5-2020

Using continuous sensor data to formalize a model of in-home activity patterns

Beiyu Lin
The University of Texas Rio Grande Valley

Diane J. Cook

Maureen Schmitter-Edgecombe

Recommended Citation

Lin, Beiyu, Diane J. Cook, and Maureen Schmitter-Edgecombe. 2020. “Using Continuous Sensor Data to Formalize a Model of In-Home Activity Patterns.” Journal of Ambient Intelligence and Smart Environments 12 (3): 183–201. https://doi.org/10.3233/AIS-200562.

Using Continuous Sensor Data to Formalize a Model of In-Home Activity Patterns

Abstract. Formal modeling and analysis of human behavior can properly advance disciplines ranging from psychology to economics. The ability to perform such modeling has been limited by a lack of ecologically-valid data collected regarding human daily activity. We propose a formal model of indoor routine behavior based on data from automatically-sensed and recognized activities. A mechanistic description of behavior patterns for identical activity is offered to both investigate behavioral norms with 99 smart homes and compare these norms between subgroups. We identify and model the patterns of human behaviors based on inter-arrival times, the time interval between two successive activities, for selected activity classes in the smart home dataset with diverse participants. We also explore the inter-arrival times of sequence of activities in one smart home. To demonstrate the impact such analysis can have on other disciplines, we use this same smart home data to examine the relationship between the formal model and resident health status. Our study reveals that human indoor activities can be described by non-Poisson processes and that the corresponding distribution of activity inter-arrival times follows a Pareto distribution. We further discover that the combination of activities in certain subgroups can be described by multivariate Pareto distributions. These findings will help researchers understand indoor activity routine patterns and develop more sophisticated models of predicting routine behaviors and their timings. Eventually, the findings may also be used to automate diagnoses and design customized behavioral interventions by providing activity-anticipatory services that will benefit both caregivers and patients.

KEYWORDS

Human dynamics, Population modeling, Pareto distribution, Pervasive environment, Activity recognition

1 INTRODUCTION

The wealth of data that can now be collected by ambient sensors facilitates the development of new models of human behavior supported by empirical evidence. In this paper, we propose formal models of human activities for indoor environments. Specifically, we analyze and model the sequences and timings of basic everyday activities for smart home residents. Offering such models provides a basis for making claims regarding human behavior and differentiating behavior strategies for population subgroups (healthy, dementia). We validate our models by using multiple years of ambient sensor data collected in smart homes. We find that activity arrival rates can be mathematically modeled and that model parameters differ between healthy older adults and older adults with chronic health issues. These analyses allow researchers to better understand the impact of health conditions on routine behavior and can be used to predict diagnosis categories for individuals based on automatically-sensed activity patterns.

Due to limitations with real-world data collection methods, previous models for human activities did not provide sufficient information about the dynamic property of human behaviors. They typically assumed that human activities can be modeled by Poisson processes and that the inter-arrival time, or the time interval between two successive activities, follows an exponential distribution. This assumption models activities as occurring at a constant rate [1]–[4]. However, this model does not capture the fluctuation that may occur in activity arrival rates. With the advent of quantifiable mobile data that can be collected unobtrusively and continuously, researchers recently proposed the use of heavy-tailed distributions to describe human dynamics [5]–[9].

Our approach in this paper is to build a general model of human activities that involves real-time data collection in everyday environments based on ambient sensor data collected in smart homes. We perform our data collection on subjects inside their own homes. The data collection reflects routine human behavior without requiring any alteration to the environment or activities, facilitating an ecologically valid analysis. We analyze the inter-arrival times of automatically-labelled smart home sensor data (e.g., cooking, eating), and find activity interdependencies in subgroups (healthy older adults and older adults with chronic health problems). To investigate the relationship between behavior changes and health problems, we use a case study with 65 months of data from one smart home. This behavior-driven sensor data shows that activity routines can be modeled by non-Poisson processes. The activity inter-arrival times follow a heavy-tailed distribution, specifically a Pareto distribution.

We find that model parameters for activity arrival rates differ between healthy older adults and older adults with health issues. The resulting mathematical models open up the possibility of recognizing the development of health problems and providing efficient interventions and assistance. Once differences in patterns among subgroups are found, they can be used to better understand the impact of culture, age, and education on daily routines. The design of technology-based tools such as agent- and human-oriented software and hardware systems [8], [10]–[12] can also greatly benefit from this work. Researchers in the fields of sociology, psychology, and anthropology will also be able to align their studies with customized and personalized healthcare systems.

Our study provides evidence to support three hypotheses of human routine behaviors in home environments. First, human behavior can be described by formal statistical distributions. Second, data supporting this conclusion can be collected using ambient sensors in an ecologically valid manner. Third, the Pareto model and its properties, such as the 80/20 rule, can be useful for the study of human dynamics and investigation of hypotheses because of its ability to model human behavior patterns. Our study first analyzes and models inter-arrival times of identical indoor behaviors based on both 99 smart homes and subgroups of older adults. We further study the inter-arrival times of activity sequences from one smart home. The findings of this study will offer the potential to automate diagnoses and design customized behavioral interventions.

2. LITERATURE REVIEW

Maturing pervasive computing technologies have sparked a new wave of human behavior analysis and resulted in new theories regarding human behavior patterns. Barabási's study of the timing of consecutive electronic and physical mail messages sparked a model of human dynamics as a heavy-tailed distribution [5], [13]. A queuing model and heavy-tailed distribution were introduced in Barabási's study to explain the large time gap between sent messages after a burst of responses.

After Barabási's discovery, scientists use heavy-tailed distributions to explain human behavior in diverse domains, ranging from social science to health care. In the social network field, heavy-tailed distributions are used to characterize the dynamics of popularity based on diverse digital platforms, such as Wikipedia, blog posts, Android applications, Web pages, and Twitter [14]–[20]. As an example, Li et al. show that the behavior-based popularity of Android applications follows the Pareto principle [17]. Tsompanidis et al. also discover that web traffic flow size can be explained by the Pareto distribution [19]. Similarly, researchers presented a list of social and organizational power laws, one kind of heavy-tailed distribution, to describe human behavior [21]–[23]. Specifically, the power law distribution identifies the number of inter-firm relationships observed from linkages between firms: suppliers, customers, and owners [22], [23].

Further, scientists use heavy-tailed distributions to model and predict human mobility [24]–[30]. For example, GPS-based human movement patterns can be captured by heavy-tailed flights for different transportation modes, including walking/running and car/taxi [28]–[30]. Regardless of transportation modes, the distribution of user's moving distances, from visited locations to the target location, can be modelled by the Pareto distribution [27].

Besides heavy-tailed distributions, other mathematical models are also used to uderstand a more varied set of human activities than basic movements. A mixture of Gaussian intensities model is introduced to explain activities, such as exercising and eating, that have time-varying, interdependent, and periodic properties [31]. The temporal granularity algorithm, considering behaviors happened within a time interval instead of at an exact timestamp, is used to identify frequent behavioral patterns, such as receiving a call, sending/receiving a text message, and holding a meeting [32].

In addition to mathematical formalisms, researchers adopt machine learning methods to model aspects of human behavior [33], [34]. For example, inverse reinforcement learning (a method which flips the problem of traditional reinforcement learning and learns an agent's rewards by observing its behavior) models human driving routines to help aggressive drivers improve their driving style [33] and to find taxi driver's preferences on working regions and times [34].

Given the development of these diverse models to understand human behavior patterns, we propose to extend previous work further by modeling indoor behavior patterns based on ambient sensor data. Although researchers have analyzed raw sensor data and design features from the raw sensor data to understand human behaviors in smart environments [35]–[38], the substantial number and diversity of raw sensor event patterns are difficult to provide a rich vocabulary to express human behavior. Analyzing the labeled activities from sensor event sequences resolves the concern.

In this paper, we model generalized human behavior based on ambient sensor data collected in smart homes. Other work has similarly focused on labeling and analyzing smart home-based behavior. Some of this prior research introduces data-driven techniques for recognizing or predicting daily activities in smart home environments based on continuous sensor data [39]–[46]. Based on raw or activity-labeled sensor data, other studies analyze and assess and individual's physical and mental health stated associated with the observed behavior [47], [48]. Outside of the health domain, researchers have also analyzed behavior patterns from ambient sensor data to predict the associated energy consumption [49], a useful step in designing energy-efficient automated buildings. However, these techniques do not offer a mechanistic description of indoor behavioral patterns. Furthermore, they have not yet attempted to describe behavior patterns at a population level.

Our work explores several research problems. First, we utilize activity recognition methods to label sensor data in real time with corresponding activity labels. Second, based on this labeled data, we analyze activity inter-arrival times and construct heavy-tailed distributions, specifically Pareto distributions, to describe routine patterns for smart home residents in everyday environments. Third, we investigate the patterns of selected activities both at a group level with 99 smart homes and between subgroups of older adults (healthy, chronic condition) to have a generalized understanding of behavior patterns and their differences across a population. Fourth, we analyze the information from our model to determine its value as an indication of a person's health status.

3. METHODOLOGY

We propose a method to formally model activity timings from behavior-based sensor data. First, we monitor sensor events and automatically label the events with corresponding activity names using machine learning models [41]. We then use change point detection (CPD) [50] to segment data into sequences that represent single, uninterrupted activities. Once the data is segmented, we apply a well-known statistical method, extreme value theory (EVT) [28]to remove noise. We use the remaining data to perform distribution fitting of the histogram of activity inter-arrival times. We model the data for 82 different probability distribution functions (pdf) and determine the best distributions based on minimizing the summation of the squared errors (SSE). We utilize non-Poisson processes to model the inter-arrival times of human behavior routines and postulate that activity inter-arrival times can be approximated by Pareto distributions. The modeling steps are illustrated in Fig. 1.

We repeat these modeling steps for each activity separately both for a complete sample of 99 smart home residents as well as for two subgroups of older adults within the sample: healthy and chronic cognitive or physical health conditions. We test the hypothesis that differences in health status between subgroups may be significantly reflected by patterns of each activity. Analyzing inter-arrival times for one activity at a time is reflective of analysis techniques used in previous studies [51]–[54]. Furthermore, focusing on a single activity allows us to understand the potential relationship between the activity of interest and population subgroups as well as identify differences in model parameters between activities. On the other hand, analyzing individual activity may prevent us from holistically examining a person's entire behavior routine. Thus, we additionally analyze the entire activity sequence patterns for one of the smart homes.

3.1. Data Collection and Processing

In this study, we collect data from 99 smart homes to investigate routine behavior patterns for selected activities. We provide details on the first four steps of the process in Fig. 1: data collection, activity labeling, segmentation, and noise filtering.

3.1.1. Data Collection

Data are collected using the CASAS Smart Home in a Box (SHiB) [41], [55]. In each smart home, four types of ambient sensors are installed: infrared motion, magnetic

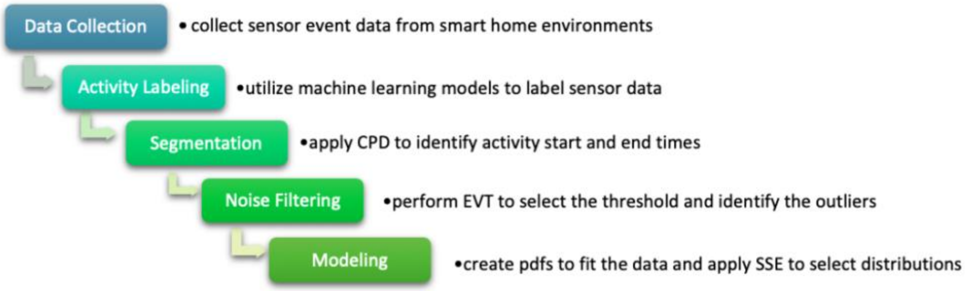

Data Collection
• collect sensor event data from smart home environments
Activity Labeling
•utilize machine learning models to label sensor data
Segmentation
•apply CPD to identify activity start and end times
Noise Filtering
•perform EVT to select the threshold and identify the outliers
Modeling
•create pdfs to fit the data and apply SSE to select distributions

Figure 1. The steps of population-based activity modeling. Ambient sensor data is collected from smart homes, labeled with corresponding activities, and segmented. The data is cleaned then modeled based on probability density functions to select appropriate distributions.

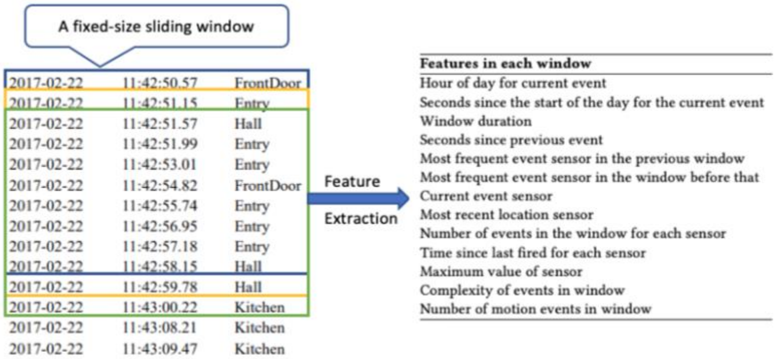
(door) contact, light level, and temperature level. These sensors are discrete event sensors and thus only generate a message (sensor event) when there is a change in a state, such as a refrigerator door opening or closing. The sensors are installed throughout the house in each room including the kitchen, living room, dining room, bedrooms, bathrooms, office, and laundry room. Infrared motion sensors include narrow-area and wide-area sensors. Narrow-area motion sensors detect heat-based movements within a one-meter diameter area. They are attached to the ceiling above specific objects or areas in the home, such as above a participant's favorite chair or above a sink. Wide-area motion sensors perceive movements occurring anywhere in an entire room. These sensors are placed on ceiling corners in large rooms, such as the dining or living room. Magnetic sensors detect the use of doors and cabinets, such as in entering or leaving the home or accessing items within kitchen or bathroom cabinets. Temperature sensors can be useful in detecting activities that change the heat level in an area of the home, such as showering or cooking. Similarly, light sensors can help us identify activities occurring within a home as well as seasonal effects of light levels.

3.1.2. Activity Labeling of Sensor Data

Activity labeling provides us with a rich vocabulary to express human behavior. We employ automated activity recognition techniques to label collected data with eleven activity classes. The set of activities that we categorize and use in this analysis are seven activities: Relax, Cook, Eat, Personal Hygiene, Wash Dishes, Sleep, and Work. We use a separate class, Other Activity, to recognize unidentified sensor events.

We apply automated activity recognition (AR), a heavily-investigated challenge [31], [34]–[38], to map a sequence of captured sensor events onto one of the activity classes. Our AR steps are based on an approach that has been previously-validated for real-time activity recognition from streaming sensor data [39]. First, we extract features from the raw data collected from the discrete event sensors (see Fig. 2). We move a fixed-size sliding window over the time-ordered sensor data and compute feature values for each window [33], [39], [40]. Within each window, sensor events may be widely spread apart in time. To take this into account, a time-based weighting factor is applied to account for the relative temporal distance between sensor events. Second, after training a random forest classifier with ground truth pre-labeled sample data, the resulting model can provide activity labels for data based on features extracted from sensor sequences. The sequence of sensor events in a window provides a context for labeling the last (most recent) sensor event. The approach we utilize is distinctive in that it is designed to provide activity labels in real-time from ambient sensor data collected continuously in real homes and to build a generalizable model based on training ground truth-labeled data. After training, the resulting model can provide activity labels for data obtained in new smart home settings. Our approach to activity recognition yields an average of 95% accuracy and 0.78 F1 scores for activity labeling based on the three-fold cross-validation method [41].

After applying AR, all unidentified sensor events are assigned to the class, Other Activity. The drawback of putting all undefined activities into a single Other Activity class is that we cannot distinguish those activities from each other. Each of them may shed light on the behavioral routines for the residents but because they are grouped together, they cannot be analyzed as individual activities. Therefore, for the Other Activity category, we employ a k-means clustering algorithm to divide the category into a specified number, *k*, of clusters. The value of *k* is chosen using an elbow curve method, which shows the minimized sum of squared distances of samples from the closest cluster center. Thus, for each home, our activity classes include both the seven predefined and the cluster-generated activity classes, which are labeled cluster$_1$ through cluster$_k$.

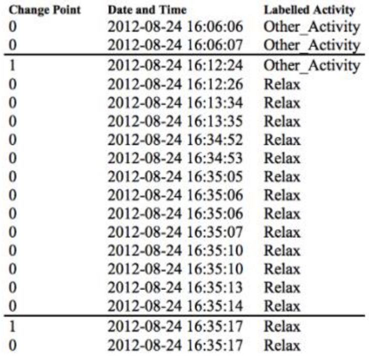


2017-02-22	11:42:50.57	FrontDoor
2017-02-22	11:42:51.15	Entry
2017-02-22	11:42:51.57	Hall
2017-02-22	11:42:51.99	Entry
2017-02-22	11:42:53.01	Entry
2017-02-22	11:42:54.82	FrontDoor
2017-02-22	11:42:55.74	Entry
2017-02-22	11:42:56.95	Entry
2017-02-22	11:42:57.18	Entry
2017-02-22	11:42:58.15	Hall
2017-02-22	11:42:59.78	Hall
2017-02-22	11:43:00.22	Kitchen
2017-02-22	11:43:08.21	Kitchen
2017-02-22	11:43:09.47	Kitchen

Features in each window
Hour of day for current event
Seconds since the start of the day for the current event
Window duration
Seconds since previous event
Most frequent event sensor in the previous window
Most frequent event sensor in the window before that
Current event sensor
Most recent location sensor
Number of events in the window for each sensor
Time since last fired for each sensor
Maximum value of sensor
Complexity of events in window
Number of motion events in window

Figure 2. Using a fixed-size sliding window to extract features from the raw sensor data.

The activity recognition algorithm labels each sensor event with a corresponding activity class. The algorithm does not, however, indicate the beginning or ending of each activity occurrence. This information is valuable for our analysis because we want to consider the inter-arrival times of each activity's start as part of a person's overall routine. To segment labeled data into individual activities, we utilize an unsupervised method referred to as Change Point Detection (CPD). CPD identifies the point in time where the state of the underlying process changes [56], [57]. CPD can be used to detect real-time activity transitions or changes in the data between two successive windows of sensor events [58]. An example of sensor events with corresponding times, activity labels, and detected change points is shown in Fig. 3, where transitions are indicated by horizontal lines. The first event in an activity segment (the first sensor event after a change point) is considered the start of an activity and the last event in a segment (the last sensor event before a change point) represents the end of an activity. In this study, we use a CPD method that is based on Bayesian online learning [59]. Given the segmented data, we label each segment with the most majority activity label for that segment. The labeled activity in each segment provides us the time and activity information and allows us to calculate the inter-arrival times of each activity (in hours), defined as the time between two successive start times of the activity.



Change Point	Date and Time	Labelled Activity
0	2012-08-24 16:06:06	Other_Activity
0	2012-08-24 16:06:07	Other_Activity
1	2012-08-24 16:12:24	Other_Activity
0	2012-08-24 16:12:26	Relax
0	2012-08-24 16:13:34	Relax
0	2012-08-24 16:13:35	Relax
0	2012-08-24 16:34:52	Relax
0	2012-08-24 16:34:53	Relax
0	2012-08-24 16:35:05	Relax
0	2012-08-24 16:35:06	Relax
0	2012-08-24 16:35:06	Relax
0	2012-08-24 16:35:07	Relax
0	2012-08-24 16:35:10	Relax
0	2012-08-24 16:35:10	Relax
0	2012-08-24 16:35:13	Relax
0	2012-08-24 16:35:14	Relax
1	2012-08-24 16:35:17	Relax
0	2012-08-24 16:35:17	Relax

Figure 3. A sample of CPD application to the sensor data. A change point value of 1 indicates a transition/activity start time, 0 indicates no change. The 0 right before the next transition is the end time of an activity. A transition is detected and shown by a horizontal line.

3.1.3. Participant Information

In addition to collecting sensor data for 99 smart homes, we also store four additional parameters for each home: the number of residents as well as resident ages, education levels, and physical and mental health statuses (where available). Our sample includes single-resident (46%), two-resident (18%), three or more-resident homes (4%), and a not-reported category (32%). Residents can be categorized as young (age <35, 14%), middle-aged (age 35-64, 9%), senior (age >64, 65%), and a not-reported category (12%). Education levels of residents in our dataset varies, including a high school diploma (10%), bachelor's degree (19%), master's level (20%), doctorate degree (15%), and a not-reported category (36%). Our entire 99 smart home dataset includes people who are healthy (57%) and those with targeted health ailments (23%) such as mild cognitive impairment (MCI) (9%), as well as those whose health status was not reported (20%).

While we have collected a large set of sensor data for this analysis, the data may not be representative of the population as a whole. Thus, we employ different indices to determine how representative our data are of the national population. Information statistics (Shannon index) and dominance (Simpson index) indices are utilized to identify and quantify both the richness (number of subgroups present) and abundance (the number of individuals per subgroup) of our smart home dataset in comparison with the US population. We also utilize mean and variance analysis to investigate the composition of the dataset. Further, we utilize Jaccard's coefficient index, a value between 0 (not similar) and 1 (identical), to compare the similarities between our smart home dataset and the US population. The data of the US population in 2010 is collected from the census government website [60]–[63].

For the information statistics, Fig. 4 shows that the value of Shannon indexes for age in our sample (1.01) is close to that at the national average (1.03), reflecting the richness and abundance of our smart home dataset. We further use the Simpson index to analyze the dominant subgroup (see Fig. 5) as well as mean and standard deviation to analyze the composition in both datasets (see Fig. 6). In Fig. 5, for the category of education level, both our smart home dataset and the national population have a Simpson index greater than 0.6, reflecting diverse education levels and no dominant subgroup. For age and number of residents, the Simpson index in our smart home dataset is less than 0.5 (dominant subgroups may exist). The mean and deviation charts (see Fig. 6) show that the mean value of age in our sample (76) is over twice that at the national level (37). In respect to household size, our sample does include fewer residents on average (1.4) than that in the national population (2.6). Our dataset is more representative of senior residents and low-population homes. The values of the Jaccard index for the categories of age(s), number of residents, and education level are 0.14, 0.14, and 1, respectively. That is, in the category of education level, our dataset is similar as the national population. A complete list of home descriptive parameters is provided in the supplementary material.

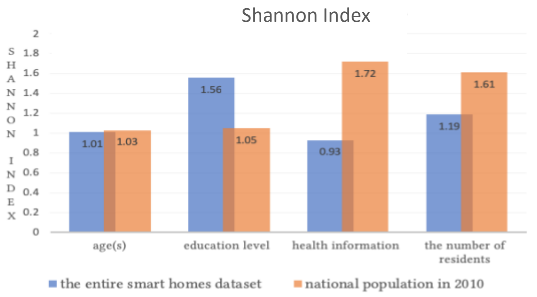


Figure 4. Shannon Index (information statistics index) of smart home dataset and national population in 2010.

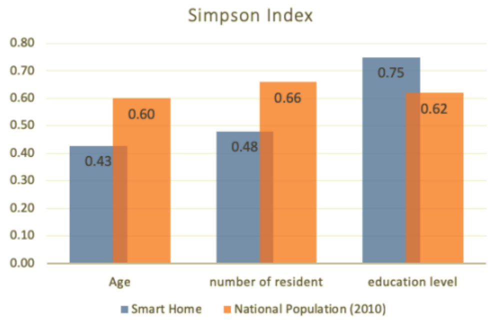


Figure 5. Simpson Index (dominance index) of smart home dataset and national population in 2010.

To analyze differences in behavior models between different population subgroups for individual activities, we select two subgroups among our smart home participants who are matched in terms of age and number of residents. The first subgroup consists of senior residents who are healthy and living alone (Subgroup H). This represents a baseline group for a comparison to senior residents who are living alone and have chronic health ailments. Most of these residents had multiple health problems. The most significant limiting conditions included mild cognitive impairment (N = 4) and mild dementia (N = 3). Further, one resident has Parkinson's Disease, 4 people have mobility limitations, 2 residents have lung problems (chronic hypoxia/chronic obstructive pulmonary disorder), 2 participants have atrial fibrillation, and 1 resident has macular degeneration. . To keep our sample sizes as large as possible, we did not constrain education level for these participants. There are 16 homes included in Subgroup H and 17 homes included in Subgroup NH. To study an entire sequence of activities, we selected a home with over five years of collected data. The resident of this selected home also experienced health changes during the data collection period. Specifically, the resident has vision and mobility problems.

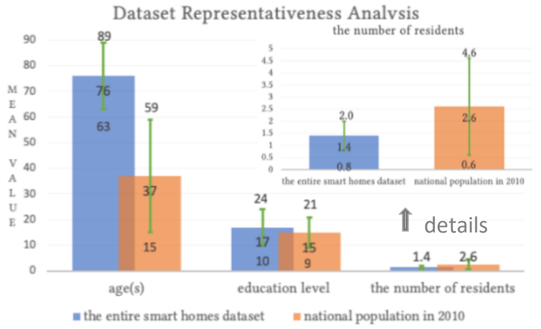


Figure 6. Mean and variance of each category from smart home dataset and national population in 2010.

3.2. Data Cleaning

Before we fit a model to our smart home data, we preprocess the inter-arrival times of activity segments to remove noise. Noise can arise in smart home data due to issues including sensor failure, visitors in the home, activity recognition/segmentation errors, or changes that are made to the environment. While some of the outliers may represent behavioral changes that need to be captured, others represent errors in the data collection process and are best removed.

Our study focuses on large sets of continuous real-valued data. Longitudinal data collected from real homes is also subject to noise resulting from imperfect sensors and related system issues. As a result, we first apply outlier detection to the activity inter-arrival times by using a threshold exceedances approach, a principal approach found in extreme value theory (EVT). This approach allows us to test a range of threshold values u and identify outliers which have values above the threshold. For each candidate threshold, we fit a distribution for the excesses, the difference between the outlier and the threshold. Lower thresholds tend to bias the excess model by categorizing a large amount of data as outliers. Higher thresholds lead to a greater variance of the excess distribution because of the small number of outliers. The standard rule is to choose a threshold as low as possible so that the excesses fit a reasonable distribution [64], [65].

Here, a reasonable distribution is governed by two factors. First, we strive for a balance between bias and variance of the excesses distribution. Mean residual life plots help visualize this balance. A mean residual life plot graphs the mean value of excesses as a function of the threshold value. We select the threshold value at the lowest threshold value to show linearity in the plot. Linearity indicates that the bias and variance of the excess distribution are nearly evenly balanced. As an example of this approach, Fig. 7 hows the mean residual life plot with 95% confidence intervals for the inter-arrival times of the "Personal Hygiene" activity for our sample of 99 homes. In Fig. 7(a), the x-axis shows a range of threshold values (u) from the minimum to the maximum values observed Personal Hygiene inter-arrival times. The y-axis shows the mean excess (the mean of the excess times above the threshold) for each threshold value.

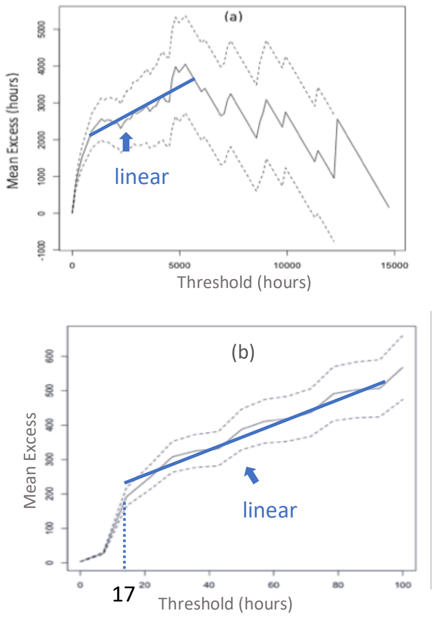


Figure 7. (a) The mean residual life plot of Personal Hygiene inter-arrival times for the entire dataset. Dashed lines represent the 95% confidence interval. The solid line plots the threshold value against the mean excess. Both the threshold and excess values are in hours. The graph is near-linear at u = 1000 hours; (b) The mean residual life plot of the Personal Hygiene inter-arrival time for a range of thresholds from 5 to 20. Dashed lines represent the 95% confidence interval. The solid line represents the threshold value against the mean of excesses. This graph is near-linear at u = 17 hours.

To balance the bias and variance of the excess distribution, we identify a threshold value where the solid line in Fig. 7(a) shows linearity. We notice that the graph appears to be approximately linear around u = 1000 hours. Since the value at the 99th percentile of the entire dataset of the inter-arrival time of Personal Hygiene is 13.1 hours, the value 1000 (hours) can be considered a high threshold. To reduce the variance of excess model fitting, we choose a threshold based on a mean residual life plot for a range of thresholds from 5 to 20 (hours) as shown in Fig. 7(b). This plot suggests an upper threshold (the maximum inter-arrival times) of approximately 17 hours with 2835 outliers out of 366,441 data points, or 0.774% of the total number of data points.

In the second step, we refine the threshold choice to ensure that the shape parameter (affecting the shape of a distribution instead of shifting or stretching/shrinking the distribution) and modified scale parameter (affecting the stretching/shrinking of a distribution) of the excess distribution are quantifiably stable. To do this, we select the lowest threshold value, near the approximated threshold from the first step, for which both the estimated shape parameter remains near-constant and the estimated modified scale parameter is near-linear [64].

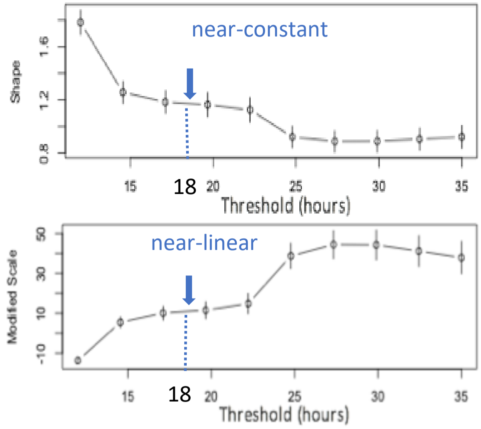


Figure 8. Parameter estimates against a range of thresholds from 10 to 35 hours from the Personal Hygiene inter-arrival times. We select the lowest value (u = 18 hours, the maximum inter-arrival times) of thresholds, near the approximated threshold (u = 17 hours, the maximum inter-arrival times), for which the estimated shape (affecting the shape of the distribution rather than shifting or stretching/shrinking it) parameter remains near-constant and the estimated modified scale (affecting the stretching/shrinking of a distribution) parameter is near-linear.

Using the same example of Personal Hygiene inter-arrival times as in the first approach, Fig. 8shows shape and re-parametrized scale parameters as a function of alternative threshold values. Based on the plots in Fig. 8, which offer a model-based analysis of excess, we choose a threshold of u = 18 hours (the maximum inter-arrival times) with 2730 outliers out of 366,441 data points (0.745%). The perturbations among the excesses are small relative to sampling errors based on the stability of parameter estimates in Fig. 8.

4. MODEL FITTING

Modeling human behavior from smart home sensors provides a unique perspective not only to investigate behavioral norms but also to compare these norms between population subgroups. Once differences in patterns are discovered, they can be used to better understand the impact of personal characteristics, such

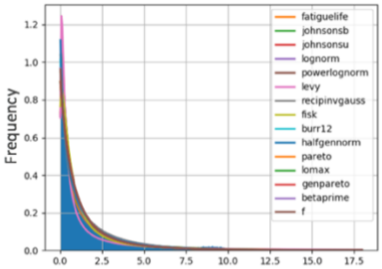


Figure 9. The top 15 fitted distributions, selected by SSEs, among 82 probability distributions.

as age, health conditions, and education on daily routines. They can also be used to automate diagnoses and predict additional behavioral features of individuals within a group. In this section, we provide details of our model fitting procedures. To illustrate the process, we focus on the Personal Hygiene activity observed from all the sampled smart homes.

We use the data below the selected EVT threshold to determine which distribution best describes the sensor-based activity data. We model data histograms using 82 different probability distribution functions. In our study, we employ the Freedman–Diaconis (F-D) rule to select the size of the bins for the histogram. In general, three well-known rules are used to calculate bin widths: the Sturges, Scott, and F-D rules. The Sturges rule is applicable when the data is from symmetric and Gaussian distributions [66]. The Scott rule works well for non-Gaussian distributions but refers to sample sizes between 50 and 500. For larger samples as in our study, the Scott rule estimates a smaller number of histogram bins, leading to over-smoothing. Over-smoothed histograms provide limited information on the shape of the underlying distribution [67]. The F-D rule is a robust method that substitutes the estimated standard deviation in the Scott rule with a multiple of the interquartile range. Thus, the F-D rule ensures 35% more bins than from the Scott rule as well as keeps the property of the Scott rule for non-Gaussian distributions [68], [69].

Next, we use the smallest summation of the square errors (SSE) value to compare distributions and select the distributions that best fit the data. SSE is calculated by determining the difference between the data and the fitted distribution. Here, we give an example of the distribution fitting and selection process for the Personal Hygiene inter-arrival times from all smart homes. The results of modeling the remaining activities and comparing the two subgroups are provided as supplementary material. While we use 82 distributions to fit the histogram, Fig. 9 shows the top 15 fitted distributions for the Personal Hygiene inter-arrival times from all the smart homes. The Pareto distribution is selected as one of these top distributions. Fig. 10 summarizes the SSEs between the smart home data and the top 15 fitted distributions. The Pareto distribution has the same order-of-magnitude errors (10^{-1}) as the other top distributions. We hypothesize that the Pareto distribution provides a close approximation to the top fitted distribution for Personal Hygiene inter-arrival times based on the sampled 99 smart homes. We test this hypothesis by both visualizing the fitting and determining the significance of the difference in fit between the Pareto distribution and the top-fitting distribution.

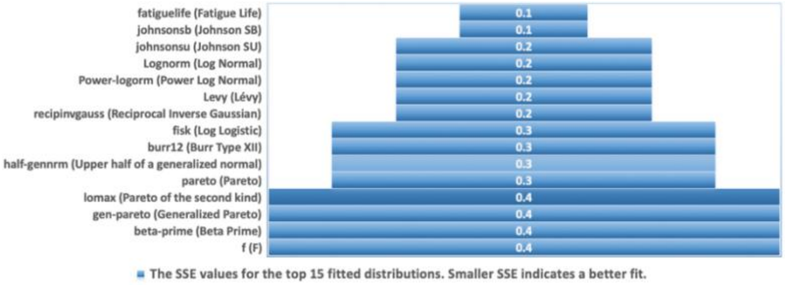


Figure 10. The SSE values for the top 15 fitted distributions. Smaller SSE values indicate a better fit.

First, the figure on the left side of Fig. 11 shows the fit of the Pareto distribution and the top-fitting distribution for the Personal Hygiene inter-arrival times across all smart homes. Figures on the right side, (b) and (c), of Fig. 11, respectively, show portions of the fitted Pareto distribution for the Personal Hygiene inter-arrival times from 0 to 6.5 (hours) and from 3 to 18 (hours). Based on Fig. 11, the fitted Pareto distribution well approximates the shape of the histogram of the Personal Hygiene inter-arrival time in the entire dataset,

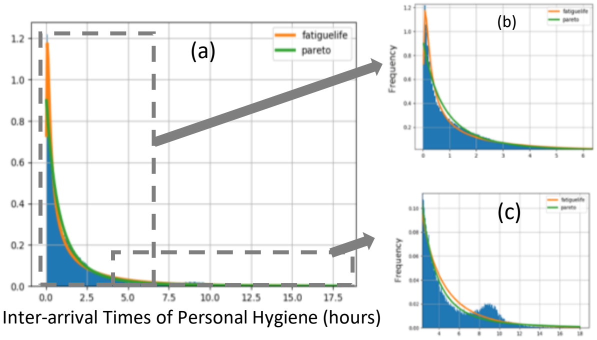


Figure 11. (a) Distribution fitting between the Pareto distribution and the top-fitting distribution. Bins are indicated by the x-axis. The y-axis represents frequency, the amount of data included in each bin divided by the total amount of data. The SSE of this Pareto distribution is 0.3. (b) Portion of the graph corresponding to shorter inter-arrival times (from 0 to 6.5 hours). (c) Portion of the graph corresponding to longer inter-arrival times (from 3 to 18 hours).

though the distribution did not capture everything from the histogram. For example, a hump exists (see Fig. 11 (c)) around hours 8 through 10 with frequencies 0.01 to 0.02. Because, we are capturing a general view of indoor behavior patterns rather than modeling each detail of a single activity, we may also be observing overfit.

Second, to validate that the Pareto model provides a statistically significantly-similar fit to the top-fitting distribution, we perform a t-test analysis with the null hypothesis that the two distributions have identical fit scores. Given the activity inter-arrival times and the estimated distribution parameters, we generate the values of probability density functions for the Pareto distribution and the top-fitted distribution. Next, we split the two sets of values into 60 subsets and perform a paired t-test on the means for each subset.

For Personal Hygiene inter-arrival times across all sampled homes, the p value is 0.153 with the t-statistic -1.443. For the null hypothesis that the two distributions have identical average scores, a small p value (<= 0.05) leads to rejecting the null hypothesis and a large p value (>= 0.05) leads to accepting the null hypothesis. Thus, we accept the null hypothesis, and the Pareto distribution can be considered approximately as strong as the top-fitting distribution for the Personal Hygiene inter-arrival times from the entire collection of smart homes. Similar results were observed for all selected seven activities.

Based on both the visualization and t-test, the Pareto distribution provides a strong fit for this activity data and the properties of the Pareto distribution, for example the 80/20 rule, provide opportunities for future analyses and investigations of hypotheses that model indoor behavior patterns.

5. RESULTS

To understand the general principles behind human behavior in everyday environments and to compare the behavioral norms between population groups, we perform the same procedures described in Sections 3 and 4 for each recognized activity both across all 99 smart homes and among two older adult subgroups (Subgroup_H and Subgroup_NH). In addition, using the same procedures as in Sections 3 and 4, we study a holistic behavior routine in one home as a combination of all detected activities.

For the data from 99 smart homes, before performing outlier detection on the inter-arrival times of each activity, we visualize some statistics to gain an intuitive understanding of the data (see Fig. 12). In these graphs, we observe that the maximum value of the inter-arrival times of each activity is relatively large (>= 10_3 hours). There are multiple possible explanations for these large values, including sensor failures and the resident's absence from the home during travel. The 99th percentile of inter-arrival times for each activity is in the range of 10_1 to 10_2 hours. That is, approximately 99% of occurrences for each activity exhibit small inter-arrival times, thus only the top 1% of inter-arrival times demonstrate these large values of interest. The mean inter-arrival time for each activity is in the range of 10_0 to 10_1 hours. For example, the mean value of the inter-arrival times of two successive Cook activities from 99 smart homes is 3.5 hours, which gives us a generalized view about the gap between two successive Cook activities. In Fig. 13, we notice that over 99.5% of the data are kept after removing outliers. In addition, the threshold value of each activity is above their mean value (except activity Eat) and sometimes above the 99th %.

After filtering the outliers (using methods described in Section 3), we perform model fitting as described in Section 4. The summarized result of fitting of activity inter-arrival times is shown in Fig. 14 and Table 1. In Fig. 14, we noticed that the shape parameter of the Pareto distribution for Sleep inter-arrival times from the entire smart home dataset is less than one. This means that the

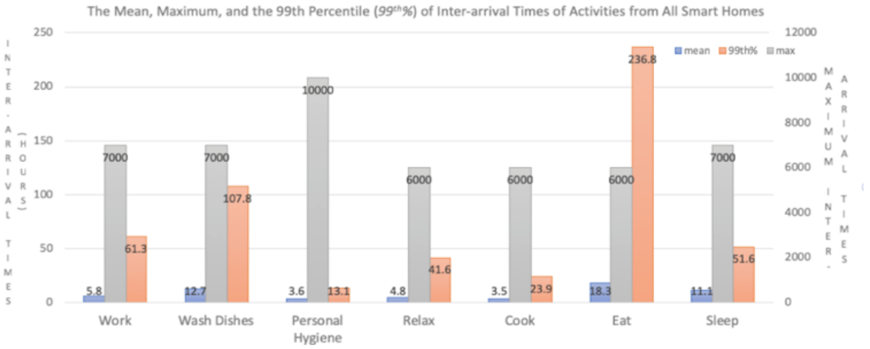


Figure 12. A summary of inter-arrival times of all activities for all smart homes before performing outlier detection. The results include the mean value of the dataset (mean), the 99th percentile (99th%) and maximum value of the dataset (max).

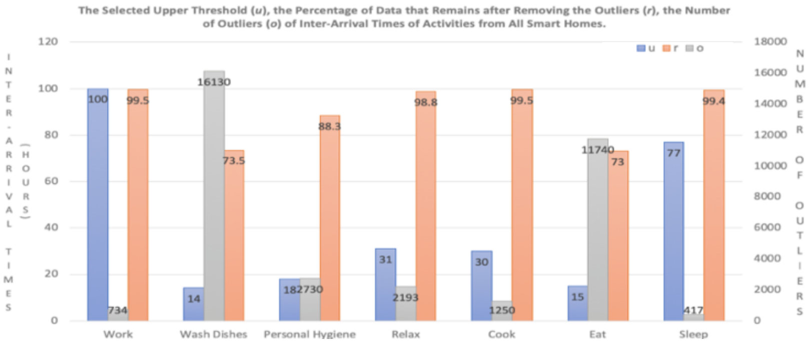


Figure 13. Summarized results of all activities for all smart homes. The results include the selected upper threshold (u), the percentage of data that remains after removing the outliers (r), and the number of outliers (o).

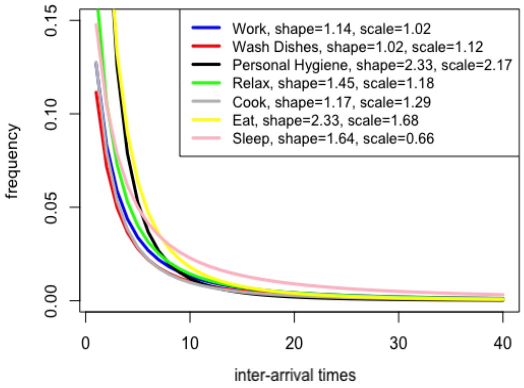


Figure 14. The Pareto distribution of each activity given simulated inter-arrival times from 0 to 40 hours.

expected start time of a Sleep activity relative to the previous Sleep occurrence approaches infinity. This result occurs because the mean value of the dataset is influenced by the largest single value. This may occur in finite-size samples when an outlier causes the mean to become arbitrarily large. The Sleep activity arrival times therefore cannot be adequately captured by distributions and thus we will use quantiles to describe the data spread of the Sleep activity.

Based on Fig. 15, we observe that for each activity, the SSE of the Pareto distribution fit is relatively small, in the range of 10^{-3} to 10^{-1}. Furthermore, the Pareto distribution for each activity is approximately as strong as the top-fitting distribution based on the large t-test p values (>= 0.05) in Table 1 (the null hypothesis is that the two distributions have identical mean scores).

To further interpret the behavioral norms, we compare the selected thresholds (u) and the Pareto shape parameters (α) for the entire sampled population and among population subgroups (see Figures 16 and 17). In Fig. 16 (a), the threshold of the inter-arrival time of Work, Relax, Cook, Eat, and Sleep in Subgroup NH is smaller than the corresponding threshold in Subgroup H.

One possible explanation for this difference is that individuals in Subgroup NH may be unable to sustain long periods of one activity, thus creating bursts of activities with short breaks. The phenomena might be due to physical health ailments, such as mobility or stamina difficulties that may require periods of rest. In addition, participants with cognitive limitations may be more likely to get distracted or experience difficulty remembering to quickly return to a task following an activity interruption, resulting in the need to reinitiate an activity within a short period of time.

Fig. 17 ows the values of the shape parameters of the Pareto distributions for the individual selected activities, the variations of which may be due to variations in health conditions. In Fig. 17(a), we observe that the shape parameters for activities Personal Hygiene, Relax, Cook, and Sleep in Subgroup H are smaller than the shape parameters in Subgroup NH. The smaller the shape parameter, the heavier the tail is of the Pareto distribution. That is, longer starting times between two successive activities occur more frequently in Subgroup H, possibly due to fewer interruptions.

Table 1. The p value of the t-test between the top- fitting distribution and the Pareto distribution (p).

	Top Fitted Distribution	p value		Top Fitted Distribution	p value
Work	Burr	0.540	Relax	Inverse Gaussian	0.283
Wash Dishes	Fatigue life	0.312	Cook	Lévy	0.247
Personal Hygiene	Fatigue life	0.153	Eat	Fatigue life	0.095

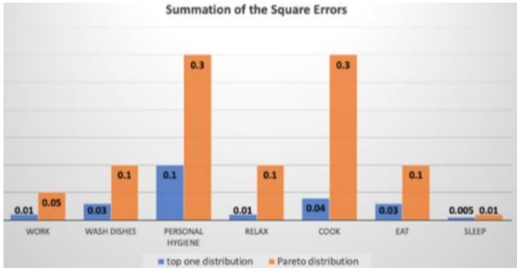


Figure 15. The SSE values of the top-fitting distribution and the fitted Pareto distribution.

Based on the above observations, we hypothesize that activity inter-arrival times may be used to predict subgroup classifications. We also notice that the Work activity (typically working at a desk or on a computer in an office area of the home) exhibits a large difference (0.36) in model shape parameters between the subgroups. To predict subgroup classifications, we currently use Work inter-arrival times from both subgroups and then utilize a random forest algorithm with 10-fold cross validation. The average accuracy of predictions is 0.814 and the standard deviation is 9.5. The precision of predicting subgroup classification is 0.784 and the recall is 0.740 .

Further, to validate that the random forest algorithm provides a statistically significantly-better prediction than that from random guesses, we perform a t-test analysis with the null hypothesis that the mean difference of the prediction results from the two algorithms are equal to zero. The p value of the t-test is 0.0001 with the t-statistic 10.43. Since a small p value (<= 0.05) leads to rejecting the null hypothesis, we concludes that the prediction results from the random forest classifier using model parameter attributes are statistically significantly-better than those from random guesses. These results indicate that the formal model does indeed reflect differences in behavior patterns or population subgroups and can help us understand behavioral impacts of traits such as chronic health conditions.

In addition, the shape parameters for activities Wash Dishes, Personal Hygiene, Cook, and Eat in the combined dataset are larger than parameters for either of the subgroups. That is, the shorter starting times between two successive activities occur more frequently in the combined dataset. This may be due to the number and diversity of residents in the combined dataset. Homes with multiple residents, young, or middle-age residents have a higher frequency of shorter inter-arrival times than that for single senior residents. We also noticed that the shape parameter of Work in the combined dataset is smaller than either of the subgroups (larger gaps between two successive Work activity occurrences in the combined dataset exist), possibly due to the fluctuation of residents' schedules, such as when the residents' are travelling, while seniors often have more stable schedules.

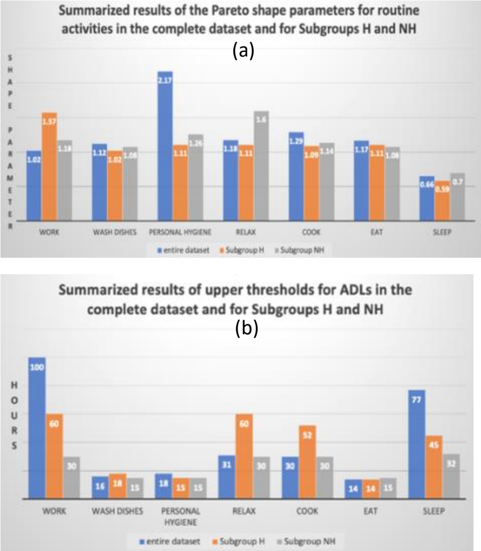


Figure 16. Summarized results of upper thresholds for ADLs in the complete dataset and for Subgroups H and NH. (a) Upper thresholds as a function of activity category. (b) Upper thresholds as a function of the subgroup.

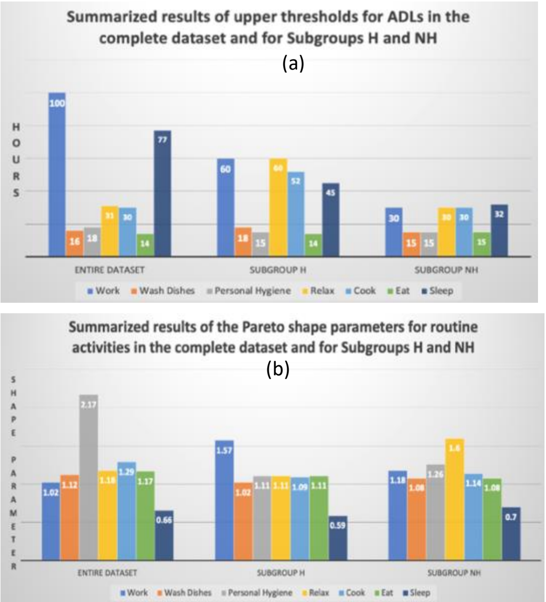


Figure 17. Summarized results of the Pareto shape parameters for routine activities in the complete dataset and for Subgroups H and NH. (a) Pareto shape parameters as a function of activity category. (b) Pareto shape parameters as a function of the subgroup.

In Fig. 17(b), we observe that in the combined dataset, the activities Relax and Eat have approximately the same value of the shape parameters (1.18 and 1.17, respectively). In Subgroup NH, activities Wash Dishes and Eat share the same value of the shape parameter (1.08 for both). In Subgroup H, four activities, Personal Hygiene, Relax, Cook, and Eat, have almost the same values of the shape parameters (1.11, 1.11, 1.09, and 1.11, respectively). Given the similar shape parameters, we propose to utilize bivariate or multivariate Pareto distributions (its cumulative density functions are shown in equations 1-3) to describe the combination of activities in each group. That is, the interdependencies of certain activities exist both at 99 smart homes and among subgroups. For example, in Subgroup NH, we can use a bivariate Pareto distribution to describe the relationship between activities Wash Dishes and Eat. In our study, all Pareto distributions are Pareto Type II.

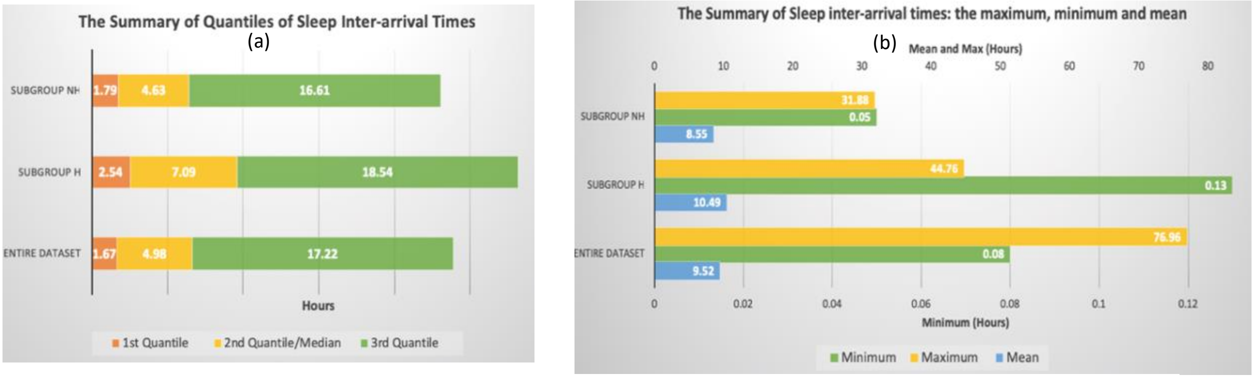


Figure 18. Summarized results of Sleep inter-arrival times. (a) includes the first quantile (1st Qu.), the median value (Median) / 2nd quantile (Median), the third quantile (3rd Qu.). (b) includes the minimum value of the dataset (Min.), the mean value (Mean), maximum value of the dataset (Max.).

The previous activities were tightly modeled as Pareto distributions. For Sleep inter-arrival times, the shape parameters are less than one for the entire sample of 99 smart homes and among population subgroups (see Fig. 17(a)). Statistically, this implies that the expected inter-arrival time approaches infinity. This result occurs because the mean value of the dataset is influenced by the largest value of the dataset. For a dataset of finite size, the sample has a finite value and so does the mean. But the more samples we have, the larger value of the mean. That is, the estimate of the mean is divergent when the size of the dataset goes to infinity. Since we cannot find a fitted distribution to adequately describe the pattern of the Sleep data below the selected thresholds, we utilize quantiles to understand the data spread (see Fig. 18). We notice that all the values in Subgroup H are greater than those in Subgroup NH. This is likely because residents with health ailments tend to experience more interrupted sleep. As this discussion highlighted, we do see differences in behaviour (and consequently in model parameters) for healthy and non-healthy subpopulations given the activities we examined. However, using the models to automate diagnoses is left for future work.

$$F_{entire}(x_{work}, x_{eat}) = (1 + \frac{x_{relax}+1.14}{1.14} + \frac{x_{eat}+1.61}{1.61})^{-1.17} \quad \text{Equation 1}$$

$$F_{subgroupNH}(x_{washDishes}, x_{eat}) = (1 + \frac{x_{washDishes}+1.31}{1.32} + \frac{x_{eat}+1.17}{1.17})^{-1.08} \quad \text{Equation 2}$$

$$F_{subgroupH}\left(x_{personalHygiene}, x_{relax}, x_{cook}, x_{eat}\right) = (1 + \frac{x_{personalHygiene}+1.66}{1.70} + \frac{x_{relax}+2.96}{2.97} + \frac{x_{cook}+1.66}{1.66} + \frac{x_{eat}+1.33}{1.33})^{-1.11} \quad \text{Equation 3}$$

Studying activity classes separately cannot provide a comprehensive view of a person's entire routine. To understand all activities comprising a routine, we select one home to investigate patterns of all activities. We combine both the predefined (and labelled) seven activities and the remaining clustered activities. The appropriate number of clusters is selected when no significant change of the sum of squared distances occurs in the elbow curve. That is, the optimal number of clusters is near the elbow part of the curve. Based on Fig. 19, we select $k = 10$. The resident is a single senior whose health status transitioned from healthy to having vision and mobility problems during the course of data collection. The time period of the experiment for this home is 65 months (from 2011-06-14 to 2016-11-10).

We perform the same process of data processing and model fitting as described in Sections 3 and 4. Fig. 20 shows that the Pareto distribution also fits the inter-arrival times for all routine activities. The threshold and shape parameter of the inter-arrival times of all activities are 5.8 hours and 1.17, respectively. The sum of square error for the fitted Pareto distribution is 1.8. Given the values of the shape (1.17) and scale (0.29) parameters, we confirm that 17% of the total number of inter-arrival points, which occurs in the tail, comprises 80% of the total inter-arrival hours, and 20% of the total number of inter-arrival points, comprises 74% of the total inter-arrival hours.

To further investigate the relationship between the fat tail of the Pareto distribution and the resident's health status, we first look at the 20% of inter-arrival times that occur in the tail and represent the large gap between two successive activities, and then manually examine the sensor data that corresponds to these gaps. Our investigation is summarized in Fig. 21. We notice that the successive activities, Sleep and Bed-Toilet Transition, occur in 40% of the cases lying in the tail. We hypothesize that the large gap and high frequency of these two activities are symptomatic of the resident's health problems. We validate the hypothesis by comparing the provided health information with the sensor-based night time walking duration (minutes) for the corresponding dates (see Fig. 22). Average night time walking duration is calculated based on the time that elapses between the end of a sleep activity and the beginning of the following bed toilet transition, given that the distance between bed and bathroom is constant and night-time bathroom trips typically involve direct routes.

We observe an increase in the average walking duration from August 2014 to November 2014. Self-reported mobile difficulty also increases during that time, provide a possible explanation for this change. We also notice that from December 2014 to March 2015 the sensor data reflects an decrease in the average walking duration, while the self-reported mobility difficulty

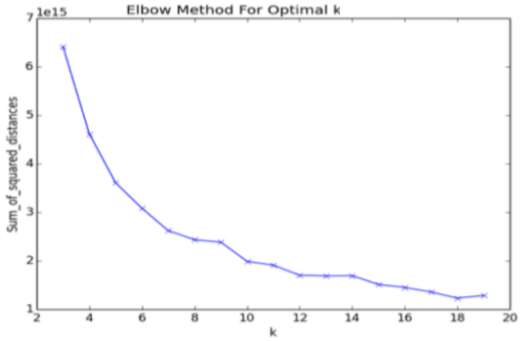


Figure 19. Elbow Curve.

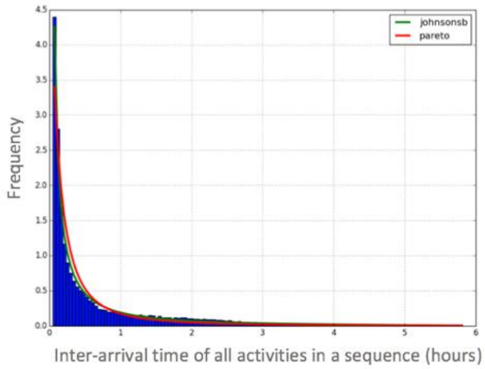


Figure 20. The Pareto distribution and the top-fitting distribution.

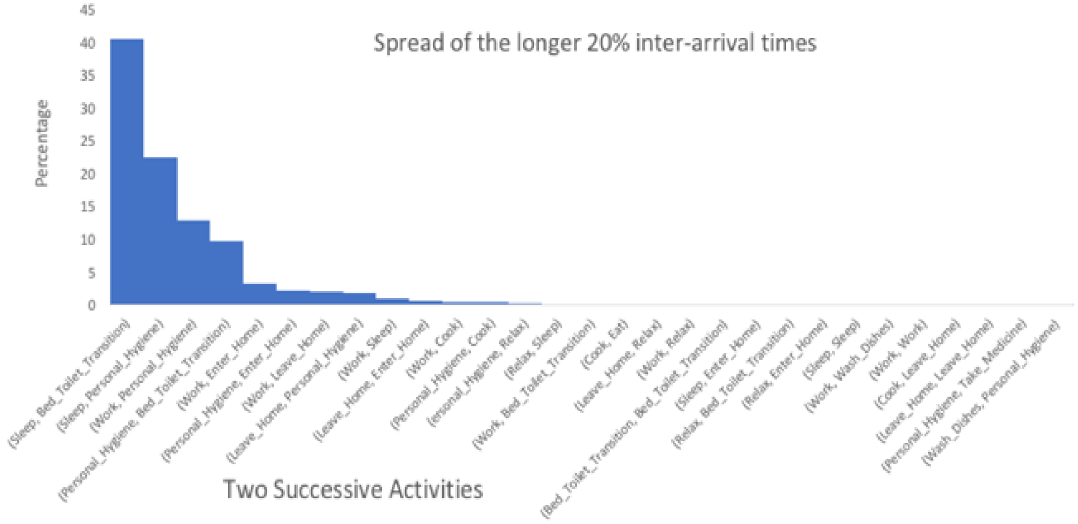


Figure 21. Spread of the two successive activities in 20% of the total inter-arrival times that occur in the tail of the distribution.

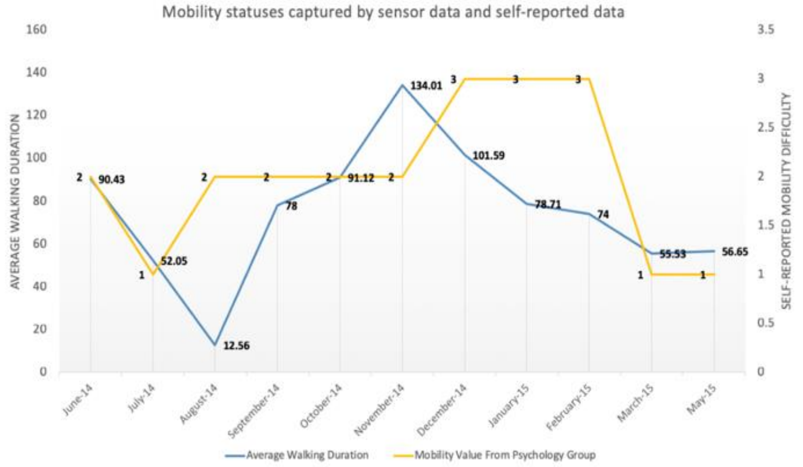


Figure 22. Compare the monthly average walking speed captured by sensor data with the self-reported mobility difficulty.

consistently drops from 3 to 1 (on a scale from 1= no difficulty to 5= tremendous difficult). The results provide evidence that the large time gaps and high frequency of Sleep and Bed-Toilet Transition activities in this particular home are related to the resident's health status.

6. DISCUSSION, LIMITATIONS, AND DIRECTIONS FOR FUTURE RESEARCH

In this paper, we propose formal methods for modeling human routines in everyday environments. We found that the Pareto distribution fits many activities, thus providing unique insights into behavior norms for the entire sampled population and behavior variations between population subgroups. Further, we discover that several activities in certain groups can be described by multivariate Pareto distributions. We also explore the pattern of all activities as a routine in one home and its relationship with the resident health ailments.

When applying our analysis to smart home data, we find that activities follow a non-Poisson process and the inter-arrival times of individual activities as well as all activities within a holistic routine fit a Pareto distribution. The findings may provide useful information to further investigate potential behavior changes that might be related to health problems. Limitations of this study include sensitivity of the models to noise in the sensor data, addressed in part by the outlier filtering process. In addition, the Pareto distribution fits the data well but does not fully describe all routine details. For example, the small hump in the histogram of the data (see Fig. 11 (c)) exists with frequency in a range of 0.01 to 0.02. A mixture model may be introduced to capture the hump for greater model detail. In Section 3, the selection of the threshold may impact the parameters of the Pareto distribution, especially when investigating the distribution's 20% tail. However, since we look at the highest percentage of two successive activities that lie in the tail, the impact of the similar threshold/shape parameters may be small. Further, the current two subgroups only consider single residents instead of multiple residents, though the experiments that evaluate the entire 99 smart homes do include multiple residents with diverse backgrounds. The problem of tracking, recognizing, and analyzing multi-resident behavior is an ongoing challenge, although Wang et al. discuss one possible strategy for multi-resident tracking in smart homes [70].

In addition, one can observe that our initial model oversimplifies the complexity of human behavior. For the purpose of this present study, we intentionally kept the model simple and focused on automatically-detected activity timings in smart environments. However, development of more sophisticated models combining other parameters including social interactions, circadian rhythms, night time relative walking speed, and movement patterns in and out of the home can further boost our ability to understand and reproduce the structure of human activities.

Additionally, further analysis of each activity and all activities in a routine will allow us to address specific questions that have been asked in the literature. For example, we could provide evidence to support or deny the hypothesis that human behavior in certain groups is random or Markovian [71]–[74]. We could also examine whether the 80/20 rule in which the largest-distance movements (20% of movement distances) occur with 80% of total inter-arrival hours applies in home environments. Finally, future work can quantify the predictability of behavior parameters for different population groups and types of sensed data.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1543656. The authors would like to thank Samaneh Aminikhanghahi for her help with change point analysis.

REFERENCES

[1] G. Last and M. Penrose, *Lectures on the Poisson process*, vol. 7. Cambridge University Press, 2017.

[2] F. Jovan, J. Wyatt, N. Hawes, and T. Krajník, "A Poisson-spectral model for modelling temporal patterns in human data observed by a robot," in *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, 2016, pp. 4013–4018.

[3] R. G. Gallager, "Poisson processes," *Stoch. Process. Theory Appl.*, pp. 74–108, 2013.

[4] A. Ihler, J. Hutchins, and P. Smyth, "Adaptive event detection with time-varying poisson processes," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 207–216.

[5] J. G. Oliveira and A.-L. Barabási, "Human dynamics: Darwin and Einstein correspondence patterns," *Nature*, vol. 437, no. 7063, p. 1251, 2005.

[6] A. Vázquez, J. G. Oliveira, Z. Dezsö, K. Il Goh, I. Kondor, and A. L. Barabási, "Modeling bursts and heavy tails in human dynamics," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 2006.

[7] R. F. Grais, J. H. Ellis, and G. E. Glass, "Assessing the impact of airline travel on the geographic spread of pandemic influenza," *Eur. J. Epidemiol.*, vol. 18, no. 11, pp. 1065–1072, 2003.

[8] A. Pieropan, C. H. Ek, and H. Kjellström, "Functional object descriptors for human activity modeling," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2013, pp. 1282–1289.

[9] O. Kwon, W.-S. Son, and W.-S. Jung, "The double power law in human collaboration behavior: The case of Wikipedia," *Phys. A Stat. Mech. its Appl.*, vol. 461, pp. 85–91, 2016.

[10] L. L. Constantine, "Human activity modeling: toward a pragmatic integration of activity theory and usage-centered design," in *Human-centered software engineering*, Springer, 2009, pp. 27–51.

[11] G. Gay and H. Hembrooke, *Activity-centered design: An ecological approach to designing smart tools and usable systems*. Mit Press, 2004.

[12] L. L. Constantine and L. A. D. Lockwood, *Software for use: a practical guide to the models and methods of usage-centered design*. Pearson Education, 1999.

[13] A. Bees, N. York, and A. Barabasi, "The origin of bursts and heavy tails in human dynamics," *Nature*, vol. 435, no. 7039, pp. 207-211., 2005.

[14] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," in *Proceedings of the 2007 SIAM international conference on data mining*, 2007, pp. 551–556.

[15] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani, "Characterizing and modeling the dynamics of online popularity," *Phys. Rev. Lett.*, vol. 105, no. 15, p. 158701, 2010.

[16] R. Kumar, M. Mahdian, and M. McGlohon, "Dynamics of conversations," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pp. 553–562.

[17] H. Li *et al.*, "Characterizing smartphone usage patterns from millions of android users," in *Proceedings of the 2015 Internet Measurement Conference*, 2015, pp. 459–472.

[18] Y. Gandica, J. Carvalho, F. S. Dos Aidos, R. Lambiotte, and T. Carletti, "Stationarity of the inter-event power-law distributions," *PLoS One*, vol. 12, no. 3, p. e0174509, 2017.

[19] I. Tsompanidis, A. H. Zahran, and C. J. Sreenan, "Mobile network traffic: A user behaviour model," in *2014 7th IFIP Wireless and Mobile Networking Conference (WMNC)*, 2014, pp. 1–8.

[20] L. Yu, P. Cui, C. Song, T. Zhang, and S. Yang, "A temporally heterogeneous survival framework with application to social behavior dynamics," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 1295–1304.

[21] T. M. Scholz, "The human role within organizational change: A complex system perspective," in *Change management and the human factor*, Springer, 2015, pp. 19–31.

[22] P. Andriani and B. McKelvey, "Perspective—From Gaussian to Paretian thinking: Causes and implications of power laws in organizations," *Organ. Sci.*, vol. 20, no. 6, pp. 1053–1071, 2009.

[23] Y. U. Saito, T. Watanabe, and M. Iwamura, "Do larger firms have more interfirm relationships?," *Phys. A Stat. Mech. its Appl.*, vol. 383, no. 1, pp. 158–163, 2007.

[24] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[25] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the levy-walk nature of human mobility," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 630–643, 2011.

[26] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science (80-.).*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[27] W.-Y. Zhu, W.-C. Peng, L.-J. Chen, K. Zheng, and X. Zhou, "Modeling user mobility for location promotion in location-based social networks," in *Proceedings of the 21th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1573–1582.

[28] S. Hong, *Human movement patterns, mobility models and their impacts on wireless applications*. North Carolina State University, 2010.

[29] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Sci. Rep.*, vol. 5, p. 9136, 2015.

[30] R. Gallotti, A. Bazzani, S. Rambaldi, and M. Barthelemy, "A stochastic model of randomly accelerated walkers for human mobility," *Nat. Commun.*, vol. 7, p. 12600, 2016.

[31] T. Kurashima, T. Althoff, and J. Leskovec, "Modeling interdependent and periodic real-world action sequences," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, 2018, pp. 803–812.

[32] R. Rawassizadeh, E. Momeni, C. Dobbins, J. Gharibshah, and M. Pazzani, "Scalable daily human behavioral pattern mining from multivariate temporal data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 11, pp. 3098–3112, 2016.

[33] N. Banovic, T. Buzali, F. Chevalier, J. Mankoff, and A. K. Dey, "Modeling and understanding human routine behavior," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 248–260.

[34] M. Pan *et al.*, "Dissecting the Learning Curve of Taxi Drivers: A Data-Driven Approach," in *Proceedings of the 2019 SIAM International Conference on Data Mining*, 2019, pp. 783–791.

[35] H. Ghayvat, J. Liu, S. C. Mukhopadhyay, and X. Gui, "Wellness sensor networks: A proposal and implementation for smart home for assisted living," *IEEE Sens. J.*, vol. 15, no. 12, pp. 7341–7348, 2015.

[36] G. Laput, Y. Zhang, and C. Harrison, "Synthetic sensors: Towards general-purpose sensing," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 3986–3999.

[37] B. Lin *et al.*, "Analyzing the relationship between human behavior and indoor air quality," *J. Sens. Actuator Networks*, vol. 6, no. 3, p. 13, 2017.

[38] A. P. Plageras, K. E. Psannis, C. Stergiou, H. Wang, and B. B. Gupta, "Efficient IoT-based sensor BIG Data collection–processing and analysis in smart buildings," *Futur. Gener. Comput. Syst.*, vol. 82, pp. 349–357, 2018.

[39] D. J. Cook, "Learning setting-generalized activity models for smart spaces," *IEEE Intell. Syst.*, vol. 2010, no. 99, p. 1, 2010.

[40] N. C. Krishnan and D. J. Cook, "Activity recognition on streaming sensor data," *Pervasive Mob. Comput.*, vol. 10, pp. 138–154, 2014.

[41] D. Cook and N. Krishnan, "Activity Learning from Sensor Data." Wiley, 2014.

[42] E. Kim, S. Helal, and D. Cook, "Human activity recognition and pattern discovery," *IEEE pervasive Comput.*, vol. 9, no. 1, pp. 48–53, 2009.

[43] A. Benmansour, A. Bouchachia, and M. Feham, "Multioccupant activity recognition in pervasive smart home environments," *ACM Comput. Surv.*, vol. 48, no. 3, p. 34, 2016.

[44] J. Wan, M. J. O'grady, and G. M. O'hare, "Dynamic sensor event segmentation for real-time activity recognition in a smart home context," *Pers. Ubiquitous Comput.*, vol. 19, no. 2, pp. 287–301, 2015.

[45] G. Fairchild, K. S. Hickmann, S. M. Mniszewski, S. Y. Del Valle, and J. M. Hyman, "Optimizing human activity patterns using global sensitivity analysis," *Comput. Math. Organ. Theory*, vol. 20, no. 4, pp. 394–416, 2014.

[46] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, 2019.

[47] A. Helal, D. J. Cook, and M. Schmalz, "Smart home-based health platform for behavioral monitoring and alteration of diabetes patients," *J. Diabetes Sci. Technol.*, vol. 3, no. 1, pp. 141–148, 2009.

[48] D. J. Cook, M. Schmitter-Edgecombe, L. Jönsson, and A. V Morant, "Technology-enabled assessment of functional health," *IEEE Rev. Biomed. Eng.*, vol. 12, pp. 319–332, 2018.

[49] C. Chen, D. J. Cook, and A. S. Crandall, "The user side of sustainability: Modeling behavior and energy usage in the home," *Pervasive Mob. Comput.*, vol. 9, no. 1, pp. 161–175, 2013.

[50] S. Aminikhanghahi, T. Wang, and D. J. Cook, "Real-time change point detection with application to smart home time series data," *IEEE Trans. Knowl. Data Eng.*, 2018.

[51] M. Yuan, "Human dynamics in space and time: A brief history and a view forward," *Trans. GIS*, vol. 22, no. 4, pp. 900–912, 2018.

[52] J. J. Davis and E. G. Conlon, "Identifying compensatory driving behavior among older adults using the situational avoidance questionnaire," *J. Safety Res.*, vol. 63, pp. 47–55, 2017.

[53] C. Li, W. K. Cheung, J. Liu, and J. K. Ng, "Automatic Extraction of Behavioral Patterns for Elderly Mobility and Daily Routine Analysis," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, p. 54, 2018.

[54] A. F. Costa, Y. Yamaguchi, A. J. M. Traina, and C. Faloutsos, "Modeling temporal activity to detect anomalous behavior in social media," *ACM Trans. Knowl. Discov. from Data*, vol. 11, no. 4, p. 49, 2017.

[55] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, "CASAS: A smart home in a box," *Computer (Long. Beach. Calif).*, vol. 46, no. 7, pp. 62–69, 2013.

[56] C. Truong, L. Oudre, and N. Vayatis, "A review of change point detection methods," *arXiv Prepr.*

arXiv1801.00718, 2018.

[57] D. Picard, "Testing and estimating change-points in time series," *Adv. Appl. Probab.*, vol. 17, no. 4, pp. 841–867, 1985.

[58] S. Aminikhanghahi and D. J. Cook, "Using change point detection to automate daily activity segmentation," in *Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on*, 2017, pp. 262–267.

[59] R. P. Adams and D. J. C. MacKay, "Bayesian Online Changepoint Detection," 2007.

[60] "Census Age Information." [Online]. Available: https://www.census.gov/data/tables/2010/demo/age-and-sex/2010-age-sex-composition.html.

[61] "Census Disability Characteristics." [Online]. Available: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_1YR_S1810&prodType=table%0D.

[62] "Census Educational Information." [Online]. Available: https://www.census.gov/data/tables/2010/demo/educational-attainment/cps-detailed-tables.html%0D.

[63] "Census Households Information." [Online]. Available: https://www.census.gov/data/tables/time-series/demo/families/households.html%0D.

[64] S. Coles, J. Bawa, L. Trenner, and P. Dorazio, *An introduction to statistical modeling of extreme values*, vol. 208. Springer, 2001.

[65] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.

[66] R. J. Hyndman, "The problem with Sturges' rule for constructing histograms," *Monash Univ.*, no. July, pp. 1–2, 1995.

[67] Ž. Ivezić, A. J. Connolly, J. T. VanderPlas, and A. Gray, *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*. Princeton University Press, 2014.

[68] I. Salgado-Ugarte, M. Shimizu, and T. Taniuchi, "Practical rules for bandwidth selection in univariate density estimation," *Stata Tech. Bull.*, vol. 5, no. 27, pp. 5–19, 1995.

[69] D. Freedman and P. Diaconis, "On the histogram as a density estimator:L2 theory," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 4, pp. 453–476, 1981.

[70] D. J. C. Tinghui Wang, "Towards Unsupervised Multi-Resident Tracking in Ambient Assisted Living: Methods and Performance Metrics," in *Assistive Technology for the Elderly, 1st Edition*, N. S. Subhas Mukhopadhyay, Ed. Academic Press.

[71] M. Rosenblatt, *Markov Processes, Structure and Asymptotic Behavior: Structure and Asymptotic Behavior*, vol. 184. Springer Science & Business Media, 2012.

[72] K. Doty, S. Roy, and T. R. Fischer, "Filtering and smoothing state estimation for flag Hidden Markov Models," in *American Control Conference (ACC), 2016*, 2016, pp. 7042–7047.

[73] M. Xue and S. Roy, "Spectral and graph-theoretic bounds on steady-state-probability estimation performance for an ergodic Markov chain," *J. Franklin Inst.*, vol. 348, no. 9, pp. 2448–2467, 2011.

[74] S. Roy, "Scaled consensus," *Automatica*, vol. 51, pp. 259–262, 2015.