GfKl 2004 Contest: Correspondence clustering of Dortmund city districts

Stefanie Scheid

Department for Computational Molecular Biology,
Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany,
stefanie.scheid@molgen.mpg.de

Abstract. We combine correspondence analysis (CA) and K-means clustering to divide Dortmund's districts into groups that are associated to particular variables and thus represent a social cluster. CA visualizes associations between rows and columns of a frequency matrix and can be used for dimension reduction. Based on the first three dimensions after CA mapping we find a stable partition into five clusters. We further identify variables that are highly associated with the cluster centers and thus represent a cluster's social condition.

1 Introduction

The city of Dortmund is regarded as a highly heterogeneous urban area. The 170 districts comprise the city center (the ancient Dortmund) and incorporated suburbs that are nowadays either connected to the city center or maintained as village-like entities. The city covers industrial areas as well as fields and forests.

The GfKl 2004 Contest provides data of 200 variables for each district. The question is whether the city's heterogeneity is reflected in the data and whether it is possible to classify the districts with respect to their social environment. Since the definition of "social environment" is wide, we do not only regard typical social variables like amount of unemployed adults. By defining general preselection rules based on correlation and variability, we arrive at a set of variables that does not only represent social but general residential information.

The reduced data set is submitted to correspondence analysis (CA). CA is a useful tool to visualize associations between rows and columns of a frequency matrix, that is between districts and variables. Row and column vectors are simultaneously mapped into a space where the similar direction of vectors reflects their association. Proximity of districts alone reflects close relationship with respect to similar associated variables. We apply K-means clustering after mapping to correspondence space. We derive a stable cluster pattern of districts that can be interpreted with regard to the corresponding variables.

Alosaus	Alosdeut	BJ01bis19	BJ19bis48
BJ49bis57	BJ58bis62	BJ63bis72	BJ73bis82
BJ83bis92	BJ93bis01	BJbis1900	Fins
GebBest10ins	GebBest1und2ins	GebBest3ins	Gebzus
GenNeuins	Hhins	HWBinsA	HWBinsD
InstdtUmZuzus	Kins	LKW	Mins
PKW	Sozempfzus	sozvpflBeschAus	sozvpflBeschDeut
Sterbzus	WanZuzus	WGebäuins	Zugmaschine

Table 1. List of remaining variables after preprocessing.

2 Material and methods

Material. The data set consists of 170 districts covering Dortmund and 200 variables collected in 2002. The variables give complete inventory of population (German, foreign, births, deaths, movements), unemployment, social welfare, buildings (stock, construction, covered area) and motor vehicles. All variables are measured as absolute frequencies except those representing areas. Area variables are given in square meters, the total area is given in hectares. We exclude the variables "total population" (`HBWins`) and "total area" (`Flächeha`) and use them for scaling.

The data set is accompanied by geographical information. For each district a planar polygonal representation is given to account for spatial relations.

Variable preselection and scaling. Many variables are linear combinations of others or are otherwise highly correlated. We scan variables belonging to one topic separately. For each group of variables with mutual correlation coefficients exceeding ± 0.7 one representative is selected manually. If possible, the variable containing a grand total is preferred. The four variables regarding welfare recipients are merged into one variable (`Sozempfzus`).

The remaining variables are scaled with their median absolute deviation. Variables with median absolute deviation of zero are removed from further consideration. The preselection process results in 32 variables, none of them contains area (see Table 1).

The city districts differ with respect to population and area. Each district is scaled with its density, that is inhabitant per hectare. After preprocessing and scaling, the remaining data matrix contains informative variables comparable on the same scale.

Correspondence analysis. Given a matrix of absolute frequencies, we can compute the χ^2 test statistic for homogeneity. Similar to principal component analysis, CA provides the mapping of variables into a lower space while preserving a considerable percentage of the χ^2 statistic. CA maps row and column vectors simultaneously into the same space. A row and a column vector are positively associated if they point to the same direction, that is share a small angle. The more remote they are from the origin, the more associated they are. Row and column vectors pointing to opposite directions can be interpreted as negative associations. In the following, we introduce the main concepts of CA. For detailed theory and further concepts see for example Mirkin (1996, chap. 2.3.3) and Nakayama (2001).

Let $\mathbf{N}$ be a matrix of absolute frequencies with r rows, c columns, entries n_{ij} and grand total n. Hence, $\mathbf{F} = n^{-1}\mathbf{N}$ is the matrix of corresponding relative frequencies. We further define $\mathbf{D}_r$ as the $r \times r$ diagonal matrix of mean row profile $(n_{1.}/n, \ldots, n_{r.}/n)$ and $\mathbf{D}_c$ as the $c \times c$ diagonal matrix of mean column profile $(n_{.1}/n, \ldots, n_{.c}/n)$, where $n_{i.}$ and $n_{.j}$ are row and column sums of $\mathbf{N}$.

The distance between two columns j and j' of $\mathbf{N}$ can be measured in χ^2 distance d^2:

$$d^2(j, j') = \sum_{i=1}^{r} \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}} \right)^2 . \tag{1}$$

The χ^2 distance is directly connected to the Euclidean distance: If we apply the transformation $\tilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1}$, then $d(j, j')$ is the Euclidean distance between columns j and j' of $\tilde{\mathbf{F}}$.

The objective of CA is to find a mapping of the columns of $\mathbf{N}$ from space $\mathbb{R}^r$ to a lower dimensional space $\mathbb{R}^m$ with $m < r$. The relative positions of two columns in χ^2 distance before mapping and Euclidean distance after mapping are preserved. A suitable mapping is found by singular value decomposition of the matrix $\mathbf{S}$ with:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2} - \mathbf{D}_r^{1/2}\mathbf{1}_r\mathbf{1}_c^T\mathbf{D}_c^{1/2}. \tag{2}$$

Singular value decomposition decomposes $\mathbf{S}$ into $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{W}^T$. The matrices $\mathbf{V}$ and $\mathbf{W}$ contain left and right singular vectors respectively. Matrix $\mathbf{\Lambda}$ is the diagonal matrix of singular values $\lambda_1 \geq \cdots \geq \lambda_k > 0$ with $k \leq \min(r, c)$.

The columns of $\mathbf{N}$ can be presented in space $\mathbb{R}^k$ as the columns of $\mathbf{C}$ with:

$$\mathbf{C} = \left(\lambda_1 \mathbf{D}_c^{-1/2} w_1, \ldots, \lambda_k \mathbf{D}_c^{-1/2} w_k \right), \tag{3}$$

where w_s denotes the sth column of $\mathbf{W}$.

Applying the same considerations to the rows of $\mathbf{N}$ it follows that the rows can be presented as the columns of $\mathbf{R}$ with:

$$\mathbf{R} = \left(\lambda_1 \mathbf{D}_r^{-1/2} v_1, \ldots, \lambda_k \mathbf{D}_r^{-1/2} v_k \right), \tag{4}$$

where v_s denotes the sth column of $\mathbf{V}$.

The total sum of squares of $\mathbf{S}$ is equal to the χ^2 statistic of $\mathbf{N}$ divided by n. The value χ^2/n is called *inertia*, a term that interprets relative frequency as mass. Due to singular value decomposition, the total inertia is equal to the sum of squared singular values of $\mathbf{S}$. We compute the proportion of inertia explained by the first singular value as $\lambda_1^2 / \sum_{s=1}^{k} \lambda_s^2$. Regarding a chosen proportion of explained inertia, we represent the data in a lower dimensional space $\mathbb{R}^m$ by only considering the first m columns of $\mathbf{C}$ and $\mathbf{R}$.

K-means clustering. The K-means algorithm of Hartigan and Wong (1979) divides data points into K clusters. Initially, K data points are chosen randomly as cluster centers. In an iterative process, the algorithm assigns a data point to a cluster if this minimizes the within-cluster sum of squares of Euclidean distances to the center. The centers are updated by replacing them with the mean of within-cluster points. The iteration stops in a local minimum, that is, no reallocation of points results in a smaller within-cluster sum of squares.

The number of clusters K is chosen such that the resulting partition of points into clusters is optimal with respect to a cluster validation index. A suitable index is based on the silhouette score s_i for data point i:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{5}$$

where a_i denotes the average Euclidean distance between point i and its within-cluster points. For all clusters not containing point i, the average distance between i and within-cluster points is computed. The term b_i denotes the minimum of these values, that is the average distance of i to its nearest neighboring cluster. The silhouette score ranges from -1 to 1. A score near 1 supports the allocation of i to an appropriate cluster. A point allocated to an inadequate cluster has a score near -1.

A general validation index for a specific K-means partition is given by the global silhouette score, that is the mean silhouette score over all points. The highest global silhouette score leads to the optimal number of clusters.

Two runs of K-means can lead to different partitions if the initial centers were chosen with random seeds. The process is stable if a reasonable percentage of several runs with K random centers results in identical partitions. Therefore, we propose a second measure for cluster validation: A process is stable if the resulting partitions are highly concentrated, that is, a small number of non-identical partitions is observed with high frequency. We allow a difference in one point to call two partitions identical. Given l non-identical partitions with observed relative frequencies $\hat{\pi}_j$, $j = 1, \ldots, l$, we measure concentration via normalized entropy NE_K:

$$NE_K = -\sum_{j=1}^{l} \hat{\pi}_j \frac{\log \hat{\pi}_j}{\log l}. \tag{6}$$

A low value of normalized entropy corresponds to high concentration. Therefore, the lowest value of normalized entropy leads to the optimal number of clusters.

The relationship between χ^2 and Euclidean distance in CA suggests that clustering on Euclidean distances after CA mapping is equivalent to clustering on original χ^2 distances. Clustering can be performed directly on the

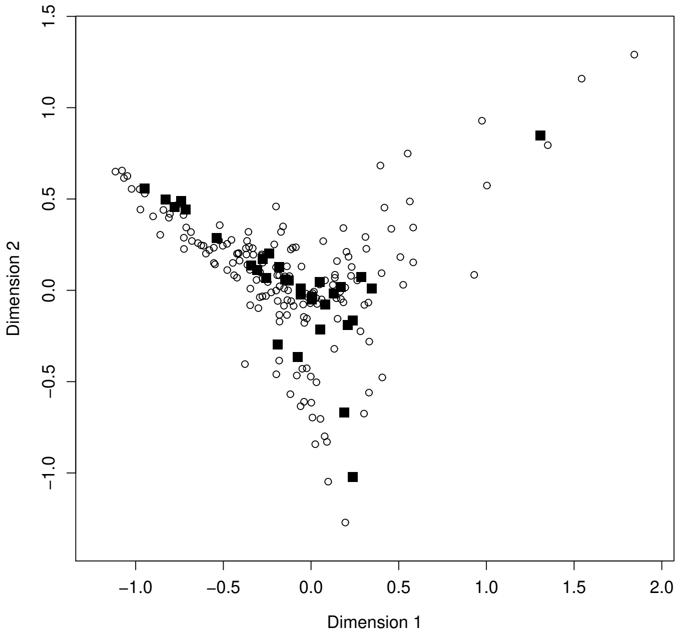

Fig. 1. Biplot of districts (circles) and variables (squares) based on first two dimensions after CA mapping.

original data set using χ^2 distances. Nevertheless, a stable partition is often gained after dimension reduction. We apply CA first and choose the number of dimensions and the number of clusters based on global silhouette and normalized entropy scores.

We use the statistical software `R`, version 1.7.1 with incorporated functions `kmeans` and `silhouette`, for data analysis (Ihaka and Gentleman (1996)). Source code is available on request.

3 Results

After preprocessing, the scaled data matrix containing 170 districts and 32 variables is submitted to CA. Figure 1 presents the two-dimensional biplot of districts and variables after mapping. Districts and variables that point to the same direction are positively associated. From this low dimensional plot we can already assume more than three clusters of districts which are associated with the same variables. Many districts scatter around the origin and are not associated with a particular variable.

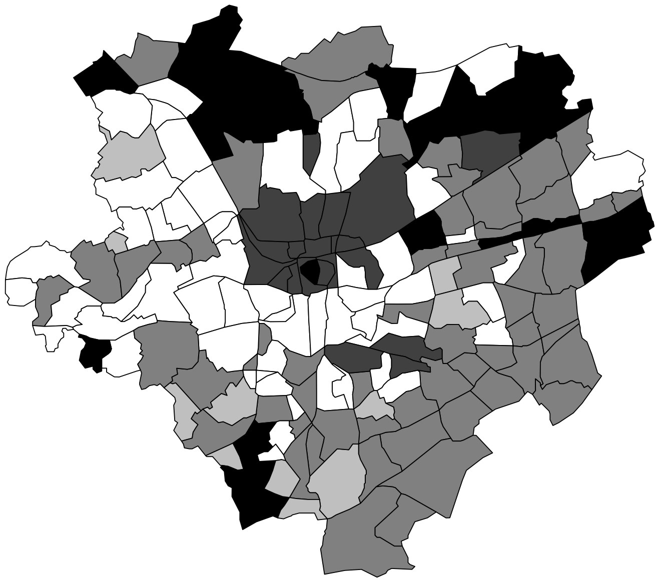

Fig. 2. Map of Dortmund city districts after correspondence clustering. Districts of one color belong to one cluster.

As successive CA dimensions explain fewer percentages of inertia, we focus on the first m dimensions that lead to high silhouette scores. The performance of K-means also depends on the number of cluster centers. We maximize the global silhouette score with respect to the number of CA dimensions and the number of clusters. For K ranging from two to ten and the number of dimensions ranging from three to eleven we conduct 1 000 partitions each and compute the mean global silhouette score for each combination. The maximal mean score is reached for partitions with four clusters on three dimensions followed closely by five clusters on three dimensions. The latter has a three times lower standard error. Regarding the combination of mean score and standard error we choose five clusters on three dimensions. The first three dimensions together explain 58.5% of total inertia.

The 5-means clustering on three dimensions is also optimal regarding its concentration. We conduct 10 000 runs for each combination of dimensions and cluster centers as above. The minimal normalized entropy is reached for five clusters and three dimensions where the algorithm converges at high rate to the cluster pattern shown in Figure 2.

In addition, a randomization test for spatial correlation is performed to test whether neighboring districts in CA are also near in reality. A spatial correlation test compares two distance matrices corresponding to the same variables by correlation coefficient and assigns a randomization p-value

Fig. 3. Maps of inner city cluster (21 districts), western circle (57 districts) and eastern circle (59 districts).

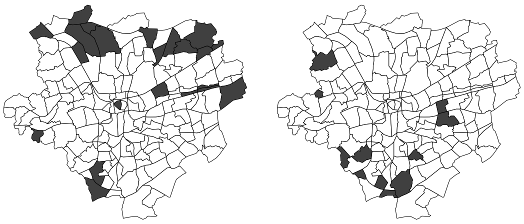

Fig. 4. Maps of northern suburbs (22 districts) and small cluster (11 districts).

(Manly (2001, chap. 9)). One distance matrix contains the mutual Euclidean distances based on the first three CA dimensions. The entries of the second matrix are 1 if two districts are real neighbors, that is, if they share at least one point in their polygonal representation, and 0 if not. The two matrices are weakly negatively correlated but show a significant p-value < 0.0001 based on 10 000 randomizations. Therefore, clustering on CA mapping corresponds to the simple neighboring relationship of districts.

Dortmund is divided into a small, two mediate and two large clusters. As suggested by the spatial correlation analysis, most members of a cluster are connected geographically. The inner city districts cluster together and are surrounded by the two large clusters. The latter span a circle of outer city districts, accumulating the western and eastern districts respectively. A fourth cluster contains mostly northern rural suburbs but also eastern industrial areas and a part of the city center. The smallest cluster is scattered over Dortmund with emphasis on south-eastern districts. Figures 3 and 4 show maps of each cluster versus the others.

Finally, we identify those variables that are highly associated with the cluster centers. In three-dimensional space we compute length of variable vectors and angles between centers and variables. The smaller the angle and the further apart from the origin the variable point is, the higher is the association to the corresponding cluster. We select variables with an angle smaller than 0.3 in radiant and a vector length greater than 0.4.

The analysis shows that all clusters are associated with one or more variables except the western circle which is simply not associated with any vari-

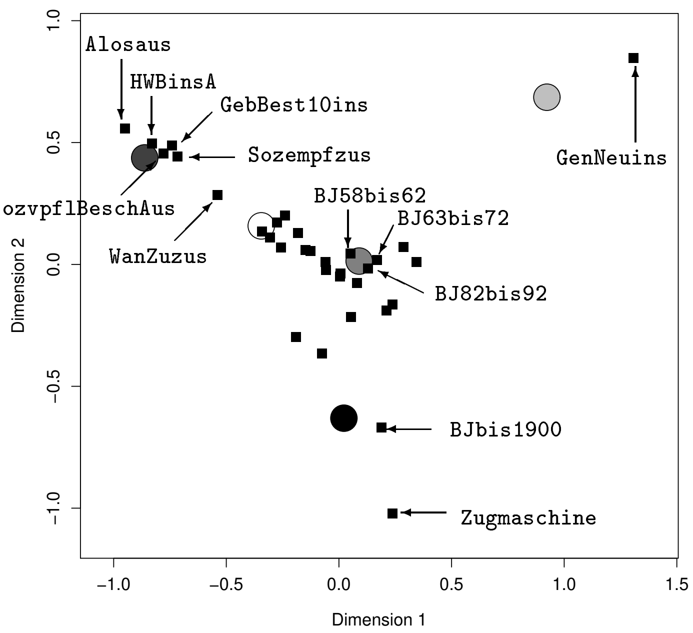

Fig. 5. Two-dimensional biplot of cluster centers (circles) and variables (squares). Colors of cluster centers correspond to colors in Figure 2. Labeled variables are positively associated to nearest cluster center based on first three CA dimensions.

able. Figure 5 shows a biplot similar to Figure 1 but with districts represented by their cluster centers. Variables that are highly associated with the nearest cluster center are labeled.

Six variables are positively associated to the inner city cluster: Social welfare recipients (`Sozempfzus`), unemployed foreigners (`Alosaus`), foreign population (`HWBinsA`), foreigners subjected to social insurance contribution (`sozvpflBeschAus`), immigration (`WanZuzus`) and residential building stock with 10 and more apartments (`GebBest10ins`). Districts of the inner city cluster are separable from other districts with respect to their similar high relative frequencies of these variables. Before interpreting these variables in a social context one has to keep in mind that all remaining variables were selected as representatives of their group.

The eastern circle is associated to building stock variables spanning the sixties and eighties as years of construction (`BJ58bis62`, `BJ63bis72` and `BJ82bis92`). Districts in this cluster have similar high relative frequencies of postwar building stock. The cluster of northern suburbs shows high frequencies of very old houses (`BJbis1900`) and tractors (`Zugmaschine`).

The small cluster is positively associated with the variable `GenNeuins`, that is total building permits for residential buildings with living space. The interpretation of building permits with respect to social environment is difficult: High numbers of new buildings either represent development areas that usually attract young families or represent dense housing areas of lower social level.

4 Conclusion

After necessary preprocessing steps, the remaining data set was submitted to correspondence analysis. Examination of the first two dimensions after mapping already suggested that the districts do not form a homogeneous entity. Considering the first three new dimensions a stable partition into five clusters was found: The inner city and adjacent districts, a western and an eastern circle, a cluster of northern suburbs and industrial areas and a small cluster scattered over the outer city area.

For four clusters we identified highly associated variables which represent populational structure and building stocks. However, the interpretation regarding the social environment is restricted to the included variables. The western city cluster is not particularly associated with any variable and represents the cluster of remaining districts after separation of other clusters.

The analysis reflects Dortmund's social conditions in 2002. To monitor changes in social or residential conditions, next year's data can be submitted to the same procedures with reduced variables as in 2002. Few changes in 2003 will probably result in a similar partition, whereas dramatic changes may lead to an altered number of clusters and other associated variables.

References

HARTIGAN, J. A. and WONG, M. A. (1979): Algorithm AS 136: a k-means clustering algorithm, *Applied Statistics, 28, 100–108.*

IHAKA, R. and GENTLEMAN, R. (1996): R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics 5(3), 299–314.* Software: http://www.r-project.org/.

MANLY, B.F.J. (2001): *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall, London.

MIRKIN, B. (1996): *Mathematical Classification and Clustering*, Kluwer Academic Publishers, Dordrecht.

NAKAYAMA, T. (2001): Tests for redundancy of some variables in correspondence analysis, *Hiroshima Mathematical Journal, 31, 1–34.*