Are you moving predictably?

Miriam Wagner, Martin Breuer, Moritz Werthebach, Timo Bergerbusch, and Walter Schikowski

RWTH Aachen, Templergraben 55, 52062 Aachen, Germany

Abstract. We analyze movements in the urban environment of the Columbian city Medellín. Each movement is given as spatiotemporal pattern with additional information about the reason of movement, means of transportation and the corresponding person concerning the socio-economic status (strata), the age and gender. The objective of the paper is twofold: first we try to find groups of movement patterns and see, whether they correspond to the socio-economic status, with this information, since we have the economic status of each person in the data set, we will apply supervised learning to classify patterns into socio-economic status. Neither decision trees nor neural networks match our performance expectations, which leads to the conclusion that we need more data and information about the data in order to find a proper social stratification and to predict the given socio-economic status accurately, if even possible.

Keywords: Data Mining · Clustering · Rapidminer · Cluster · Neural Nets

1 Introduction

Within the time of Industry 4.0 and various data sources the question arises, if one can define who we are by the collected data. In particular, is it possible to determine the wealth of a person, only given movements of one single day? For this we considered the dataset stated in [1]. There we have a set of 124979 rows of movement data of various persons from the Columbian town Medellín, which is the second largest Colombian town with an estimated population of 2.5 million as of 2017 [5] . All the data was collected at a single day, with possibly multiple entries referring to the same person.
The data entries consist of data about the movement, like endpoints, length, and also some meta parameters like gender, age or the so-called strata of the person. The strata defines the socio-economic group, reflecting the affluence and therefore impose the ancillary costs. Those costs are defined by Colombians laws, which classifies households to regulate the access to public utility services, having as a result six socio-economic stratas [1].
Our goal is to ascertain, if there is a correlation between the movements and the strata, in order to be able to reproduce or predict the strata based on the movements. This is done in two steps:

First, by clustering using the k-means algorithm, considering different distance measures, with optionally including the principal component analysis (PCA). And second, a decision tree and neural net by training and testing.

2 Methods

2.1 Preprocessing

In order to classify the given data into smaller test sets or mask different aspects, we have to perform some analysis.
We observe that even though we have 124979 individual lines defining a movement, there is one line encounter a `NotANumber`-exception and therefore gets excluded from further usage.
We provide the `testDataGenerator` python script. Through flags and input arguments, the script is able to create all test sets used by our clustering and neural net approaches.
We observe the following distribution over the whole dataset:

strata	1	2	3	4	5	6	Σ
abs	6963	52265	49404	8772	5536	2038	124978
%	5.57	41.82	39.53	7.02	4.43	1.63	100

There is an upper bound on equal distribution through strata 6. It has at most 2038 individual elements. Furthermore, we have to make sure that two different data points, which belong to the very same person, are assigned to the same cluster. To do so, we compute the value `ID`, which identifies each person and can be used to combine movements that are considered to be from the same person. I.e. two movements correspond with the very same person, if and only if they are consecutive in the original dataset and have the same strata, age and gender. This approach is taken since the surveys are concatenated sequentially and it is unlikely, that multiple consecutive movements with same strata, age, gender belong to two different persons.

strata	1	2	3	4	5	6	Σ
abs	3153	23367	21418	3497	2083	595	54113
%	5.83	43.18	39.58	6.46	3.85	1.1	100

Following, we introduce vectors representing single persons. Since strata 6 is the smallest strata with 595 persons, it limits the size of an equally distributed dataset, where each data point coincides with one person.

Moreover, we define aggregated strata as the combination of strata 1 and 2 to "strata_1", strata 3 and 4 to "strata_2" and strata 5 and 6 to "strata_3".
Stratified person data
As stated before, additionally of simple IDs for every person we expand the parsing by using a data encapsulating in a class called `Person`. This class stores the ID, the parameters defining a person, and all movements from that person. Then we are able to compute the following vector, with 850 entries, for further

usage, that combines all movements of the person:

$$\underbrace{\#o_1, \ldots, \#o_{413}, \#d_1, \ldots, \#d_{413}}_{2 \cdot 413}, \underbrace{AM, MD, PM, MN}_{4}, \underbrace{\#r_1, \ldots, \#r_7}_{7},$$
$$\underbrace{\#MoT_1, \ldots, \#MoT_7}_{7}, \underbrace{S_{Dest}, S_{Dist}, G, A, strata, strataGrouped}_{6}$$

with the following abbreviations ($1 \leq i \leq 413$, $1 \leq j \leq 7$):

o_i:	the i-th origin data point	MoT_j:	the j-th mean of transportation
d_i:	the i-th destination data point	S_{Dest}:	sum of all durations
AM:	movements at time stamp AM	S_{Dist}:	sum of all distances
MD:	movements at time stamp MD	G:	the gender
PM:	movements at time stamp PM	A:	the age
MN:	movements at time stamp MN	$strata$:	the strata (used for comparison)
r_j:	the j-th reason	$strataGrouped$:	the aggregated stratas

2.2 Unsupervised learning

Clustering the data is in the field of *Data Mining*, *Cluster Analysis* or *Clustering* a process of grouping data objects from a dataset into multiple groups. The essential criterion, for the quality of the clustering, is *similarity*, such that data objects are similar to other objects in the same cluster and dissimilar to objects from other clusters.
In the scope of this work, we decided to use the well-known partitioning method k-means. In general, given n data objects, partitioning methods distribute the data objects into $k \in \mathbb{N}$ clusters with $k \leq n$, using a distance measure to evaluate the respective similarity. Note that the number k of clusters has to be chosen manually a-priori and given to the partitioning process.

k-means is a *centroid based technique*, meaning that each cluster is represented by a data point, which at the same time is the centre of the cluster. A distance measure is then used to assign every remaining data object from the data set to the best fitting cluster. This is done according to its similarity to the centre of this cluster and its dissimilarity to the centres of any other cluster.
The data objects within a dataset are considered to reside in an euclidean space. Thus, the euclidean distance is used to calculate a score for the similarity of two data points. When using k-means, the quality of a cluster C_i, $i \in [k]$, can be evaluated by computing the sum of squared distances between all data points p in the object space and the centroids $c_i \in C_i$ of every cluster. This method is known as *within-cluster variation* [4] and defined as follows:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2, \tag{1}$$

Given k, the first step of k-Means is to select k random points as cluster centres. Those do not have to be actual data points. In order to achieve this, each cluster centroid is redefined as the mean of all objects within that cluster. By considering the updated centroids, every data point is reassigned to the now best fitting cluster. This iterative process will continue until no better clustering can be found or a maximal chosen step size is reached.

As a tool we consider RapidMiner, since it has many modules already efficiently implemented and is easy to adapt.

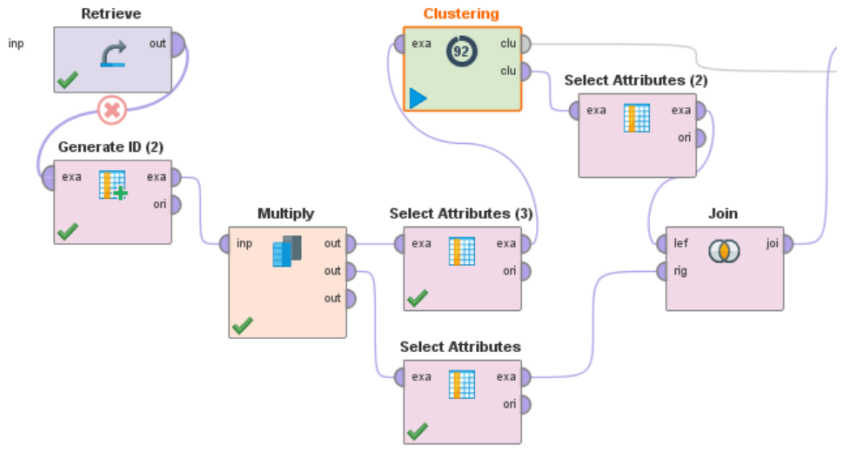


Fig. 2: Process of k-means clustering

The Process (c.f. Figure 2) contains the following steps:

Retrieve : imports the data into the process
Generate ID : creates an ID so we can make the comparison step at the end through joining the sets based on the ID
Multiply : creates two identical data sets
Select Attributes : removes the strata before the clustering step, everything except cluster and ID after the clustering and just keeps ID and strata for the join step
Clustering : runs the k-means clustering algorithm. The number of clusters has to be fixed
Join : For comparing the clustering result and the strata, we join the two filtered data sets on the ID

In the clustering block, we can choose between different distance measures and maximal step numbers. We keep the default configuration and choose the mixed euclidean distance measure.

Gower distance, in contrast to the popular distance measures, can also handle the mixed data within the given dataset. The Gower distance measure distinguishes between three types of variables: binary, categorical and numerical. The distance can be calculated as follows:
Given two data points x and y, each form a tuple of v variables of arbitrary type, the similarity coefficient is given by:

$$S_{xy} = \sum_{k=1}^{v} s_{xy,k} / \sum_{k=1}^{v} \delta_{xy,k} \tag{2}$$

where $s_{xy,k}$ denotes a *score for the similarity* of the two variables at the k-th entry in the data points x and y. Note, that the definition of the score depends on the type of the variable, as defined below. In the divisor, $\delta_{xy,k}$ basically represents the possibility of comparing the two variables at index k, such that it evaluates to 1, if the two variables are comparable and to 0 if not. For example, variables are not comparable, if values are undefined in the data points or the variable types do not match. Within this work, the dataset is complete, therefore, $\sum_{k=1}^{v} \delta_{xy,k} = v$. Thus, the similarity coefficient in equation (2) can be interpreted as the average value of all similarity scores. With respect to the variable type, the similarity score $s_{xy,k}$ is defined as follows:

Binary: The score for binary variables is basically the result of an logical conjunction operation. As pointed out above, 0 values are considered to be not a match, thus not even considered to be comparable. Hence, the values result as in the table

i	1 1 0 0
j	1 0 1 0
$s_{xy,k}$	1 0 0 0
$\delta_{xy,k}$	1 1 1 0

Categorical: The similarity score of categorical variables is 1, if the variables are completely identical in x and y, and 0, if they differ.
Numerical: For numerical variables, the similarity score is calculated by

$$s_{xy,k} = 1 - \frac{|x_k - y_k|}{range(k)}$$

where $range(k)$ is the total range of values, that the numerical attribute at index k can accept. This can be a global maximum of acceptable values for attribute k, or chosen on the basis of the dataset.

Principal component analysis (PCA) uses an orthogonal transformation to convert a dataset, with correlated data, into a dataset without correlations and decreased dimension. This is a helpful technique to get a better understanding of the data complexity, as well as decreasing the time consumption in further steps.

The first component for the PCA transformation, the weight matrix, has to satisfy the following:

$$\begin{aligned} w_1 &= \text{argmax}_{\|w\|=1}\{\sum_i (t_1)_i^2\} \\ &= \text{argmax}_{\|w\|=1}\{\sum_i (x_i w)^2\} \end{aligned}$$

which is achieved by the corresponding eigenvector.

The next components need to fullfill:

$$\begin{aligned} \hat{X}_k &= X - \sum_{s=1}^{k-1} X w_s w_s^T \\ w_k &= \text{argmax}_{\|w\|=1}\{\|\hat{X}_k w\|^2\} \\ &= \text{argmax}_{\|w\|=1}\{\frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w}\} \end{aligned}$$

Based on the weight matrix a mapping is possible from the original space to a new space in which the dataset is uncorrelated.

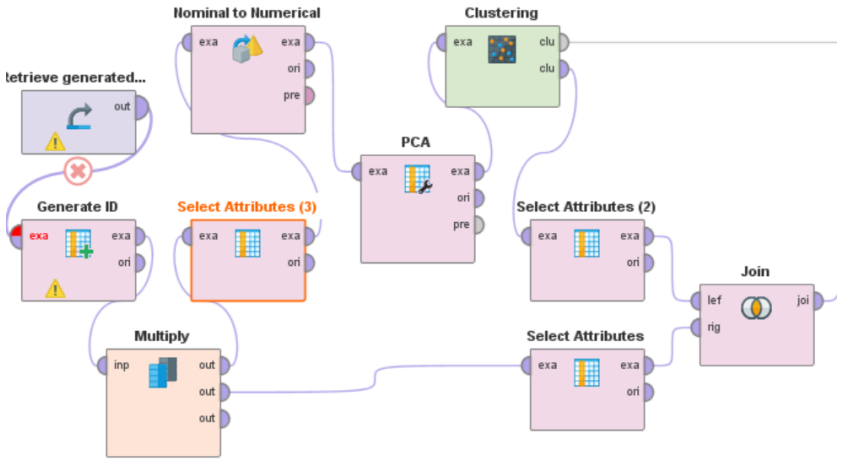


Fig. 3: RapidMiner process for PCA and k-means

The Process does the following steps, visible in Figure 3:

Retrieve : imports the data in the process.
Generate ID : creates an idea for the join later
Select Attribute : gives the chosen attributes back
Nominal to Numerical : changes the nominal data to numerical data, so that we can apply PCA in the next step
PCA : applies the PCA reduction to the data threshold 0.95
Clustering : clusters the given data
Join : joins the clustering outcome and the strata by the ID

2.3 Supervised learning

Decision Trees are a good manner to figure out, which parts of the data set have the most influence on the decision. Therefore, labelled data is needed, which we have given with the strata. We again used RapidMiner for building trees, based on different data sets.

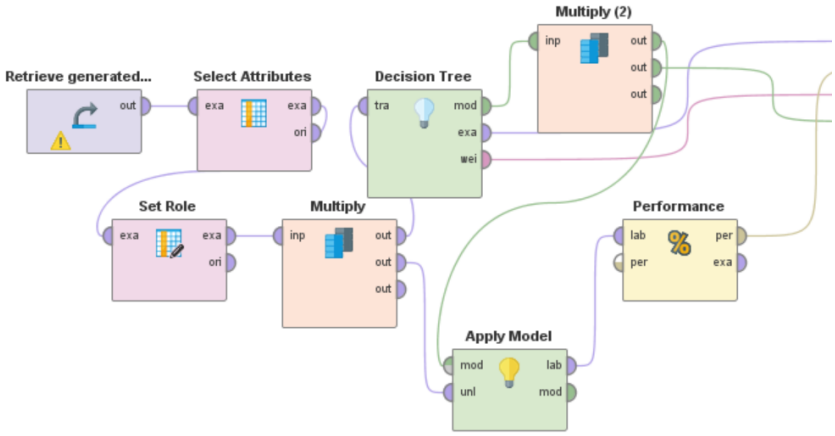


Fig. 4: Process for decision trees in RapidMiner

RapidMiner does the following steps, illustrated in Figure 4:
Retrieve : includes the dataset
Select Attributes : makes it possible to have a look at grouped strata or normal strata
Set Role : gives strata the label role, so that the decision tree has those as leafs
Multiply : clones the data set
Decision Tree : creates the decision tree
Apply Model : creates the labelled data set for the **Performance** step
Performance : gives the performance result of the created model
We tried several configurations regarding the splitting criterion and confidence threshold, but the results were similar.

Neural Net are based on biology. In detail, based on the neurons and synapses within the human brain. There a neuron fires to any further connected neurons, if a weighted sum minus a bias value is above a certain threshold. Through iteration-wise computing the result and comparing it to the expected outcome, called label, and the use of various machine learning techniques the net is able to adjust the weights and biases [6].

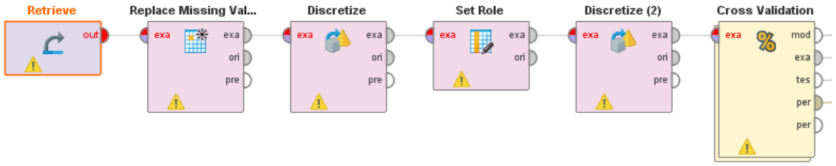


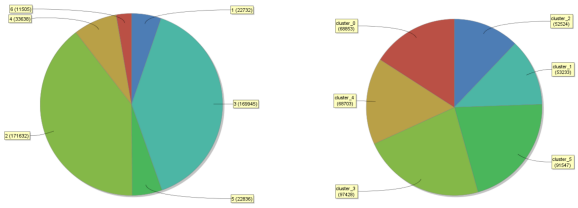
Fig. 5: RapidMiner process for neural nets

Retrieve : includes the dataset
Replace Missing Values : ensure applicable data
Discretize : translates numerical to nominal data
Set Role : gives strata the label role, so that the decision tree has those as leafs
Cross Validation : models the neural network

3 Results

3.1 Natural clusters

We look for clusters and compare those with the strata. Also we try to observe, if the possibly found clusters have special properties.
In the first step we try to cluster the *Original Data* in *6 cluster*. Therefore we retrieve the original data set in RapidMiner and choose k as 6.

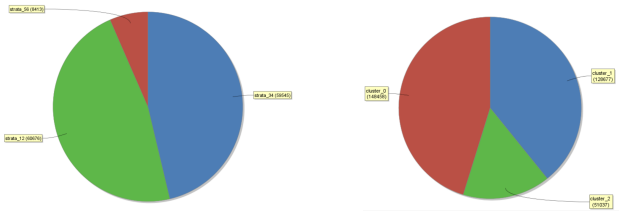


(a) Strata (b) Cluster

Fig. 6: Distribution of original data

Figure 6 is the result of the first try represented as pie charts. Figure 6a shows the strata distribution and Figure 6b the resulted cluster distribution. It can already been seen, that the distributions are not similar. For more steps the outcome is analogous.

The next idea is: are 6 cluster too specific? Based on the grouped strata we look at 3 clusters.



(a) Strata (b) Cluster

Fig. 7: Distribution of original data for just 3 clusters

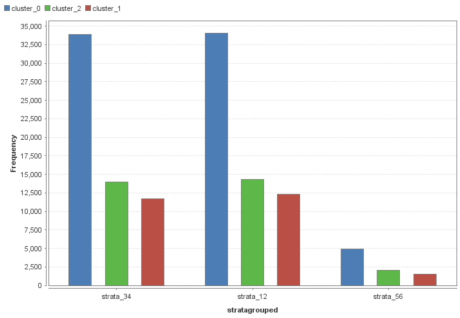
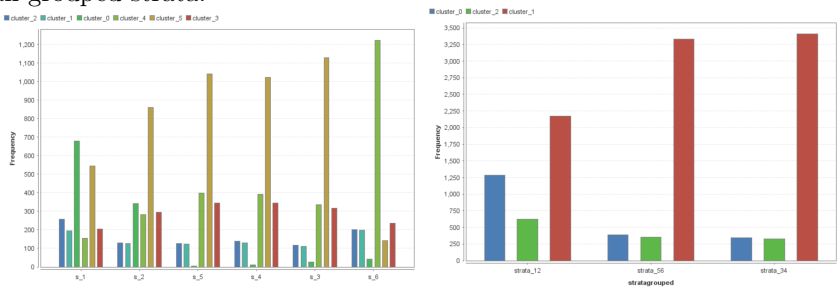


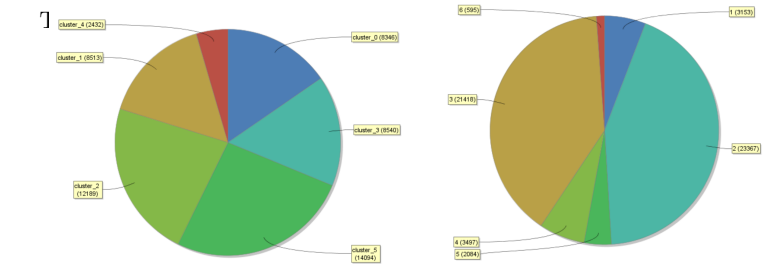
Fig. 8: Distribution of the clusters within the grouped strata

Figure 7 looks promising, so we had a look at the cluster distribution in the 3 grouped stratas, visualized in Figure 8. It can be seen, that all clusters appear in all grouped strata.



(a) For 6 clusters (b) For 3 clusters

Fig. 9: Distribution of the clusters in between the strata dataset size 2038



(a) Strata (b) Cluster

Fig. 10: Distribution of stratified person data

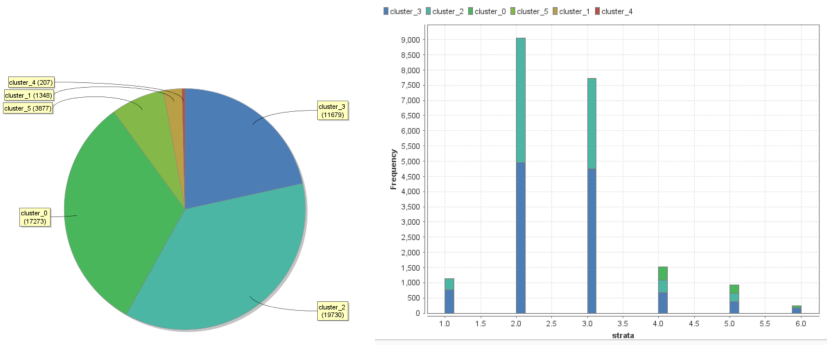


Fig. 11: 1000 steps clustering in 6 clusters on maximal stratified person dataset

Different dataset sizes and equal distribution of the stratas do not effect the result. In Figure 9 the clustering results for the maximal sized dataset can be seen. Once aggregated and once normal. Again we could not observe a significant correlation.

After those not convincing results, we apply the process to the *stratified person data*.

The results for the whole data set without equalization are shown in Figure 10. We change the number of steps to 1000 for comparison and the result, illustrated in Figure 11, lets us assume, that 3 clusters would fit better.

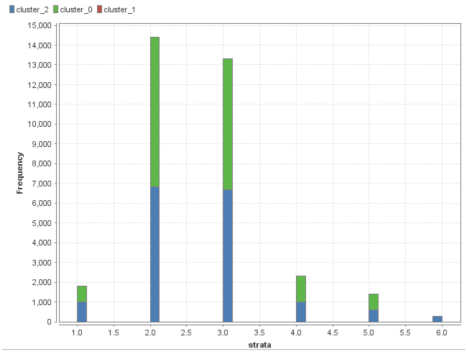
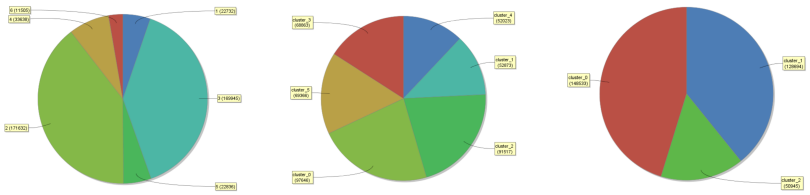


Fig. 12: 1000 steps clustering in 3 clusters on maximal stratified person dataset

Figure 12 shows clearly, that again no correlation can be found. Applying the procedure to the different datasets, we generate similar results.
To save time, we apply the PCA process on the *original data* without strata and a variance threshold of 0.95. We get two attributes, so we can assume, that the data is strongly correlated. As expected the computing time decreases a lot for the whole process of clustering.



(a) Strata (b) 6 cluster (c) 3 cluster

Fig. 13: Distributions of the original data. Clustering with the PCA process

For the clustering with the preprocessing step PCA, we get distributions, which can be seen in Figure 13, similar to the distributions we get from the k-means process with the same configurations. For a clustering of 3 clusters we again get results, that are not meaningful (Figure 13c).
Using different dataset results in comparable unrewarding results.

3.2 Supervised learning

Decision Trees

In the first step we apply the process on the *original data set* and the resulting tree is just the leaf 'strata 2'. So we try it with different other data sets and the best result we get is for *stratified person data* equally distributed and just 200 data entries in every strata.

accuracy: 99.33%

	true s_1	true s_2	true s_5	true s_4	true s_3	class precision
pred. s_1	23	2	0	0	0	92.00%
pred. s_2	0	175	0	0	0	100.00%
pred. s_5	0	0	199	0	0	100.00%
pred. s_4	0	0	1	199	1	99.00%
pred. s_3	0	0	0	0	0	0.00%
class recall	100.00%	98.87%	99.50%	100.00%	0.00%	

Fig. 14: Performance for stratified person data with 200 in every strata

accuracy: 30.17%

	true s_1	true s_2	true s_5	true s_4	true s_3	true s_6	class precisi...
pred. s_1	561	554	304	399	474	79	23.66%
pred. s_2	0	0	0	0	0	0	0.00%
pred. s_5	0	0	0	0	0	0	0.00%
pred. s_4	0	0	0	0	0	0	0.00%
pred. s_3	0	0	0	0	0	0	0.00%
pred. s_6	34	41	291	196	121	516	43.04%
class recall	94.29%	0.00%	0.00%	0.00%	0.00%	86.72%	

Fig. 15: Performance for stratified person data with 585 in every strata

In Figure 14 the best and in Figure 15 the worst outcome can be seen for 6 clusters. For all other configurations the outcome is similar, so just small data sets have an acceptable result and larger datasets have no sufficient results for our question.

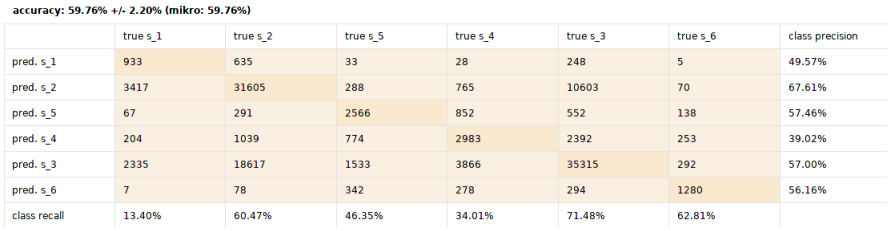
Neural Net

In this section we want to further improve the accuracy of predicting the strata of one person by using neural networks. Like in the previous section, where we used decision trees, we need ground truth to be able to train the neural networks. For all the neural net computations we consider person vector data sets of different sizes (c.f. Section 2.1).

We do this, because results on the normal datasets had an unacceptable performance, since only single movements and not complete paths of individuals are considered. An example training and performance measure is given in Figure 16, where unprocessed data is used. The performance is measured using 10-fold cross validation, i.e. the data is split into 10 subsets, where in each iteration exactly one data set is used as test set and the other 9 as training set. The average value of those accuracy values leads to the total accuracy of the neural net.

In the following we consider 3 neural nets $\mathcal{N}_1, \mathcal{N}_2$ and $\mathcal{N}_3$, all having 4 hidden layers, 50 epochs and 10 iterations. We call the aggregated strata sets $\mathcal{N}_i^\star$, for $i \in \{5, 10, 20\}$ denoting the number of neurons. This builds a superset of the original stratas and since the stratas themselves are logically connected, this

task should be easier to fulfill.



accuracy: 59.76% +/- 2.20% (mikro: 59.76%)

	true s_1	true s_2	true s_5	true s_4	true s_3	true s_6	class precision
pred. s_1	933	635	33	28	248	5	49.57%
pred. s_2	3417	31605	288	765	10603	70	67.61%
pred. s_5	67	291	2566	852	552	138	57.46%
pred. s_4	204	1039	774	2983	2392	253	39.02%
pred. s_3	2335	18617	1533	3866	35315	292	57.00%
pred. s_6	7	78	342	278	294	1280	56.16%
class recall	13.40%	60.47%	46.35%	34.01%	71.48%	62.81%	

Fig. 16: An example of a neural net trained without person vector data.

For each neural net we are using equally distributed data sets with 100, 200 and the maximal amount of 595 individuals per strata. For every neural net and every set size, we perform 5 independent runs and calculate the average over those accuracy values in order to have a sophisticated, comparable statement.

Name	# Neurons	AG	Set size 100	200	595
$\mathcal{N}_5$	5	✗	60.03	59.92	60.18
$\mathcal{N}_5^\star$	5	✓	87.6	89.7	71.05
$\mathcal{N}_{10}$	10	✗	75.83	73.54	69.56
$\mathcal{N}_{10}^\star$	10	✓	92.93	93.48	74.58
$\mathcal{N}_{20}$	20	✗	75.45	71.14	61.87
$\mathcal{N}_{20}^\star$	20	✓	92.87	94.4	78.32

Fig. 17: The accuracy values of the neural nets

The size of larger nets, in terms of neurons, is counter-productive, since, if we take 50 neurons per layer, we have $14 \cdot 50^4 \cdot 6 < 849 \cdot 50^4 \cdot 6 \approxeq 31.837.500.000$ synapses for which the input dataset would be too small to perform sufficient training.

4 Conclusion

As an overall result we observe, that we can not determine the strata based on the information we have. Using various datasets, we witness, that having equally distributed datasets leads to an overall higher accuracy. This rules out the bias observed in the original data, where strata 2 and 3 are very dominant (c.f. Section 2.1), but also reduces the set size from originally 124978 to $6 * 2038 = 12228$ entries equally distributed over all stratas.
During clustering we observe, that using a different distance measure formula

would not lead to a huge difference. Also, we detect that reducing the number of clusters leads to better results, but decreases the significance of the statement, that could be made.
Using a different neural net architecture will most likely not chance the accuracy value drastically. As stated, more complex nets need more training data, which is restricted for the reasons mentioned above.
What we can observe is, that using the stratified vectors, we are able to increase the performance and accuracy, but still have no sufficient predictions. Therefore we see: the more data we have, the higher the variance within the stratas is. We cannot draw lines to distinguish between different stratas. Data points, which were outliers in the smaller sets, are now not considered to be outliers, because many others have the same characteristics. So the lines, we were able to draw for smaller datasets, are blurring.

5 Discussion

Regarding the results, of not predicting the strata and also not observing any meaningful clusters throughout the data in given and stratified form, we come uo with various aspects having an impact on the data. We state 4 main influencing aspects, explaining the variance throughout the data.

5.1 Representativity of the day

The day the data was collected on (c.f. Section 1) is not mentioned. Therefore we do not know, if it is representative. The people, asked to enter their movement, could had an exceptional day, like a vacation day, a doctors appointment or a broken car and therefore behave different from an usual day.
Also the day itself is not mentioned, so we cannot determine, if it was even a working day or a weekend day. This influences the behaviour drastically.

5.2 People living over/under standards

There are a lot of people spending more or less money than they actually have. So for example there are people not earning a lot of money, but still having a car. Or people earning a lot, but spend it just on holidays or save it for bad times.

5.3 Individuality of lifestyle

Obviously, people are individual in their way of life. So there are people, with enough money to buy for example a car, but refuse to do so in order to reduce CO_2 emission, or simply dislike driving. On the other end of the spectrum, some people, who have little money, still own and drive a car daily in order to go to work, since they cannot afford an other apartment.
One can think of multitude scenarios of people behaving different caused by their individuality.

5.4 Inverse behavior in a strata

As an example, take a look at the strata 6, which denotes the richest people considered. Their we have inter alia two groups:

1. Group: Hard working people, laboring 60+ hours per week to earn money. One can imagine that they have to move quite a lot since they are always busy.
2. Group: People, who are rich by birth, which do not have the necessity to work or move at all. They could possibly stay home and do not have to leave at all.

Those two groups are completely opposite, but still belong to the same strata. Now also considering strata 1, we can find the exact same movement behaviour in there as well. Within the poorest there are people that are wandering around the town via feet or bike, since they have no objective, like work. Also there are people, who do not move at all, because of the same reason.
So we can see that based on the information we have, it is very unlikely to have a clear correlation between the given data and the strata. However, given more information about the previously mentioned aspects and more entries a correlation can not be ruled out in general.
Future research could include a component analysis of the decision tree and neural network in order to ascertain the influencing parts and therefore improve the data gathering. Also one could use different network visualization tools in order to infer patterns. An example would be to take the origin (destination) sectors as nodes, an edge, if there is a movement between, directed or undirected, and a different thickness or color gradient, based on the number of times this edge being taken.

Overall we can conclude that with the data we get, we are not able to find a correlation.

References

1. Lotero, Laura, et al. "Rich do not rise early: spatio-temporal patterns in the mobility networks of different socio-economic classes." Royal Society open science 3.10 (2016): 150654.
2. Hudson, Rex A. Colombia: A country study. Government Printing Office, 2010.
3. Gower, John C, *A general coefficient of similarity and some of its properties*, Biometrics, pp. 857-871, 1971.
4. Han, Jiawei and Pei, Jian and Kamber, Micheline, *Data mining: concepts and techniques*, Elsevier, pp. 444-454, 2011.
5. Estimates and projections of the total national, departmental and municipal population by area 1985-2020 (XLS). NADS. Retrieved 1 September 2014.
6. Nasrabadi, Nasser M. "Pattern recognition and machine learning." Journal of electronic imaging 16.4 (2007): 049901.