Interpreting Presence Sensor Data and Looking for Similarities Between Homes Using Cluster Analysis

John Loane, Brian O'Mullane, Brennon Bortz, R. Benjamin Knapp
CASALA
Dundalk Institute of Technology
Dundalk, Ireland.
john.loane@casala.ie
brian.omullane@casala.ie
brennon.bortz@casala.ie
ben.knapp@casala.ie

***Abstract*—In order to model older people's behaviour in the home we must first understand it. In this paper we examine data from eight purpose-built aware homes over a six-month period, looking at presence in rooms to try to determine patterns amongst the older residents. We look for homes that have similar movement patterns using cluster analysis. We also examine how movement over days clusters within individual homes. Our analysis begins to show the possibilities of distinguishing between residents in their homes based on patterns of movement.**

Keywords-AAL; Pervasive Sensor Data; Scatter Plot; Cluster Analysis; Silhouette Coefficent

I. INTRODUCTION

The population is living longer and with this there is a push towards improving quality of life of older people as well as allowing them control and autonomy while aging. In 2002, Oeppen and Vaupel's research began to consider the possibility of the lack of a limit on life expectancy, given then current and now continued trends in aging. Nine years ago, female life expectancy in the record-holding country (Japan), had risen for 160 years at a steady pace of almost three months per year, and this upward trend has continued. Furthermore, Oeppen and Vaupel provide evidence against the arguments that these results are fogged by other trends such as declines in infant mortality, showing that in the second half of the 20th century improvements in survival *after* age 65 showed a marked increase—as much as doubling in some developed countries [1]. The problems associated with people living independently while aging and declining in health are widely reported in the literature. They include vision decline [2], hearing loss [3], diminished motor skills and reduced cognition effects [4]. There are many commercial products on the market that can help with some of the effects of this decline, such as pendant alarms to call emergency services after a fall. To truly care for older people in their homes it is crucial to begin to arrest this decline by becoming cognisant of and responsive to its early indicators.

Ambient assisted living offers a potential solution to this problem and hence is an active area of research. It involves embedding low impact pervasive sensors, such as presence sensors and door usage sensors in homes, to help build a picture of behaviour and detect when this behaviour changes over time, which may be an indicator of decline. But what does this behaviour look like? How can one use statistical techniques to help model this behaviour? Is this behaviour unique to an individual or can certain generalisations be made across all the aging population? In this paper we present data gathered from a range of sensors embedded the homes of 8 older people over a six-month period. We look for commonalities as well as uniqueness in behaviour across these people and draw conclusions about the techniques as well as the subjects. Our aim is to build models that will ultimately predict a resident's behaviour. As a first step towards this modelling we must understand the data and patterns contained therein.

II. BACKGROUND AND RELATED WORK

A primary activity following from the main research foci at the Centre for Affective Solutions for Ambient Living Awareness (CASALA) is our work with a number of older adults living at the Great Northern Haven (GNH). GNH is a demonstration housing project consisting of 16 purpose-built aware homes in Dundalk, Co Louth, Ireland. Each home is equipped with a combination of sensor and interactive technologies to support ambient assisted living for older people. Currently there are 13 homes occupied by 11 men and 4 women. To date, we have collected a vast amount of data from the 100 plus sensors embedded within each of these homes, giving a total of 2240 sensors and actuators throughout the development. The sensors include presence sensors, contact sensors on all internal and external doors and windows, electricity, water and heating usage sensors, ambient light and temperature sensors as well as an array of other ambient sensors. For this study we are examining movement behaviour, using just the presence sensors in the hallway, bedroom and living room. These sensors are standard passive infrared sensors tuned to give readings of presence in a room with a reset interval of ten seconds.

The sheer amount of data acquired from over two thousand 'always-on' sensors can be overwhelming at first. As with any pattern recognition problem, however, these issues were addressed through data pre-processing and clustering, as by other researchers for initial data exploration in this and related areas of research [5],[6]. This paper focuses primarily on knowledge gleaned through the analysis of our data clustering work, which will later serve to inform a robust behaviour recognition engine, currently in development. For data clustering, we have experimented with several techniques; specifically addressed here are primarily the results of *k*-means clustering, as well as a discussion of agglomerative hierarchical clustering [7]. *k*-means clustering begins with an initial random partitioning of the data set. As the algorithm progresses, the partitions and cluster centres are continually adjusted and patterns reassigned until a convergence requirement is met. Other research in this area has either implemented *k*-means clustering for initial data exploration—these include the promising clustering results with a *k*-means algorithm in behaviour recognition through PIR sensors in similar contexts by Lotfi et al [8], the work of McKenna et al in fall detection in supportive homes through *k*-means clustering of visual data [9], and others. As this research moves towards pattern recognition, information gained from this exploration will be used to inform the chosen recognition method, be it Dynamic Time Warping [10],[11], Hidden Markov Models [12],[13], Neural Networks [14], Support Vector Machine [15] or otherwise [16].

III. Data acquisition

All of the data collected in this project is stored in a MySQL database. Each time one of the presence sensors is triggered an entry is written to the database including sensor identification details, time of the write (accurate to the millisecond), and sensor reading value.

The first step towards analysing this data was to determine when each of the residents settled into their homes. This was calculated by searching for six consecutive days where the hall presence sensor fired more than twenty times. The move-in date was set to be seven days after initial activity in the apartment. Removing the data for these first six days removed noisy presence data associated with movers.

Two of the homes are occupied by two residents—data for these homes were removed from the examination as the presence for each resident was difficult to determine.

Overall, the analysis was carried out on data gathered between July and early December 2010. In data pre-processing, we encountered an issue with incomplete data, due to issues with third-party hardware and software that were beyond our control—these issues meant some days for which data were collected contained in the database had only partial or no data. In order to find these days we looked at the electricity sensor, which is periodic, and fires six times per minute. Any days for which there were less than 24 complete hours of data were removed from the analysis. In total, 17 days for which no data were available and 16 days for which we only partial data were available out of a possible 168 days were removed from the dataset. Hence, this analysis is based on 79 days of clean data.

Once the move-in dates are taken into account Home 4 had 98 days of data, Home 26 had 112 days of data, Homes 7, 8, 20 and 5 had 135 days of data, Home 22 had 79 days of data and Home 12 had 131 days of data. For accuracy in comparison, 79 days worth of data were considered for each residence.

The layout of the homes is shown in Figure 1. The resident must pass through the hall way sensor when moving between any of the rooms, such as when they are visiting the main bathroom from the living area during the day.

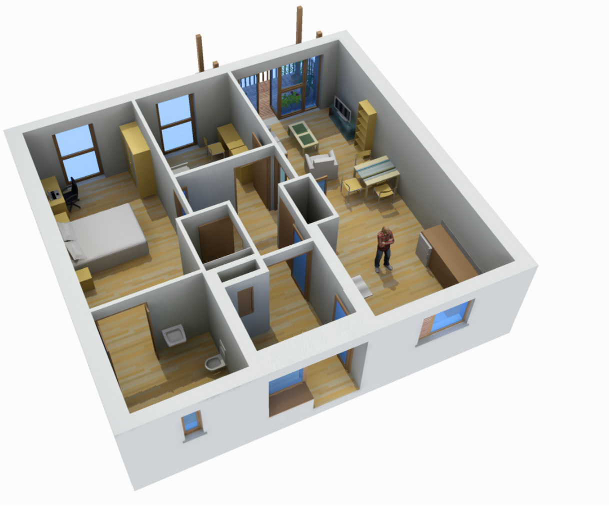

Figure 1. Typical Great Northern Haven home layout.

Each presence sensor fired on average 155 times a day and the sensor firing times gave a characteristic that was easily identifiable by manual inspection. To help visualise behaviour patterns the sensor data were represented on a spiral plot called a "last clock" which plots the data on a 24-hour clock with midnight at the top and spirals out from the centre. Each circuit represents a day (see Figure 2, Figure 3 and Figure 4).

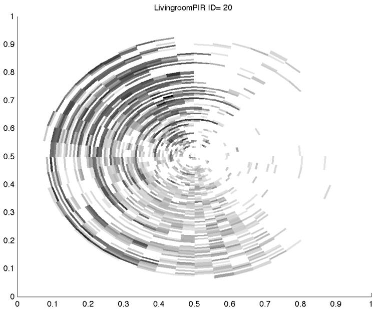

Figure 2. Living room presence sensor, showing most of the activity in the evening between 18:00 and 24:00.

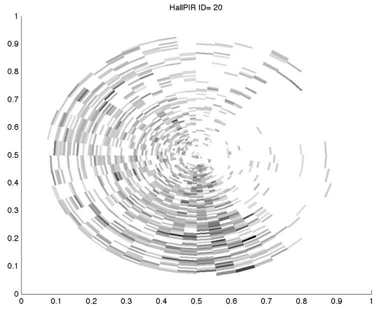

Figure 3. Hall presence sensor data, showing most of the activity during the middle of the day.

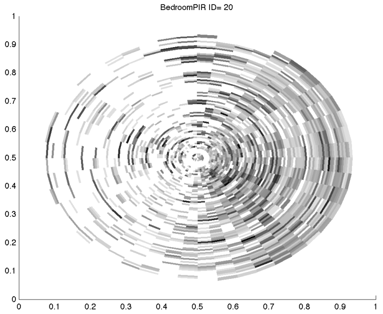

Figure 4. Bedroom presence sensor data, showing most of the data in the evening (the sensor is sensitive to motion in bed.)

IV. Methodology – Visualising the Data

The total number of sensor firings for each of the hall, bedroom and living room presence sensors for each day grouped into six-hour intervals was calculated and graphed using scatter plots. All sensors were graphed on the same scatter plot in order to allow comparison between sensors within each home.

V. Scatter Plots

A. Results

In the following scatter plots circles represent movement in the living room, triangles the bedroom and x the hall. The scales of all graphs are equalized for straightforward comparison and data exploration.

We intended to use these graphs to answer the following questions:

1) How do residents use the rooms in their home?
2) Do any residents have similar movement patterns?
3) For a given resident, is their movement increasing, decreasing or staying stable?

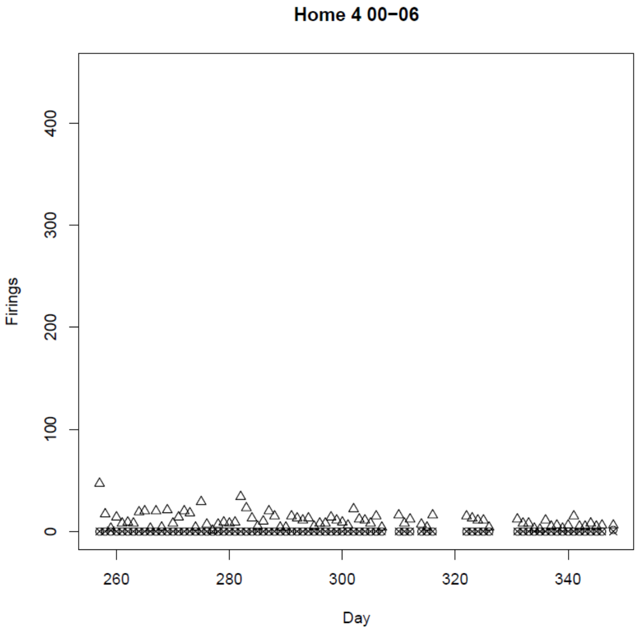

Figure 5. No activity in the living room between 00:00 and 06:00. Circles indicate living room, triangles indicate bedroom, and *x*s indicate the hall.

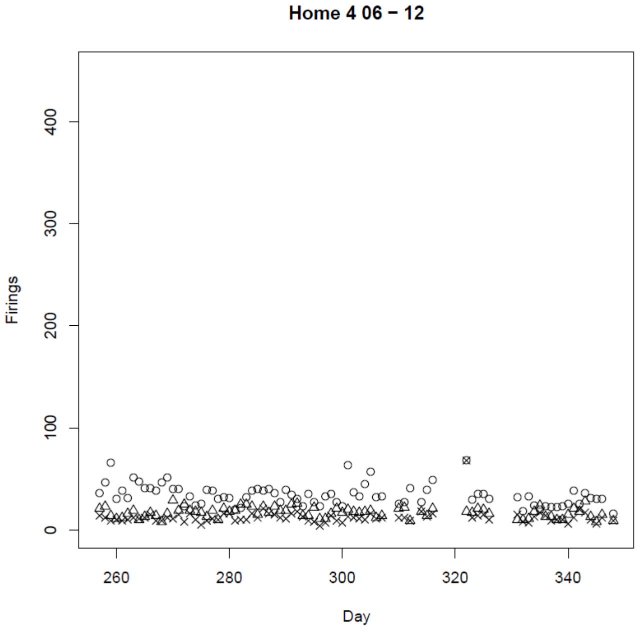

Figure 6. Activity between 06:00 and 12:00.

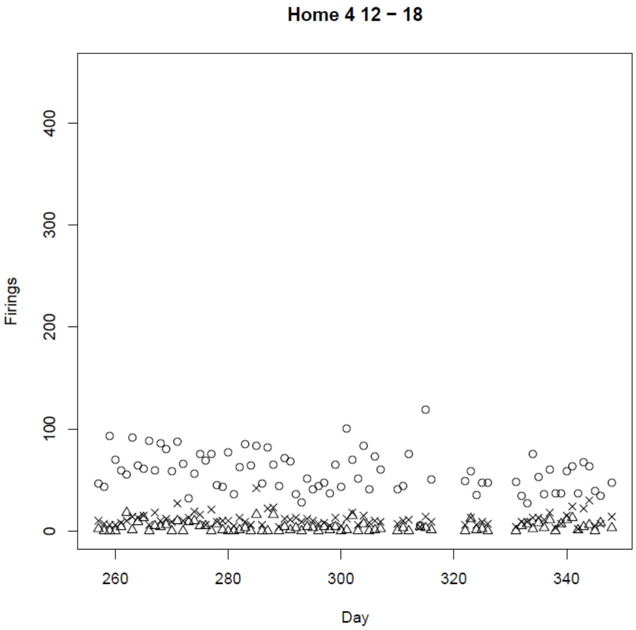

Figure 7. Activity between 12:00 and 18:00.

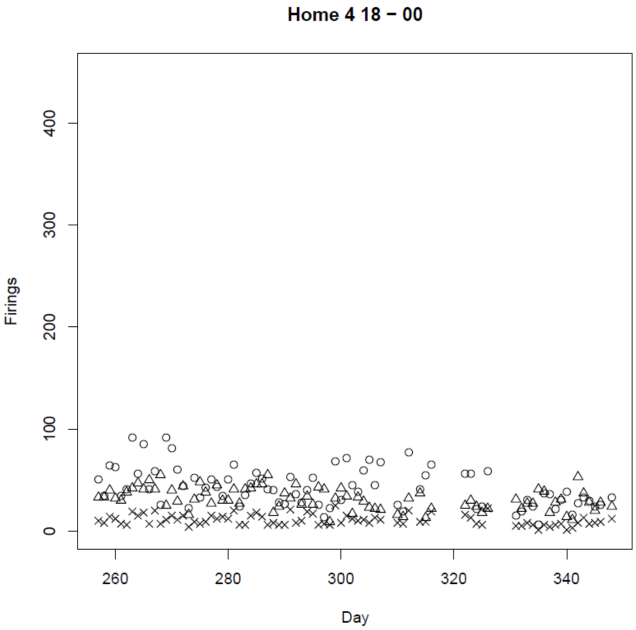

Figure 8. Activity between 18:00 and 24:00.

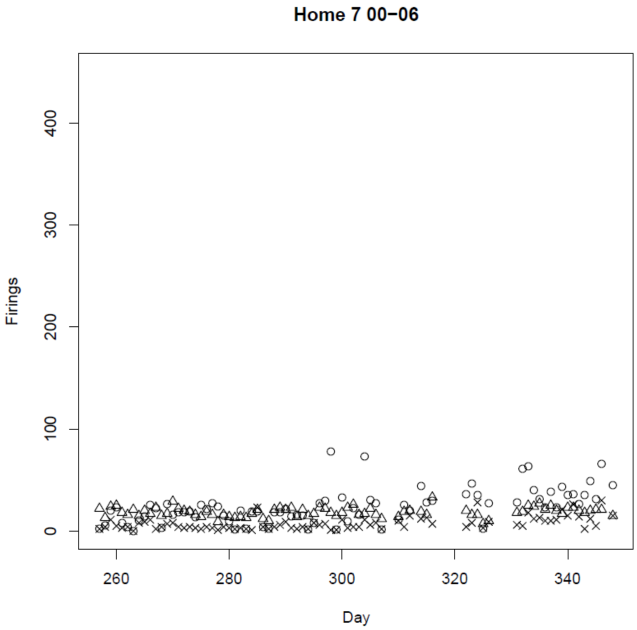

Figure 9. Activity between 00:00 and 06:00.

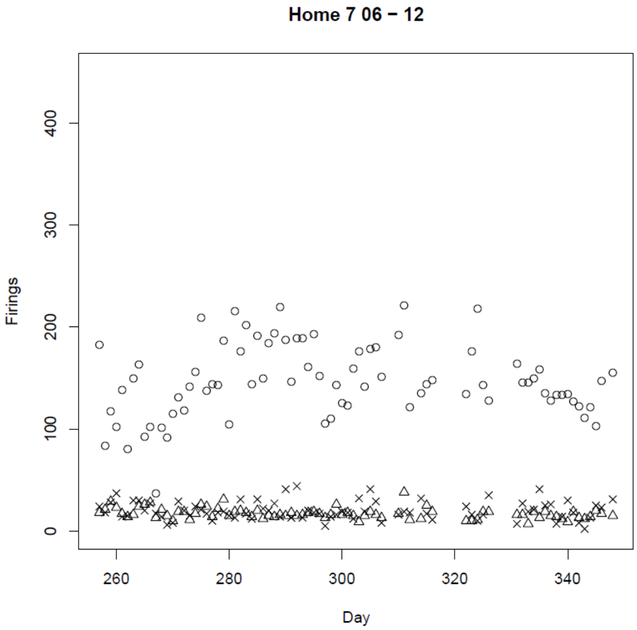

Figure 10. Activity between 06:00 and 12:00.

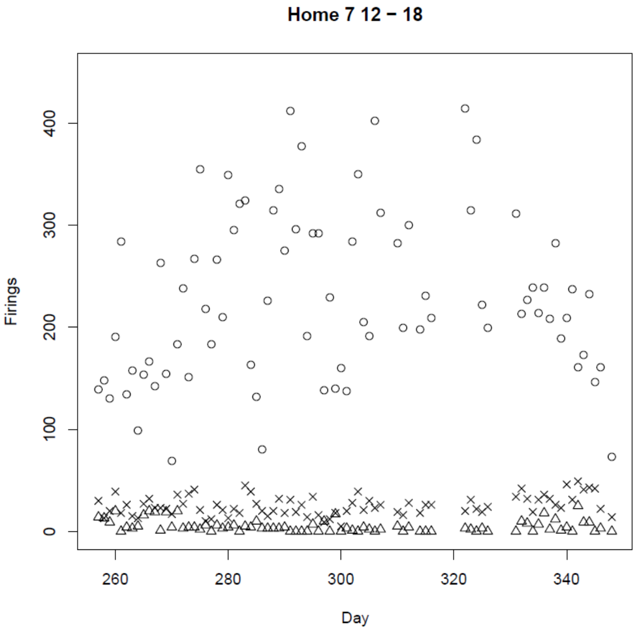

Figure 11. Activity between 12:00 and 18:00.

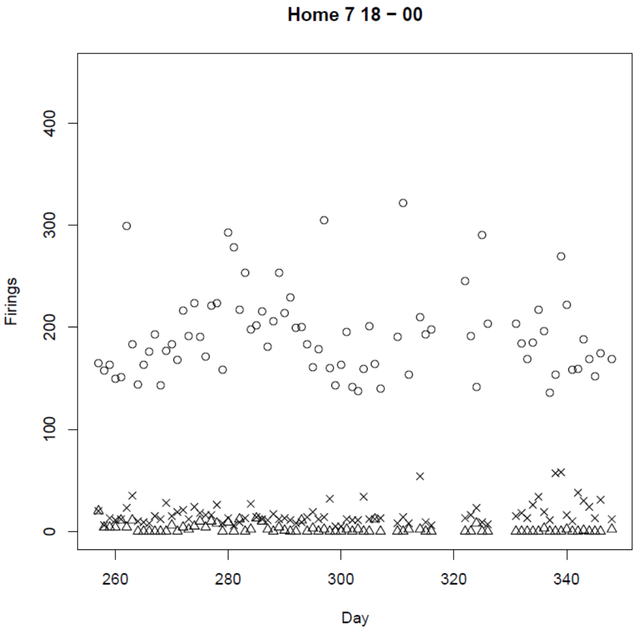

Figure 12. Activity between 18:00 and 24:00.

B. *Discussion*

The answer to Question 1 above is that most residents move around their living rooms substantially more than both the hall and bedroom. Home 7 is a particularly extreme example of this.

Question 2 is more difficult to answer solely by examination of the graphs. Homes 4 and 7 have very different movement patterns. We will return to this question later when discussing data clustering, but it is important to validate the clusters by examining the data.

Question 3 is impossible to answer due to the large variability in the data.

VI. Methodology – Cluster Analysis Looking for Similar Homes

In order to find homes with similar movement patterns a cluster analysis was carried out on the data. The first issue that arose here is that all homes did not have the same number of days of data due to the differing move-in dates. The home with the latest move-in date was used as the cut-off for the cluster analysis. All days earlier than this were omitted from the clustering. This meant that the clustering was carried out on 79 days of data. Each of those days was split day into four equal time periods: 00:00-06:00, 06:00-12:00, 12:00-18:00 and 18:00-00:00. The total number of times that the hall, bedroom and living room PIR sensor was triggered in each of these time periods was calculated. Hence, we were clustering on 948 values for each home. The cluster package in the R programming language [17] was used to carry out the analysis, employing both *k*-means and hierarchical methods.

When using *k*-means clustering the number of clusters must be known *a priori* and specified within the parameters of the clustering algorithm. In order to choose the best number of clusters we carried out the analysis on a number of different clusters. We then plotted the silhouette plot [18] for each of these numbers of clusters. By construction the silhouette coefficient ranges from -1 to 1—negative values indicate that the cluster radius is greater than the distance between the clusters. This indicates that the clusters overlap and hence the clustering is poor. We consider the average of all the silhouette coefficients of all points in each cluster and use this as a measure of the quality of our clusters. The aim is to choose the number of clusters so as to optimize the silhouette coefficient.

At the first step in the hierarchical algorithm each observation constitutes a cluster. At each step, the two closest clusters are joined to form a new cluster. Thus, the groups at each step are nested with respect to the groups obtained at the previous step. Once an object has been assigned to a group it is never removed from the group later in the clustering process. The hierarchical method produces a complete sequence of cluster solutions beginning with *n* clusters and ending with one cluster containing all *n* observations. The proper number of clusters has to be selected.

At each step in the hierarchical algorithm we should join the two closest clusters. Our starting point is the dissimilarity matrix. It is easy to determine the two closest observations. Now a problem arises: how do we calculate the dissimilarity between one observation and one cluster or between two clusters? There are a large number of possible answers. In this analysis we use Ward's method [19]. This method is based upon the concept of within sum of squares. The two clusters with the smallest between sum of squares are joined.

The agglomerative process for hierarchical cluster analysis can be graphically represented using a tree diagram, also called a *dendrogram*, with cases on the horizontal axis and the dissimilarity between the clusters joined at each step on the vertical axis (the dissimilarity is normalized). If a large change in the height occurs subsequent to an aggregation at step *C* then the solution immediately prior to this step (step *C*-1) should be chosen.

VII. Cluster Analysis – Looking for Similar Homes Results

A. Results

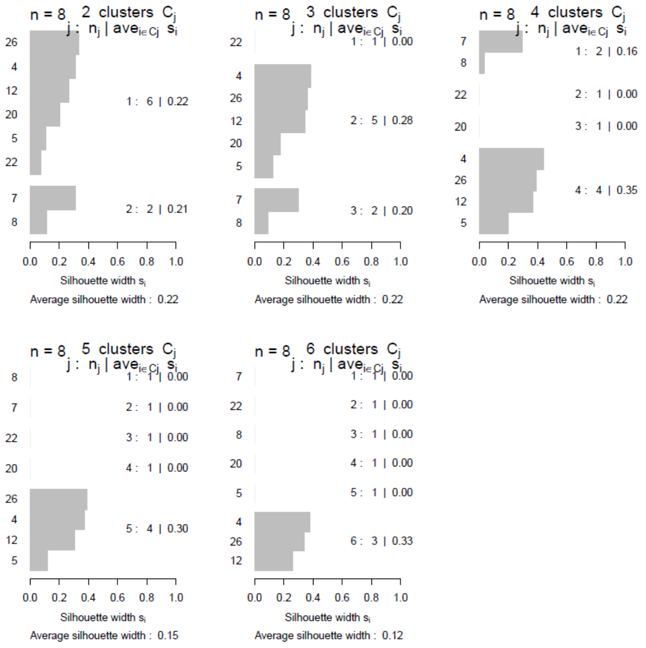

Figure 13. Results of *k*-means clustering using two, three, four, five and six clusters.

The plots in Figure 13 show the results of the *k*-means clustering. The best clustering occurs in the case of 4 clusters where we get an average silhouette coefficient of 0.35 for the cluster Home 4, Home 5, Home 12 and Home 26. Each element except Home 5 within this cluster has a silhouette coefficient above 0.3. In fact there is no strong evidence that Home 5 belongs to this cluster at all as it silhouette coefficient is insignificant. None of the other clusters are strong enough to consider.

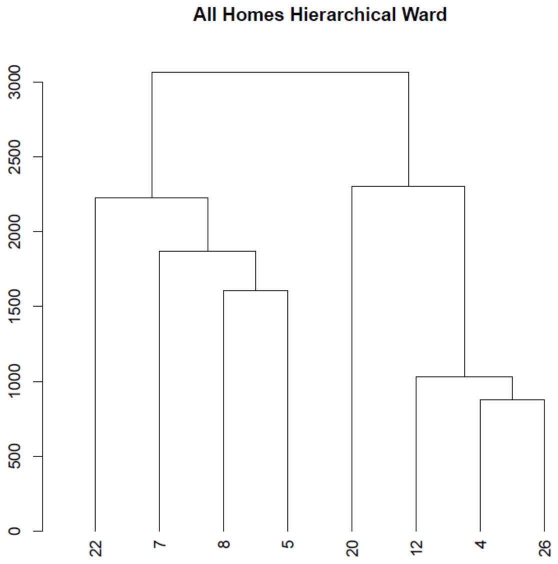

Figure 14. Dendrogram showing results of hierarchical clustering on all homes.

Once again the only cluster of any significance is that containing Home 4, Home 12 and Home 26.

B. Discussion

The results from the cluster analysis looking for similarities in movement across homes are consistent between the different clustering methods. The clustering is not particularly strong between any of the homes in the study. There is some evidence of a single cluster containing Homes 4, 12 and 26. None of the other homes cluster together.

VIII. Methodology – Cluster Analysis – Looking for Movement Patterns over Days within Homes

We also looked at each individual home and looked for patterns of movement over 79 days of data. Again *k*-means and hierarchical clustering was used to look for patterns. Each apartment had four different readings for each of the three sensors on 79 days. This data was used to cluster the days for each home.

A. Results

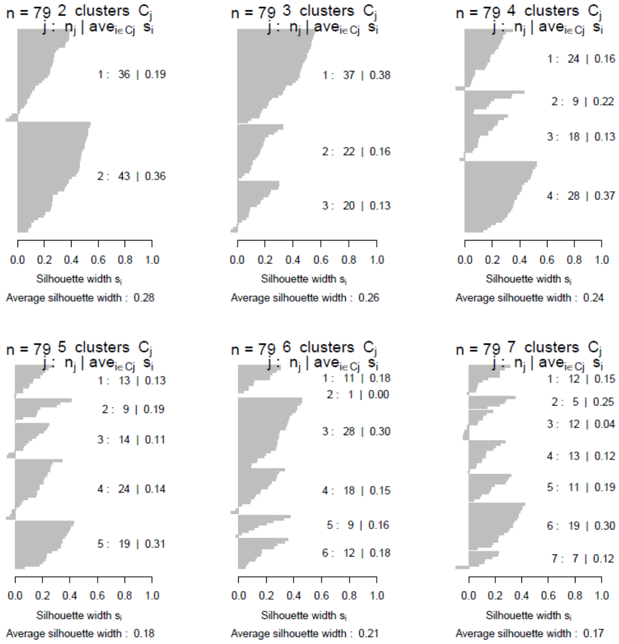

Figure 15. Results of *k*-means clustering with two, three, four, five, six and seven clusters for Home 4.

Figure 15 shows the results of *k*-means clustering for Home 4. There is quite strong evidence that this person has two distinct patterns in their movement throughout the house.

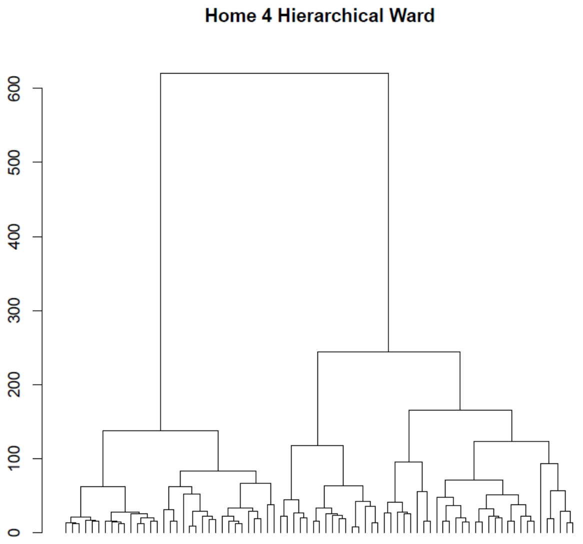

Figure 16. Results of hierarchical clustering over all days in Home 4.

Once again the results of the hierarchical clustering reinforce what the *k*-means clusters elucidated. Again, there is quite strong evidence that this individual has two distinct movement patterns throughout the house.

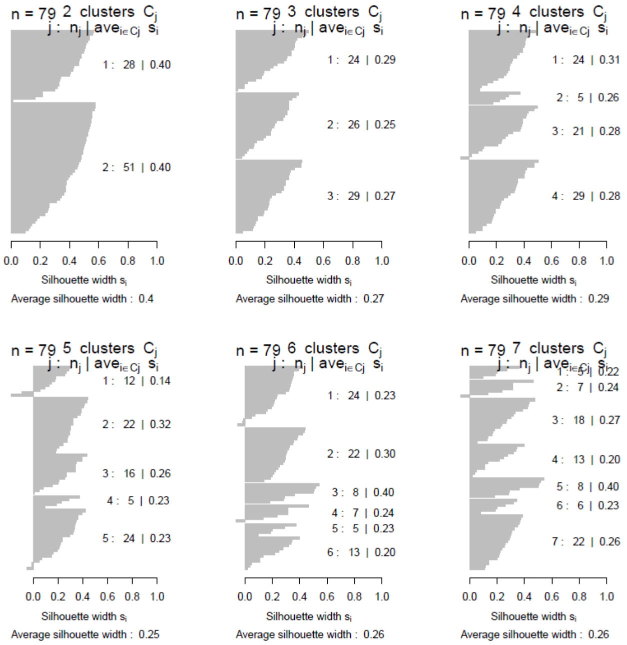

Figure 17. Results of *k*-means clustering with two, three, four, five, six and seven clusters for Home 7.

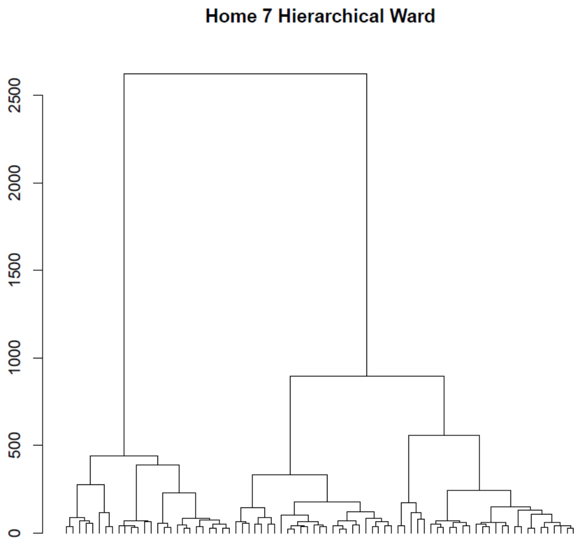

Figure 18. Results of Hierarchical clustering over all days in Home 7.

B. Discussion

The results above show that the two clustering methods give similar results. Both point to Homes 4 and 7 having two broad types of movement patterns. These two movement patterns are very dissimilar from each other. We can also compare Home 4 and Home 7. Home 7 has a stronger routine than Home 4. The larger silhouette coefficients for the clusters for Home 7 indicate that the days within each cluster are more similar. The dendrograms also tell the same story—the dendrogram for Home 7 clusters quicker than that for Home 4. (The numbers on the *y*-axis of the dendrogram just tell the relative difference between items in the cluster.)

All six other homes in the analysis showed a similar pattern of having two distinct movement types. A resident's level of routine is evident from how well the days cluster.

IX. CONCLUSION

This analysis has been a first look at a large amount of data that we are gathering. The clock plots and scatter plots demonstrate that individual homes have distinct movement patterns. These movement patterns are also quite distinct at different times of the day. Most people move very little in the living room between 00:00 and 06:00 and move a lot in the living room between 12:00 and 18:00.

The cluster analysis between individual apartments reinforces the idea that each apartment has distinct movement patterns. Only one relatively weak cluster of three homes was found.

The cluster analysis within each home yielded the most surprising results. Each of the eight homes clustered best when only two clusters were chosen for the *k*-means clustering. This was reinforced by carrying out the analysis using hierarchical clustering and studying the associated dendrograms. The clustering suggests that all eight homes have two distinct movement patterns. Some homes have much stronger clustering than others suggesting that they have a stronger routine and hence their movement patterns are more similar from day to day.

It is important to be cautious in the interpretation of the results of cluster analysis. Cluster analysis is a descriptive technique. The solution is not unique and it strongly depends upon the analyst's choices. Cluster analysis will always provide groups, even if there is no group structure. When applying a cluster analysis we are hypothesising that groups do, in fact, exist. This assumption may be false. Cluster analysis results should not be generalized. Cases in the same cluster are, it is hoped, similar only with respect to the information on which the cluster analysis was based.

REFERENCES

[1] J. Oeppen and J.W. Vaupel, "Broken limits to life expectancy," *Science*, vol. 296, 2002, p. 1029.

[2] S.M. Salvi, "Ageing changes in the eye," *Postgraduate Medical Journal*, vol. 82, 2006, pp. 581-587.

[3] K.S. Helfer and L.A. Wilber, "Hearing Loss, Aging, and Speech Perception in Reverberation and Noise," *J Speech Hear Res*, vol. 33, 1990, pp. 149-155.

[4] T. Hedden and J.D.E. Gabrieli, "Insights into the ageing mind: a view from cognitive neuroscience," *Nature Reviews Neuroscience*, vol. 5, 2004, pp. 87-96.

[5] O.C. Jenkins and M.J. Mataric, "Deriving action and behavior primitives from human motion data," *Intelligent Robots and Systems, 2002. IEEE/RSJ International Conference on*, 2002, pp. 2551–2556.

[6] S.K. Nufer and M. Buehlmann, "Intelligent, Learning Systems ABI Mark II."

[7] S.K. Nufer, M. Buehlmann, T. Delbruck, and J.M. Joller, "A mixture of experts for learning lighting control."

[8] A. Lotfi, C. Langensiepen, S.M. Mahmoud, and M.J. Akhlaghinia, "Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour," *Journal of Ambient Intelligence and Humanized Computing*, 2011.

[9] S.J. McKenna and H.N. Charif, "Summarising contextual activity and detecting unusual inactivity in a supportive home environment," *Pattern Analysis and Applications*, vol. 7, 2005, pp. 386-401.

[10] T. Darrell and A. Pentland, "Space-time gestures," *Proceedings CVPR'93*, Chambéry, France: 1993, pp. 335–340.

[11] K. Takahashi, S. Seki, E. Kojima, and R. Oka, "Recognition of dexterous manipulations from time-varying images," *Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, Texas: 1994, pp. 23–28.

[12] A.G. Kaye, M. Pavel, and T.L. Hayes, "Unobtrusive assessment of mobility," 2006.

[13] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," *Proceedings CVPR'92*, 1992, pp. 379–385.

[14] Y. Guo, G. Xu, and S. Tsuji, "Understanding human motion patterns," *Proceedings of the 12th IAPR International Conference on Computer Vision & Image Processing*, 1994, pp. 325–329.

[15] M. Rosenblum, Y. Yacoob, and L. Davis, "Human emotion recognition from motion using a radial basis function network architecture," *Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, Texas: 1994, pp. 43–49.

[16] L. Atallah and G.Z. Yang, "The use of pervasive sensing for behaviour profiling–a survey," *Pervasive and Mobile Computing*, vol. 5, 2009, pp. 447–464.

[17] "The R Project for Statistical Computing."

[18] P.J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, 1987, pp. 53-65.

[19] J.H. Ward, Jr., "Hierarchical Grouping to Optimize Objective Function," *Journal of the American Statistical Association*, vol. 58, 1963, pp. 236-244.