International Journal of Computer Vision and Signal Processing, 1(1), 16-26(2020)

Data Driven Analysis of the Behaviour of Elderly People Using k-Means and Home Automation and Power Consumption Sensors

Björn Friedrich*, Enno-Edzard Steen

Department of Health Services Research
Carl von Ossietzky University of Oldenburg, Germany

Hirohiko Suwa

Graduate School of Information Science, Ubiquitous Computing Systems
Nara Institute of Science and Technology, Japan

Andreas Hein

Department of Health Services Research
Carl von Ossietzky University of Oldenburg, Germany

Keiichi Yasumoto

Graduate School of Information Science, Ubiquitous Computing Systems
Nara Institute of Science and Technology, Japan

Abstract

Information about the mental and physical conditions of elderly people are essential to assess their ability to live alone in their own homes. Usually, those information are collected using questionnaires and geriatrics assessments. However, both methods have their limitations. People might not answer honestly to personal questions and geriatrics assessments are only measuring the capacity at a specific point in time. Moreover, questionnaires and assessments are limited in catching variability. Elderly people which have similar scores in questionnaires and assessment can still be very different in terms of mental and physical conditions. To get a distinguished impression of the condition long-term monitoring is needed. In this article we show that the behaviour of elderly people is so distinguished, that they can be identified by using k-Means clustering on a dataset comprised of motion sensor data and power consumption sensor data. Moreover, we show that the results of a combined dataset is different to considering each participant separately. We applied the algorithm to three participants of a real-world study. Even though two of the participants have similar questionnaire and assessment scores, they can be clearly distinguished from each other as well as from the third participant who has different scores.
Contribution of the Paper: Showing the difference of elderly people with similar scores in standardised geriatrics assessments and questionnaires.

Keywords: k-Means, elderly people, behaviour, comparison

IJCVSP

ISSN: 2186-1390 (Online)
http://cennser.org/IJCVSP

Article History:
Received: 29 January 2020
Revised: 9 June 2020
Accepted: 30 June 2020
Published Online: 1 July 2020

1. INTRODUCTION

The demographic change is one of the challenges nowadays. Especially in the industrial nations the older part of the population is growing and there is no change in the near future. As people become older the more assistance

*Corresponding author
Email addresses: bjoern.friedrich@uni-oldenburg.de (Björn Friedrich), enno-edzard.steen@uni-oldenburg.de (Enno-Edzard Steen), h-suwa@is.naist.jp (Hirohiko Suwa), andreas.hein@uni-oldenburg.de (Andreas Hein), yasumoto@is.naist.jp (Keiichi Yasumoto)

they need. Assistance could be a nurse visiting on a regular basis or go as far as moving to a nursing home. Moving to a nursing home cuts down the basic psychological needs of humans, especially self-determination and independence. Limitation of self-determination and independence leads to a decrease in quality of life. The mental and physical conditions are indicators for the ability of an elderly person living independently in his or her own home. The mental and physical conditions of elderly people are usually assessed by standardised geriatrics questionnaires and assessments. Both methods have their own advantages and disadvantages. Questionnaires are easy to use even for untrained people. However, questionnaires have several disadvantages. The most important one is, that the results depend on the answers of the interviewee. Considering personal questions the interviewee does not want to reveal the truth or is ashamed of the truth. Taking the Barthel Scale as an example, it has two items *presence or absence of fecal incontinence* and *help needed with toilet use* amongst other things [1]. Those kind of questions are very personal and people might not want to answer honestly. Questionnaires are also limited in capturing variability, because they have a fixed scale [2, 3].

Standardised geriatrics assessments are validated ways to assess the mobility and the functionality of the elderly people. These assessments are performed under the supervision of a medical professional. Even though an assessment gives valuable information about the physical state, it can only measure a capacity at the point in time when it is performed. Moreover, people tend to give there best in test situations [4]. Taking the difference of the self-selected gait speed and the maximum gait speed as an example, the meta-study in [5] found a difference of $0.29\frac{m}{s}$ between the self-selected gait speed and the maximum gait speed. The average self-selected gait speed was $0.58\frac{m}{s}$ and the average maximum gait speed $0.89\frac{m}{s}$. The results of a field study with elderly people revealed similar findings. The study found a mean self-selected gait speed of $1.07\frac{m}{s}$ and a mean maximum gait speed of $1.41\frac{m}{s}$ with a difference of $0.34\frac{m}{s}$ [6]. The self-selected gait speed and the maximum gait speed are dependent on several factors and so is the difference.

A promising approach to capture the variability of elderly people is long-term monitoring of their behaviour. To investigate whether long-term monitoring can give a better impression of the conditions of elderly people than questionnaires and assessments, we use a dataset collected during a study realised by the Carl von Ossietzky University. The main advantage of that dataset is the combination and frequency of questionnaire results, assessment scores and ambient sensor data. We choose a set including study participants which are similar in terms of questionnaire results and assessment scores and including participants with different results and scores. Then we cluster the data of their ambient sensors and check whether the participants can be distinguished considering the clusters.

The outline of this paper are as follows. First, we are giving an overview of the state of the art for analysing behaviour of the elderly and available datasets. Afterwards, we introduce our methodology by explaining the data acquisition, the dataset generation process, the applied clustering algorithm and the used metrics. In Section 4 we show our results. The following section features the limitations of our approach and the dataset. After discussing our results we will present our conclusions and future work.

2. RELATED WORKS

Machine Learning algorithms are a common approach to analyse smart home data. Most research is focused on activities of daily livings (ADL). ADLs can be used as indicators for mild cognitive impairment and impending dementia. The research in [7] is focused on ADL and mild cognitive impairment. The aim was to detect changes in behaviour on monthly base. Their approach was clustering power consumption data to identify changes in behaviour and frequency of ADLs. The method has been evaluated with a 7 months real-world dataset with two healthy participants. Both participants were reported to be different in behaviour. One participant followed a very structured lifestyle, whereas the other followed an unstructured lifestyle. For both participants the results of standardised questionnaires were available as well. The method showed good results for the two participant. However, they did not quantify the difference between the participants using questionnaire results. Another approach considering ADLs used a combination of clustering techniques and machine learning algorithms for ADLs [8]. They used passive infrared sensors to model the different ADLs. Clustering algorithms were used to label the data and the machine learning algorithms were applied on the labeled dataset. The best machine learning algorithm acquired an F-measure of 71.33% for classifying ADLs. For the evaluation they used real-world dataset comprised of 10 participants 20 days each. The participant's average age was 48.8 y (min = 28, max = 79). For abnormal behaviour detection Lotfi et. al. introduced a method that is using dissimilarity measurements [9]. They were using occupancy sensors and binary dissimilarity measures. They were validating their method with two participants in a real-world study. One participant was considered as sick and to show abnormal behaviour in the first part of the study. In the second part after changing medication the behaviour changed and was considered as normal. The second participant was not considered as behaving abnormally.

The results of standardised and validated geriatrics questionnaires and assessments are well accepted and understood by physicists. It is not too far to seek to estimate the scores based on home automation sensor data. However, there are not many datasets combining questionnaire and assessment results with home automation sensor data. The research in [10, 11, 12] were done on the same dataset comprised of data from 40 elderly living in 38 smart homes

over 2 y and questionnaire and assessment results. The flats were equipped with passive infrared motion sensors amongst others. The inhabitants performed the Timed Up & Go, Arm Curl [13], and Digit Cancellation test [14] and answered the Repeatable Battery for the Assessment of Neuropsychological Status [15], Prospective and Retrospective Memory Questionnaire [16], Instrumented Activities of Daily Living Compensation Scale [17] biannually. In [10] the aim was to automatically assess the score an elderly person would have in an IADL-C questionnaire. They were comparing linear regression to Support Vector Regression models. The feature were sleep duration and frequency, total walked distance, and activities of daily living derived from PIR sensors amongst others. The algorithms classified with an F-Score of up to 0.92. The prediction of mobility assessment scores and cognitive assessment scores were investigated in [11]. They were able to predict the mobility (TUG) and cognitive (RBANS) scores with an accuracy of 72% and 76% respectively. Another work focused on the detecting impending dementia by behavioural changes [12]. As indicators mobility, cognition and mood scores were used. Machine learning algorithms were applied to the data to predict score changes. A similar goal was aimed for in [18]. The authors aimed to classify whether a person was suffering from mild cognitive impairment or was healthy. The features were derived from eight different activities of daily living. The ADLs were performed in a smart home and observed by door contact and motion sensors. The ADLs were performed by 263 participants in total. Of these participants 16 were diagnosed with dementia and 51 with mild cognitive impairment. The classification accuracy was g-mean of 0.73 for the two classes dementia and healthy.

2.1. Limitations of the State of the Art

The state of the art has several limitations, when it comes to the comparison of participants. The datasets used in [8, 9] neither contain questionnaire results nor assessment scores and they do investigate the difference between their study participants. The dataset in [7] contains results of questionnaires, but they are not taken into account. The definition of the difference between the two participants was based on the impression of a person, but not quantified. Regarding assessment scores and questionnaire results a comprehensive dataset was used in [10, 11, 12]. Even though they had the information, they did not investigate the different results, if they merge the dataset and considering each participant separately. In [18] the authors used the reliable change index for inter-subject standardisation, but did no further investigation with the difference of the participants.

In general there are two approaches in the current research. The first one is to merge the data of several participants to one large dataset and the second one is to consider each participant separately. For some approaches the choice is natural. However, to best of our knowledge there is no analysis of how the participants are different in behaviour depending on their assessment scores and questionnaire results and whether that difference must be considered when doing behavioural analysis.

3. PROPOSED METHOD

In this section we describe the study for collecting the data and the preprocessing for clustering. Afterwards, we explain the used algorithm, metrics and the reasons for using them.

3.1. Data Acquisition

The data has been collected during the OTAGO study realised by the University of Oldenburg under the ethic vote Drs.72/2014. The study started on the 1st of July in 2014 and finished at the 31st of December 2015. Twenty participants (17 female, 3 male) of an average age of 84.75y ($\pm 5.19y$) participated in the study. They were in pre-frail or frail condition. The planned duration of the study was 40 weeks for each participant. Due to drop out the average participation time was 36.5 weeks. At the beginning and every four weeks the standardised geriatrics assessments, Timed Up&Go, Short Physical Performance Battery, Barthel Index and Instrumental Activities of Daily Living among others were performed [1, 19, 20, 21, 22]. Due to sickness, visitors, public holidays etc. the average days between two assessments were 31.3 days ($\pm 5.3d$). Two participants died during the study and two participants performed the assessments ten times. For the remaining 16 participants eleven assessments are available.

In addition a multi-component sensor system has been installed in the flats of the participants and each participant got a GPS receiver and an Inertial Measurement Unit (IMU). The IMU was carried by the participant for one week after each assessment day. For the sake of easy integration wireless sensors were chosen. All sensors sent their data over the air to a base station. The sensor system was mainly comprised of home automation sensors and power sensors. A concussion sensor has been placed in the bed, since the used motion sensor was not sensitive enough to measure the small movements while sleeping. A switch with four keys has been installed next to the front door of the homes. The switch was used to indicate whether the person is alone in the flat or not. The participants have been instructed to press a key to make the system aware when another person enters the flat. When the person leaves the flat again or the participant comes home, another key had to be pressed to make the system aware that only one person is inside the flat. In Figure 1 a flat of one of the participants is shown.

3.2. Dataset

The data collected during the study described in Section 3.1 must be preprocessed for clustering. Clustering algorithms are sensitive to missing values. To select the optimal subset of participants we analysed the installed

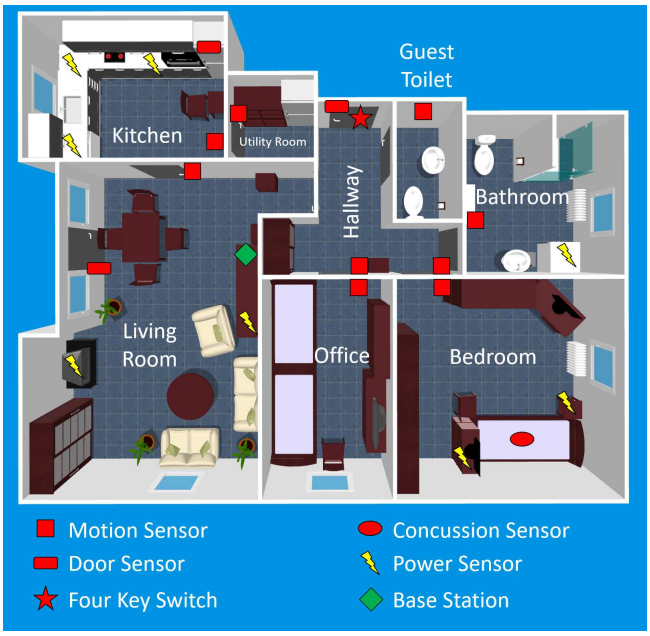

Figure 1: The layout of a flat of a participant in the OTAGO study. The positions of the sensors are marked with small symbols.

sensors for each participant and selected a subset so, that the amount of sensors and participants are maximal. The data of all participants has values for each sensor, so there are no missing values. This led to 14 sensors and three participants. For all participants Barthel Scale, IADL, Frailty Scale, TUG, and SPPB information are available.

The Barthel Scale was designed to evaluate the performance of ADLs [1]. The ten items are on an ordinal scale and the scores are 0, 5, 10, and 15 points for each item. 0 means the person is not able to perform that item and 15 means there are no impairments. The maximum score are 100 points. The Barthel Scale does not assess a person's ability to live alone, because the assessed aspects are limited to grooming, food intake and mobility. The questionnaire does not cover activities like cooking or cleaning. Therefore, other questionnaires like the Instrumental Activities of Daily Living (IADL) must be considered.

The IADL assesses eight different instrumental activities of daily living [22]. The eight items cover using a phone, doing groceries, cooking, housekeeping, washing, using means of transportation, taking medicine, and financial transactions. The interviewee can achieve 8 points in total and each of the eight items is scored with 0 or 1 point. An interviewee who achieved 8 points can perform all instrumental activities of daily living without any limitations, whereas an interviewee who achieved 0 points is completely dependent on other people. Combining the results of the Barthel Score and the IADL gives an overview about a persons ability living alone.

The Frailty Scale is an index for expressing the level of frailty of people older than 65 y [21]. The frailty index has seven levels of frailty. People with a frailty index 1 are considered as *Very fit* and belong to the most fit group for their age, whereas people with an index of 7 are considered as *Severely frail* and are completely dependent on others or terminally ill. A frailty index of 2, 3 means the person is well, with treated comorbid disease. The frailty scale must not be considered alone, but in combination with other questionnaires and assessments.

The TUG is one of the common assessments in geriatrics. The TUG test assess the mobility and the risk of falling. During the TUG test the participant starts sitting on a chair and is asked to stand up, walk 3 metres, turn around, return to the chair, and sit down again. The time needed to finish the procedure is used to derive the test score. The score ranges from 0 to 3 points, where 3 is the best score. A participant achieves 3 points if the time is less than 10 s. This participant does not have any impairments in mobility. If the time is longer than 30 s the participant gets 0 points and has strong mobility impairments. The higher the TUG test score, the better the mobility.

The SPPB test is assessing balance, gait speed and leg power [20]. To assess the balance, the participant is asked to place the feet in three different position and the time the participants can stand unsupported in that position is measured. To assess the gait speed the participant has to walk 2.44 m (8 ft) and the time needed is measured. To assess the leg power the participant is asked to sit on a chair and to stand up and sit down 5 times. The score is based on the elapsed time. The minimum score of the SPPB are 0 points and the maximum score are 12 points. The higher the score, the better the mobility. A participant with 0 points has severe mobility impairments, whereas a participant with 12 points does not have any mobility impairments.

To get a holistic view of the participants the three questionnaires Frailty Scale, Barthel Scale and IADL and the two geriatrics assessments TUG and SPPB are used for comparing and choosing the participants of our dataset. Participants 2 and 3 are very similar in terms of mobility scores, frailty score and independence. Participant 1 has lower Barthel Scale, TUG, SPPB and IADL scores. In general participants 2 and 3 are in better physical and cognitive condition than participant 1.

The door and motion sensors only have the values 0 and 1.

Table 1: Overview of the main characteristics of the three participants. The assessment and questionnaire scores are the average score over the time of the study. Participant 1 is highly dependent on other people and has severe mobility impairments. Participants 2 and 3 have slight mobility impairments and are able to perform activities of daily living on their own.

ID	Age (Sex)	Frailty Scale	Barthel Scale	IADL	TUG	SPPB
1	90 (m)	2.5	49.5	1.7	3	2.10
2	90 (f)	2.5	88.1	8.0	2.6	3.70
3	86 (f)	2.5	75.9	6.7	2.8	4.10

Where 0 is *no motion* or *door closed* and 1 *motion detected*

or *door open*. The values of the power consumption sensors are of a wider range. 0 denotes *no power* consumed and all values greater 0 are the *power consumption* in Watt. Since appliances have a great variety of power consumption the values are varying accordingly. Moreover, some appliances have standby power consumption. At first we deleted all values denoting standby state or turned off state. We assumed that the appliances are in standby mode most of the time. Hence, we defined the standby power consumption by analysing the frequency of power consumption values. It turned out that a threshold of 1 Watt is suitable for all appliances. Then we defined a vector for each measured value of each sensor. The vector has the dimension 1×14. Therefore, one sample contains 14 features, where all features are 0 except for the one of the sensor value.
The motion sensors have a sampling frequency of $7.5Hz$. Once a motion is detected the sensor is not sensing for the next $8s$. Hence, we aggregated the samples over a time window of $60s$ and computed the sum of each feature. Since we are only interested in the presence of a sensor value we replaced all values greater 0 by 1. Moreover, we overcome the difference in measurement frequency of the power consumption sensors and the motion sensors by using this approach. This also makes scaling redundant and eliminates the dominance of certain features. For example the motion sensors only have values of 0 and 1, but the power consumption sensors from 0 up to 100. Applying the Euclidean distance would lead to a dominance of the power consumption sensors, since the values are much larger than the values of the motion sensors. The same holds for the appliances. We applied this process to all three participants separately, because the values might have been recorded at the same time and at different places. Table 2 shows five samples of our final dataset and the variance of each feature. Due to preprocessing the variances are very small, but we still see that the variance of the feature TV is much larger than the variance of the feature *Lamp bedroom*. After preprocessing we had 106.299 samples in total and 35.433 samples for each participant.

Table 2: Four randomly selected samples of the dataset. One sample is comprised of 0's and 1's. One sample covers an interval of $60s$. A 0 means that there is no sensor event for the sensor in this sample or time interval and a 1 means that there is a sensor event for the sensor.

Feature / Sample	1	2	3	4
Bed	0	0	0	0
Fridge	0	0	0	0
Living room	0	0	0	0
Toilet	0	0	0	0
Front door	0	0	0	0
Rear door	0	0	0	0
Bathroom	0	0	0	0
Hallway	0	0	0	0
Bedroom	0	0	0	0
Kitchen	0	0	1	0
Lamp bedroom	1	0	0	0
Lamp living room	0	0	1	0
Kettle	1	0	0	1
TV	1	1	1	0

3.3. Clustering Algorithm

We used the k-Means clustering algorithm [23, 24]. The advantages of k-Means are its simplicity, its speed and the most important is the convergence [25, 26]. Moreover, we can save intermediate results to track the progress of the algorithm.
One disadvantage is that the number of clusters k is an input parameter and thus must be known beforehand. We overcame this disadvantage by running the algorithm with different ks. Another disadvantage is the random initialisation of the cluster centers. This may lead to different local minima. To face this problem we were running the algorithm for each k several times and pick the best result compared to our metrics. Like the most other clustering algorithms k-Means cannot handle missing values.
k-Means is designed to minimize the squared distances from the cluster centers to the data. It can be considered as an optimisation problem of the function

$$\arg\min_{S} \sum_{i=1}^{n} \sum_{x \in X} \|x - \mu_i\|^2 \tag{1}$$

where S is the set of all samples, n the number of clusters, x a sample of the subset $X \in S$ and μ the cluster center. Usually, the Euclidean metric is used, but any function that holds the characteristics of a metric is applicable. From equation (1) two of our three used metrics can be easily derived. The basic idea of all of the three metrics is to run k-Means several times with a different number of clusters and compute the metric for each run.

3.3.1. Intra-Cluster Distance

Using the intra-cluster distances the elbow method is a popular method to determine the right number of clusters. The idea is that adding another cluster to the optimal number of clusters will not explain significant more variance or enhance the result. The intra-cluster distance is defined as the mean of distances of each sample to the center of the associated cluster. In mathematical terms

$$\frac{1}{|S|} \sum_{s \in S} \|s - \mu_s\|^2 \tag{2}$$

where S is the set of all samples, s a sample and μ_s the associated cluster center. This is concurrent with equation 1. Calculating the intra-cluster distance for all number of clusters shows a monotonically decreasing progression. The number of clusters where the slope of the progression becomes linear is the best value for the number of clusters. If the slope becomes small and linear, the variance in every added cluster is very small. This is obvious in the case of

the number of clusters are equal to the number of samples. Each sample will have its own cluster center. Hence, the objective function (1) tries to minimize the distance between the cluster centers and the samples, the function would be 0, the intra-cluster distance as well and there is no explained variance.
The elbow method is not proven to determine the optimal number of clusters. So, it is reasonable to use another metric. That is when the Silhouette score comes into account.

3.3.2. Silhouette Score

The Silhouette score is a measure for quantifying the certainty of whether a sample is assigned to the correct cluster or not [27]. For computing the Silhouette value of one data sample two values are needed. The first value determines how well the sample fits to its assigned cluster or the dissimilarity of the sample i to all other values of the cluster A

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, i \neq j} d(i, j) \tag{3}$$

where *d(x,y)* is an arbitrary metric function. The second value is the minimum dissimilarity to the values of the other clusters. This value is computed in a similar fashion

$$b(i) = \min_{A \neq B} \left\{ \frac{1}{|B| - 1} \sum_{j \in B, i \neq j} d(i, j) \right\} \tag{4}$$

The Silhouette value for the sample i is computed by the formula

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

Considering the equation 5 we easily see that

$$-1 \leq s(i) \leq 1 \tag{6}$$

if $s(i) = -1$ the sample i is likely assigned to the wrong cluster. If $s(i) \approx 0$ the sample is lying close or on the decision boundary between two clusters and if $s(i) = 1$ the sample is likely to be assigned to the right cluster. This holds for the Silhouette value as well.
As well as the intra-cluster distance the Silhouette value takes the complexity of the model indirectly into account. So, the influence of the complexity is not easy to derive. That is why we consider the Bayesian Information Criterion among these two.

3.3.3. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is based on the idea that a model should not be more complex than necessary [28]. Hence, it has a penalty term with the number of model parameters as a parameter. The BIC is defined as follows

$$BIC = n \ln\left(\hat{\sigma}^2\right) + p \ln(n) \tag{7}$$

where n is the number of samples, $\ln\left(\hat{\sigma}^2\right)$ the log-Likelihood function and p the number of parameters of the model. The log-Likelihood function shows how well the data fits to the tested distribution. The penalty term grows linear with the number of model parameters and logarithmic with the number of samples. Since the number of samples is mostly fixed, the penalty term grows linear with the number of model parameters.

4. RESULTS

We applied k-Means for $k \in [1, ..., 50]$ clusters to the dataset and computed the values for the three metrics. Figure 2 shows the plot of the inter-cluster distance. As expected the graph is decreasing as the number of clusters is increasing. The best number of clusters according to the elbow method is 15.
The graph of the BIC (Fig. 3) is steadily decreasing until 30 clusters. After 30 clusters the graph is slightly increasing, but shows an asymptotic behaviour. Due to our implementation of the BIC using the log-Likelihood method, the smallest BIC value indicates the best choice for the number of clusters. In this case the best choice is 46.
The maximum of the graph of the Silhouette score depicts the best choice of clusters. The graph in Figure 4 is monotonically increasing and converges to 1. So, the best value for the number of clusters is 25 or greater.

Table 3: The last five clusters as an example. The values of the features depict the presence of a sensor event. If the value is smaller than 1, not all associated samples have an event for this sensor. For readability values smaller than 10^{-3} are replaced by 0. These values are marked by a *.

Feature/Cluster	36	37	38	39	40
Bed	0	0	0	0	0
Fridge	0	0	0	0	0
Living room	1	.99	0	0	0
Toilet	0*	0*	.13	1	0
Front door	0	0	0	0	0*
Rear door	.0028	.00069	0	.00013	0
Bathroom	0	0	.079	1	0
Hallway	1	0	0	0	0
Bedroom	0	0	.079	0	0
Kitchen	0	0	0	0	0
Lamp bedroom	0	0	0	0	0
Lamp living room	0	0	0	0	0
Kettle	.0045	0	1	0.003	1
TV	.0062	0	1	0	1

Besides clustering the merged dataset, we clustered the data for each participant separately as well. For the sake of comparison we used the same methods and metrics.
The elbow method for participant 1 suggests 13 clusters, the BIC 45, and the average Silhouette Score 40 as the optimal choice. Combining the BIC and the average Silhouette Score we choose 35 as the best number of clusters.

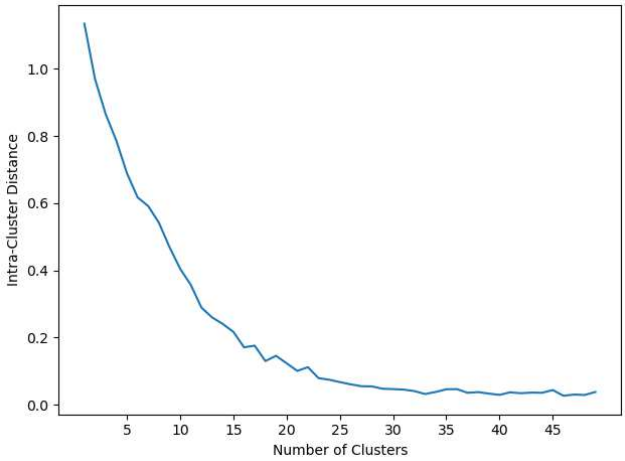

Figure 2: The values for the elbow method for all 50 different values for the amount of clusters. According to the graph 15 clusters is the best choice.

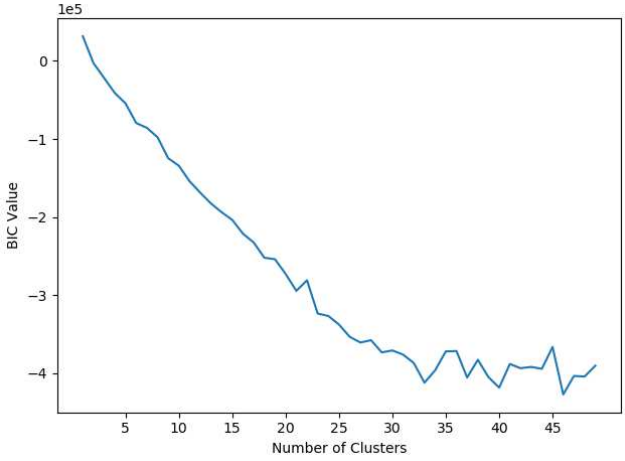

Figure 3: The Bayesian Information Criterion computed for all the amount of clusters. The BIC shows 46 clusters as the best choice.

For 35 clusters the BIC is close to its minimum and the average Silhouette Score starts to converge to its best value. The metrics are shown in Figure 7. The elbow method for participant 2 suggests 7 clusters, the BIC 12, and the average Silhouette Score 12 as the optimal choice. Combining the BIC and the average Silhouette Score we choose 12 as the best number of clusters, because the BIC has its minimum at 12 and the average Silhouette Score has its optimal value at 12. The metrics are shown in Figure 5. The elbow method for participant 3 suggests 5 clusters, the BIC more than 50, and the average Silhouette Score 5 as the optimal choice. Combining the elbow method and the average Silhouette Score we choose 5 as the best number of clusters. The graphs of the elbow method and the average Silhouette Score are constant for values larger than 5. The metrics are shown in Figure 6.

The sum of the optimal number of clusters for the participants is 52 and larger than the optimal number of clusters for the combined data. Considering the cluster where the

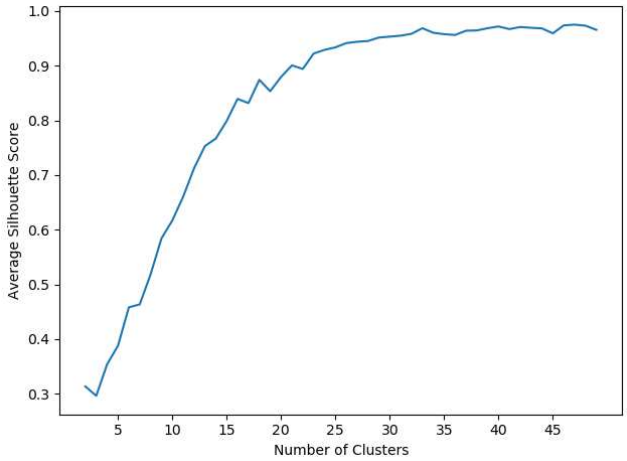

Figure 4: The average silhouette score for one cluster up to 50 clusters. The average silhouette score suggests 25 or more clusters as the best choice.

data of the participants are assigned to the sum of the clusters is 58. We calculated the pairwise smallest Euclidean distance between the clusters of all participants and the clusters of the separate clustering. Table 4 shows how often one of the 40 clusters was selected as the closest to a cluster of the separate clustering. The The majority of clusters has not been selected as closest cluster. Cluster 5 has most often been selected. The minimum distance varies for all participants. Participant 3 has the smallest distance with 0.000435, participant 2 the second smallest with 16.37 and participant 1 the largest distance with 124.42.

Table 4: The minimum distance between the set of clusters of each participant and the set of clusters of the combined dataset. The Euclidean distance between each of the cluster centers has been computed and the minimum distance for each cluster has been summed up.

ID	1	2	3
Distance	124.42	16.37	0.00435
relative Distance	3.55	1.37	0.00087

Table 4 shows the last five cluster centers. The range of the values is between 0 and 1 corresponding to the dataset. The majority of the values greater 0 are very small $\left(10^{-1}, 10^{-6}\right)$. The dimension for the feature *Bed* is 0 at all cluster centers. The dimensions for *Lamp bedroom* and *Lamp Living room* are only 0 or 1. There is no dimension that has decimal values only.

Since we balanced our dataset and kept track of the participant each sample is from, we can compute how many samples of a certain participant belongs to a cluster. Table 6 shows the labels of each cluster and how many percent of the assigned values are from each participant. For the most clusters the assignment of its label is unambiguous. The majority, 31, of the clusters are formed by samples of a certain participant. Six of 40 clusters have an assignment

Table 5: The number of times one of the 40 clusters was the closest cluster to one of the participants clusters.

Number of Selections	0	1	3	5	6	13	23
Cluster Number	8 - 40	2, 6	7	3	4	1	5

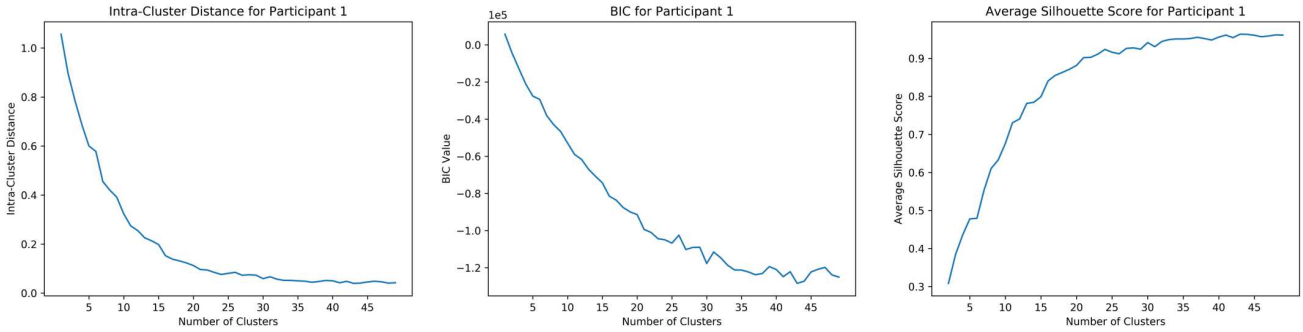

Figure 5: The three metrics computed or participant 1. The optimal number of clusters is 35. The first plot shows the intra-cluster distance, the second one the BIC and the third one the average Silhouette Score.

rate between 84% and 99% and only three clusters have an assignment rate less than 84%.

5. LIMITATIONS

The major limitation of this research is the number of participants which had to be excluded. K-Means clustering and clustering in general is sensitive to missing values. The data has been collected in a field study in the real-world. So, our data may have some errors and is not uniform. We had to exclude 17 participants to get a large uniform set of sensors and prevent missing values. The data of the three participants might not be collected at the same time, i.e. we had data of the first participant in a certain week, where no data of the other participants were available. Moreover, temporary sensor failures or false values might have occurred and cannot be detected. We also tried to filter data where more than one person were in the flat. Due to unreliable documentation of the daily life and limitations of the algorithms, we might not were able to filter all of those data.

6. DISCUSSION

We used three metrics. Each metric suggests a different number of clusters. So we chose the optimal number of clusters by majority voting. The BIC and the Silhoutte score are suggesting 40 as the optimal number of clusters. The BIC has the second smallest value for 40 clusters and the Silhouette score is nearly constant for 40 and more clusters. For the participants the optimal number of clusters are 35, 12, and 5 for participants 1, 2, and 3 respectively. After deriving the optimal number of clusters we analysed the clusters in terms of the values of the cluster centers, the amount of assigned samples of each participant for labeling the clusters, and the difference to the clusters of the separately clustered data.

We have cluster centers where an appliance is used, but the corresponding motion sensor is 0. For example there is no cluster with values greater 0 for *Kettle* and *Kitchen*. For turning on the kettle the person must enter the kitchen. Hence, there must be an event of the motion sensor. It seems reasonable that the correlation is not dominant enough to form a cluster. Due to our idea that behaviour might not be defined by activities of daily living and our data driven approach not all of the clusters can be labeled clearly. We were able to label 13 clusters, but 27 ambiguous clusters are still remaining. For example cluster 2 can be labeled as *Watch TV*. All dimensions in this cluster are 0 except for the dimensions of *TV* and *Lamp Living room*. We assume that the person is watching TV in the living room. The same holds for clusters 1 and 9. Cluster 1 might shows a person who likes watching TV and drinking tea while watching, but does not want to have the lights turned on in the room with the TV. The 1 for the features *TV* and *Kettle* and the 0 for all lamps indicates that. In contrast cluster 9 shows a person who likes to watch TV in an enlightened flat. In this case it is likely that those three clusters represent the TV watching behaviour of the three different participants. Each cluster has a different label unambiguously assigned with a ratio of about 99% of the values from one participant. Other clusters are not as obvious as the mentioned ones. The clusters 14, 21, 16 and 27 are not easy to interpret in terms of activities. Our data driven approach even detect behaviour that would not be considered as activity of daily living or activity at all. For example the clusters 4 and 25 are ways the participant is walking through the flat. Our labels and the associated clusters are shown in Table 7. Since we chose the participants, because two are similar we expected at least a few

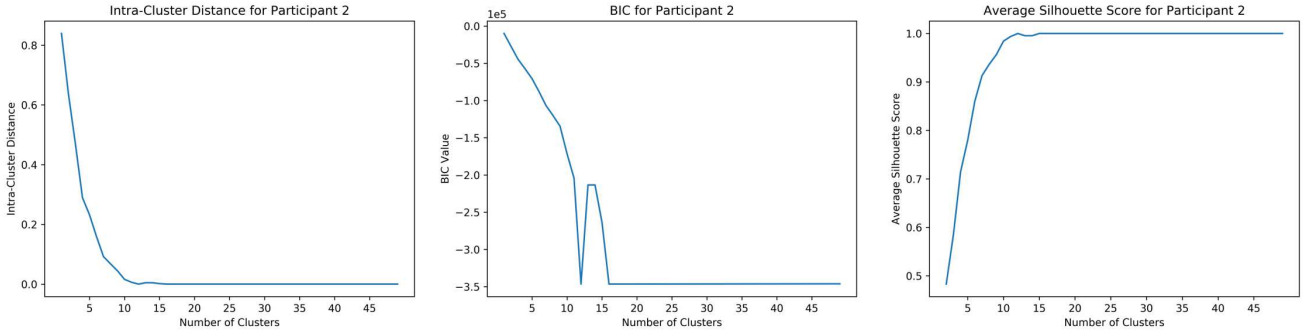

Figure 6: The three metrics computed or participant 2. The optimal number of clusters is 12. The first plot shows the intra-cluster distance, the second one the BIC and the third one the average Silhouette Score.

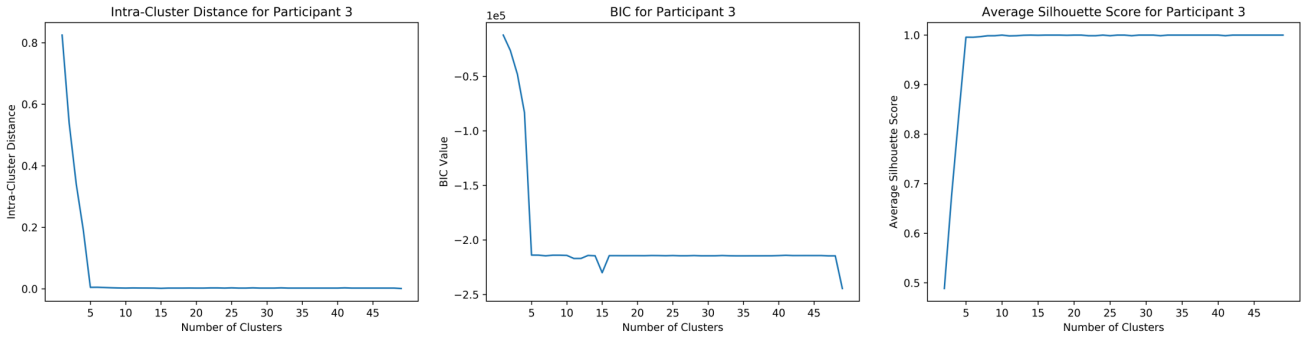

Figure 7: The three metrics computed or participant 3. The optimal number of clusters is 5. The first plot shows the intra-cluster distance, the second one the BIC and the third one the average Silhouette Score.

clusters with equally distributed samples.

Comparing the clusters centers of each participant to the cluster centers of the merged dataset, we found that the results were very different. The difference of the clusters centers are varying among the participants. The behaviour of participant 3 is captured in both approaches, because the difference between the cluster centers is 0.00435. The behaviour of participant 1 and 2 were not that present in the cluster centers of the merged set. That holds even for the relative distance. The difference of the results is reflected in the number of closest clusters as well. The majority of clusters are not selected to be the closest to the separate cluster centers, i.e. 32. The clusters representing the activity *watch TV* is the closest one with 36 selections. This means the *watch TV* behaviour is most similar between the participants. Most of the clusters describing *ambiguous* activities are very different from the clusters of the separated participants.

Considering the absolute number of clusters the results are different as well. The sum of the cluster centers of the separate clustering is 58, compared to the 40 clusters of the merged dataset. The worse the scores, the higher the number of clusters.

Since the total number of clusters for considering each participant separately is larger than the number of cluster merging the data of participants. Considering each participant separately gives a better overview about the behaviour. Most datasets are not directly comparable in terms of sensor data, participants etc. So, using this dataset as baseline might be not possible without further abstraction. However, the findings show that the difference of the results when merging the data of participants and considering each participant separately must be considered.

7. CONCLUSION AND FUTURE WORK

In this article we investigated whether elderly people who have similar in questionnaire results and assessment scores have a similar behaviour. By the use of k-Means clustering we were able to show that the behaviour is different. We merged the data of three participants to one dataset and applied k-Means to the merged dataset and to the data of each participant separately. The formed clusters were different and so was the behaviour. That finding indicates that this difference has to be considered in behavioural research. To verify and improve this finding we plan the following two steps. We used the maximum amount of sensors and that reduced the number of participants significantly. So, one step is to investigate what sensors and how many sensors are needed to capture the variability of the behaviour of elderly people. This will increase the number of participants and also gives infor-

Table 6: We analysed how many samples of each participant were assigned to each cluster. We see that the majority of samples for each cluster are from one participant. For example 100% of the samples associated to cluster 1 are from participant 3.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Label	3	2	1	1	1	2	1	3	1	1	3	3	1	1	1	1	3	2	2	2
Ratio	1	0.99	1	1	1	1	1	0.84	1	1	0.91	0.91	1	1	1	1	0.99	0.69	0.91	1

Cluster	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Label	1	1	1	2	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	2
Ratio	1	1	1	0.78	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.66

Table 7: The labels and associated clusters. We labeled similar clusters with the same label, e.g. the two clusters in *Walk in the flat* are showing to different ways. There are 27 ambiguous clusters and 13 we were able to label.

Watch TV	Walk in the flat	Entering the flat	Toilet	Cooking	Rest	Ambiguous
1,2,5,9,11, 17,19	4,25	13	22	31	6,33	3,7,8,14,15,16,18,20,21, 23,26,27,28,29,30,32,34, 35,36,37,38,39,40

mation about how many sensors are needed to be used for sophisticated behaviour monitoring of elderly people. The other step is applying different clustering algorithms. DBSCAN and Gaussian Mixture Models are algorithms that have been proven useful in behaviour analysis. The results can be used to verify that the difference of the results is algorithm independent.

We acknowledge Prof. Dr. Jürgen Bauer (University of Heidelberg) for designing and supervising the OTAGO study. We acknowledge Bianca Sahlmann (University of Oldenburg) and Lena Elgert (Peter L. Reichertz Institute, Hannover) for performing the assessments. The OTAGO study has been funded by an internal funding of the Carl von Ossietzky University of Oldenburg.

References

[1] F. Mahoney, D. Barthel, Functional evaluation: The barthel index, Maryland State Medical Journal 14 (1965) 56–61.

[2] K. Popper, The Logic of Scientific Discovery, 1959.

[3] S. Ackroyd, J. Hughes, Data Collection in Context, 1981.

[4] E. Giannouli, O. Bock, S. Mellone, W. Zijlstra, Mobility in old age: Capacity is not performance, BioMed Research International 2016.

[5] N. Peel, S. Kuys, K. Klein, Gait speed as a measure in geriatric assessment in clinical settings: A systematic review, The Journals of Gerontology: Series A 68 (1) (2013) 39–46.

[6] M. A., G. Fulk, M. Beets, T. Herter, S. Fritz, Self-selected walking speed is predictive of daily ambulatory activity in older adults, Journal of Aging and Physical Activity 24 (2) (2016) 214–222.

[7] A. Gerka, C. Lins, M. Pfingsthorn, M. Eichelberg, S. Müller, C. Stolle, A. Hein, A clustering-based approach to determine a standardized statistic for daily activities of elderly living alone, Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies 5 (2019) 264–271.

[8] T. Nef, P. Urwyler, M. Büchler, I. Tarnanas, R. Stucki, D. Cazzoli, R. Müri, U. Mosimann, Evaluation of three state-of-the-art classifiers for recognition of activities of daily living from smart home ambient data, Sensors 15 (2015) 11725–11740.

[9] S. Mahmoud, A. Lotfi, C. Langensiepen, Abnormal behaviours identification for an elder's life activities using dissimilarity measurements, in: ACM International Conference Proceeding Series, 2011, p. 25.

[10] A. Aramendi, A. Weakly, A. Goenaga, M. Schmitter-Edgecombe, D. Cook, Automatic assessment of functional health decline in older adults based on smart home data, Journal of Biomedical Informatics 81 (2018) 119–130.

[11] P. Dawadi, D. Cook, M. Schmitter-Edgecombe, Automated clinical assessment from smart home-based behavior data, Journal of Biomedical Health Informatics 20 (4) (2016) 1188–1194.

[12] A. Alberdi, A. Weakley, M. Schmitter-Edgecombe, D. Cook, A. Aztiria, A. Basarab, A. Barrenechea, Smart home-based prediction of multi-domain symptoms related to alzheimer's disease, Journal of Biomedical Health Informatics 22 (6) (2019) 1720–1731.

[13] R. R.E., C. Jones, Development and validation of a functional fitness test for community-residing older adults, Journal of Aging and Physical Activity 7 (2) (1999) 129–161.

[14] M. Taylor, The Fundamentals of Clinical Neuropsychiatry, New York: Oxford University Press, 1999.

[15] C. T. M. Randolph, E. Mohr, T. Chase, The repeatable battery for the assessment of neuropsychological status (rbans): preliminary clinical validity, Journal of Clinicla and Experimental Neuropsychology 20 (3) (1998) 310–319.

[16] J. Crawford, G. Smith, E. Maylor, S. Sala, R. Logie, The prospective and retrospective memory questionnaire (prmq): Normative data and latent structure in a large non-clinical sample, Memory 11 (3) (2003) 261–275.

[17] M. Schmitter-Edgecombe, C. Parsey, R. Lamb, Development and psychometric properties of the instrumental activities of daily living: Compensation scale, Archives of Clinical Neuropsychology 29 (8) (2014) 776–792.

[18] P. Dawadi, D. Cook, M. Schmitter-Edgecombe, C. Parsey, Automated assessment of cognitive health using smart home technologies, Technology in Health Care 21 (4) (2013) 323–343.

[19] D. Podsiadlo, S. Richardson, The timed "up & go": A test of basic functional mobility for frail elderly persons, Journal of the American Geriatrics Society 32 (1991) 142–148.

[20] J. Guralnik, E. Simonsick, L. Ferrucci, R. Glynn, L. Berkman, D. Blazer, P. Scherr, W. R.B., A short physical performance battery assessing lower extremity function: Association with self-reported disability and prediction of mortality and nursing home

admission, Journal of Gerontology 49 (1994) M85–M94.
[21] S. Searle, A. Mitnitski, E. Gahbauer, T. M. Gill, K. Rockwood, A standard procedure for creating a frailty index, BMC Geriatrics 8.
[22] M. Lawton, E. Brody, Assessment of older people: Self-maintaining and instrumental activities of daily living, The Gerontologist 9 (1969) 179–186.
[23] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
[24] S. Lloyd, Least squares quantization in pcm, IEEE Transactions on Information Theory 28 (1982) 129–137.
[25] A. Vattani, k-means requires exponentially many iterations even in the plane, Discrete and Computational Geometry 45 (2011) 596–616.
[26] D. Arthur, B. Manthey, H. Roeglin, k-means has polynomial smoothed complexity, in: Proceedings of the 50th Symposium on Foundations of Computer Science (FOCS), 2009, p. 405–414.
[27] P. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.
[28] G. Schwarz, Estimating the dimension of a model, Annals of Statistics 6 (1978) 461–464.