0.203	0.189	0.192	0.000	0.200	0.000	0.000	212	3,54 %
17	0.007	0.014	0.007	0.000	0.000	0.252	0.000	0.000	0.238	0.231	0.000	0.249	0.000	0.002	108	1,81 %
18	0.000	0.000	0.030	0.010	0.003	0.000	0.000	0.334	0.000	0.291	0.000	0.325	0.000	0.007	101	1,69 %
19	0.000	0.000	0.017	0.017	0.006	0.000	0.006	0.000	0.000	0.478	0.000	0.478	0.000	0.000	85	1,42 %
20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	19	0,32 %

Durch die Optimierung wurden 1006 Elemente neu zugeordnet, das entspricht einer Quote von 16,817% aller Elemente. Auf Basis dieser optimierten Zuordnungen wurden nun die typischen Nachbarschaften in Deutschland ermittelt. Hierzu wurde die, durch die Anzahl der Befragten in einem Cluster, gewichtete Summe der Nennungen pro Infrastruktur in jedem Cluster ermittelt. Anschließend wurde der Anteil eines jeden Infrastrukturmerkmals an der Nachbarschaft eines Clusters berechnet.

Vergleicht man die Nachbarschaften der unterschiedlichen Cluster miteinander (siehe auch Figure 16), so ist festzustellen, dass diese sich deutlich, hinsichtlich der Nachbarschaftsstruktur und der Anzahl von Bewohnern, voneinander unterscheiden.

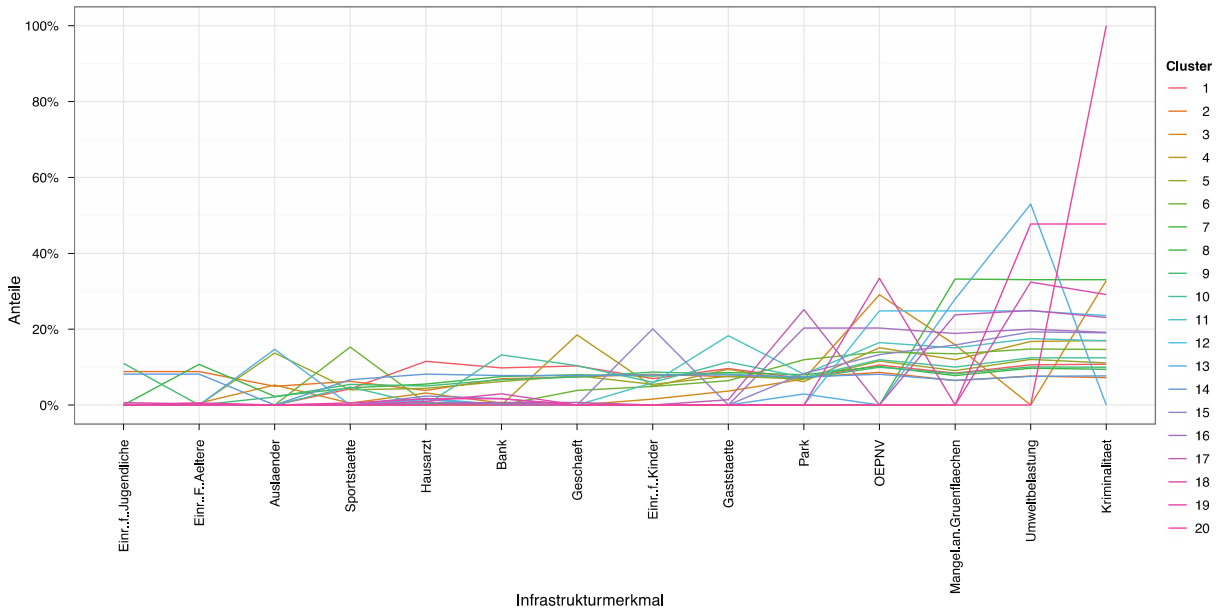

Figure 16: Typische Nachbarschaften, eigene Berechnung

So schwankt der Bevölkerungsanteil in den Clustern zwischen 0,32% im Cluster 20 und 12,42% im Cluster 9. Auffällig ist dabei auch, dass die Infrastrukturmerkmale „Alteneinrichtungen“, „Multikulturelle Umgebung“ und „Einrichtungen für Jugendliche“ nur in sehr wenigen Clustern, vorkommen. Dies deutet auf eine starke Seggregation von Ausländern, älteren Menschen und Jugendlichen bzw. Familien in bestimmten Nachbarschaftstypen hin. Die identifizierten Nachbarschaften verfügen in 3/4 aller Fälle über eine sehr gute Erreichbarkeit des öffentlichen Personennahverkehrs - eine Haltestelle ist in den Nachbarschaftstypen, in denen eine ÖPNV Anbindung vorhanden ist, in 91 % der Fälle in maximal 10 min zu Fuß zu erreichen. Betrachtet man Figure 17, so ist keine eindeutige Präferenz eines Milieus zu einem Cluster zu erkennen. Zwar schwanken die Milieuverteilungen in den einzelnen Clustern, doch gibt es keine signifikante Korrelationen zwischen den Clustern und einem Milieu (Cramer's V: 0.086).

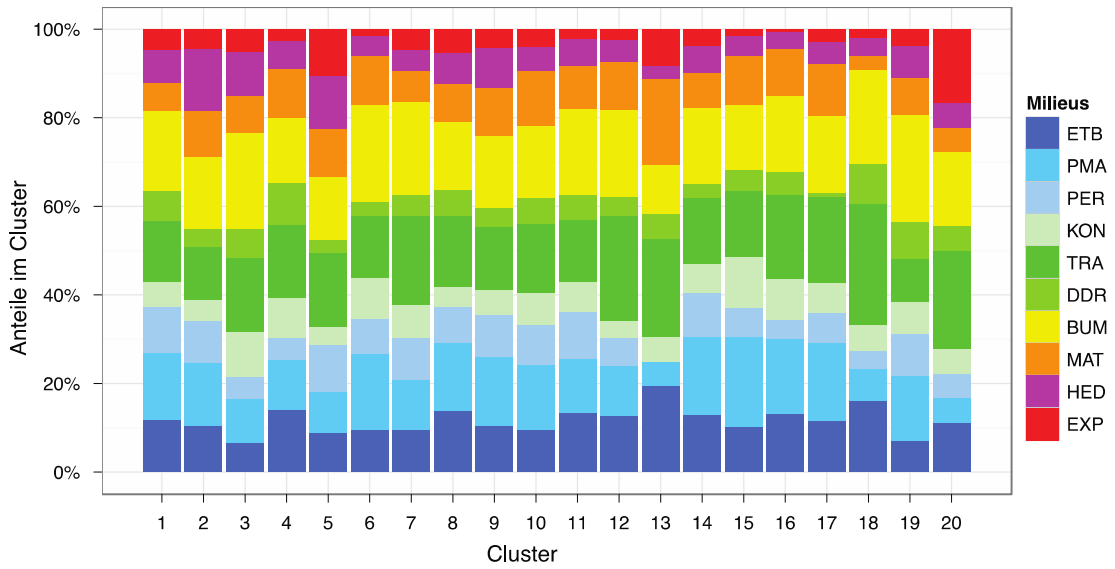

Figure 17: Milieuverteilung in den Clustern, eigene Berechnung

Führt man das beschriebene Verfahren jedoch für jedes einzelne Milieu durch und summiert die Infrasturkturmerkmale eines Milieus über alle Dimensionskombinationen, die den Bedeutungshorizont überschreiten, so kann gezeigt werden, dass sich die typischen Nachbarschaften sowohl der Lebenswelten als auch der Milieus voneinander unterscheiden (siehe Figure 18). Die Anordnung der Infrastrukturmerkmale folgt dabei der Anordnung von weniger wichtig (links) bis wichtiger (rechts), der durchschnittlichen Häufigkeit aller Milieus. Es kann gezeigt werden, dass die Ausländerrate, die empfundene Umweltbelastung sowie die Anbindung an den Öffentlichen Personennahverkehr für eine Zufriedenheit mit der Nachbarschaft im Gegensatz zu Sportstätten, Einrichtungen für Jugendliche und Einrichtungen für Ältere eine weniger wichtige Rolle spielen. Dabei zeigen sich jedoch milieuspezifische Unterschiede. So ist z.B. ein Mangel an Grünflächen für Experimentalisten, bei der bürgerlichen Mitte die Versorgung mit Hausärzten oder der Versorgung mit Geschäften für die postmateriellen Milieus ein wichtiger Indikator für die Zufriedenheit mit dem eigenen Quartier.

Auf Grund der vorliegenden Ergebnisse liegt die Vermutung nahe, dass das vorgestellte Verfahren geeignet ist die Milieuzugehörigkeit eines Haushaltes auf Grund der des jeweiligen Nachbarschaftsprofils zu bestimmen. Um dies zu überprüfen soll in den folgenden Forschungsarbeiten versucht werden an Hand des Nachbarschaftsprofils eines Haushaltes das wahrscheinlichste Milieu zu bestimmen. Hierzu sollen Nachbarschaftsprofile von Haushalten mit den Profilen der Milieus verglichen werden und der Haushalt demjenigen Milieu zugeordnet werden, zu dessen Profil die geringste Profildistanz besteht. Die Zuordnung wird dann mit der Milieuangabe verglichen. Sollte es zu einer hohen Übereinstimmung

kommen, so kann dieses Verfahren angewendet werden um Wohnpräferenzkarten auf Basis von Infrastrukturmerkmalsinformationen zu erstellen. Die Erstellung solcher Karten ist dann Gegenstand weiterer Forschung.

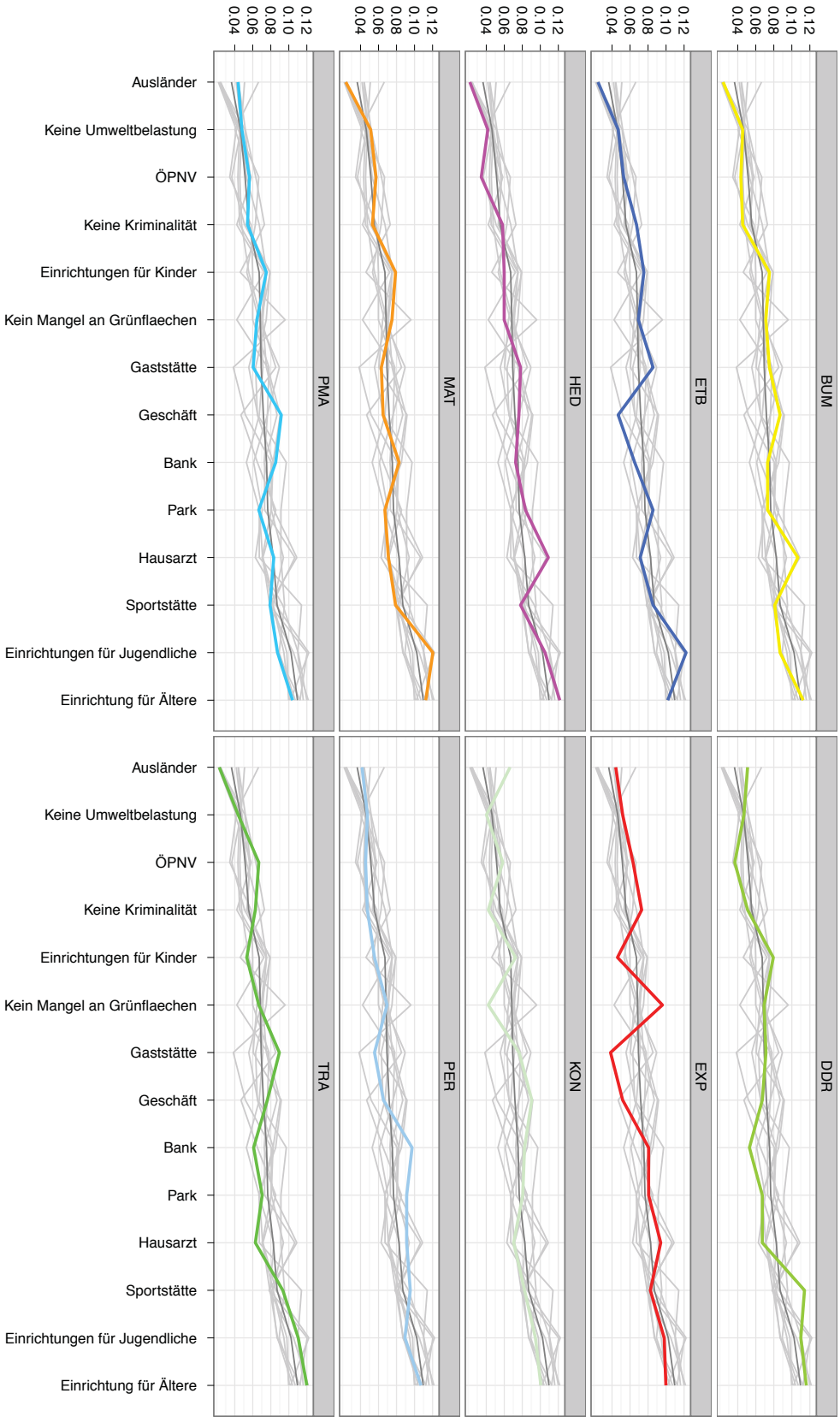

Figure 18: Nachbarschaftsstruktur der Milieus, eigene Berechnung

References

Martin Becker. Lebensqualität im Stadtquartier. PhD thesis, Albert-Ludwigs- Universität zu Freiburg i. Br., Freiburg i. Br., 2003.

A. W. Bowman and A. Azzalini. R package sm: nonparametric smoothing methods (version 2.2-4). University of Glasgow, UK and Università di Padova, Italia, 2010. URL http://www.stats.gla.ac.uk/ adrian/sm.

Ingrid Breckner. Künftige Anforderungen an Wohnumfeld- und Freizeitqualitäten in den Städten von Nordrhein-Westfalen. Bericht für die Enquêtekommission "Zukunft der Städte in NRW" des Landtags Nordrhein-Westfalen. Landtag Nordrhein-Westfalen, Hamburg, 2003.

Ingrid Breckner. Eliten, Minderheiten und soziale Milieus als regionale Entwicklungsressourcen. In: Elmar Hönekopp, Rolf Jungnickel, and Thomas Straubhaar (eds.), Internationalisierung der Arbeitsmärkte, Volume 282 of Beiträge zur Arbeits- und Berufsforschung, pages 209–230. Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit., Nürnberg, 2004.

William P. Butz and Barbara B. Torrey. Some frontiers in social science. Science, 312(5782):1898–1900, 2006.

Douglas J. Carroll, Paul E. Green, and Catherine M. Schaffer. Interpoint distance comparisons in correspondence analysis solutions. Journal of Marketing Research, 23:217–280, 1986.

Douglas J. Carroll, Paul E. Green, and Catherine M. Schaffer. Reply to greenacre's commentary on the carroll-green-schaffer scaling of two-way correspondence analysis solutions. Journal of Marketing Research, 26(3):366–368, 1989. ISSN 0022-2437.

Katherine Faust. Using correspondence analysis for joint displays of affiliation networks. In Peter J. Carrington, John Scott, and Stanley Wassermann, editors, Models and Methods in Social Network Analysis, chapter 7, pages 117–147. Cambridge University Press, New York, 2005.

J.R. Frick, S.P. Jenkins, D.R. Lillard, O. Lipps, and M. Wooden. Die internationale Einbettung des sozio-oekonomischen panels (soep) im rahmen des cross-national equivalent file (cnef). Vierteljahrshefte zur Wirtschafts- forschung, 77(3):110–129, 2008. ISSN 0340-1707.

GdW. Mietwohnungen in Deutschland - ein attraktives und wertbeständiges Marktsegment., August 2004.

Jan Graffelman. calibrate: Calibration of Scatterplot and Biplot Axes, 2010. URL http://CRAN.R-project.org/package=calibrate. R package version 1.7.

Michael J. Greenacre. Theory and applications of correspondence analysis. Academic Press, 1984.

Michael J. Greenacre. The carroll-green-schaffer scaling in correspondence: A theoretical and empirical appraisal. Journal of Marketing Research, 26:358– 365, 1989.

Frank E Harrell and with contributions from many other users. Hmisc: Harrell Miscellaneous, 2010. URL http://CRAN.R-project.org/package=Hmisc. R package version 3.8-3.

Karl-Dieter Keim. Milieu in der Stadt. Ein Konzept zur Analyse älterer Wohnquartiere. Kohlhammer, Stuttgart, 1979.

J Lemon. Plotrix: a package in the red light district of r. R-News, 6(4):8–12, 2006.

Ulf Matthiesen. Die Räume der Milieus. Neue Tendenzen in der sozial- und raumwissenschaftlichen Milieuforschung, in der Stadt- und Raumforschung. edition sigma, Berlin, 1998.

Erich Neuwirth. RColorBrewer: ColorBrewer palettes, 2011. URL http://CRAN.R-project.org/package=RColorBrewer. R package version 1.0-5.

R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL http://www.R-project.org/. ISBN 3-900051-07-0.

William Revelle. psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois, 2011. URL http://personality-project.org/r/psych.manual.pdf. R package ver- sion 1.0-97.

Sachverständigenrat. Analyse: Entwicklung der personellen Einkommensverteilung in Deutschland. Jahresgutachten 2006/2007, pages 428–447, 2006.

Joachim Scheiner and Christian Holz-Rau. Travel mode choice: affected by objective or subjective determinants? Transportation, 34(4):487–511, July 2007.

Statistisches Bundesamt. Datenreport 2002. Bundeszentrale für politische Bildung, Bonn, 2002.

Statistisches Bundesamt. Wirtschaftsrechnungen. Einkommens- und Verbraucherstichprobe. Einkommensverteilung in Deutschland, 2006.

Statistisches Bundesamt. Wirtschaftsrechnungen. Einkommens- und Verbrauchsstichprobe. Einnahmen und Ausgaben privater Haushalte, 2007.

Gert G. Wagner, Joachim R. Frick, and Jürgen Schupp. The german socio- economic panel study (soep). Schmollers Jahrbuch, 127(1):139–169, 2007.

Susan C. Weller and A. Kimball Romney. Metric scaling: Correspondence analysis. Sage Publications, London, 1990.

Hadley Wickham. Reshaping data with the reshape package. Journal of Statistical Software, 21(12), 2007. URL http://www.jstatsoft.org/v21/i12/paper.

Hadley Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009. http://had.co.nz/ggplot2/book.