2017 Summer Curricular Practical Training (CPT) Report

Jinlin He’18, *Grinnell College*

Mentor: Dr. Jacob Simmering, PhD, *University of Iowa Health Ventures*

Faculty Sponsor: Dr. Charlie Curtsinger, PhD, *Grinnell College*

Part A: Project Report (Time Series Clustering on Fitbit Data)

Background

MapTrek game is a walking game designed to encourage more walking exercise amongst users\*\*. The users participate in the game by wearing Fitbit which syncs data with a mobile application that tracks their step counts. By walking more in real life, participants move their virtual avatar in the game further in the designated route. Each route corresponds to a place of interest, for example, Rome, Italy. At any point of the route, the game provides the users with a real view at that location, and the view changes as users make progress in their walking.

Goal

The ultimate goal in this study is to understand individuals’ walking patterns from their Fitbit step counts. If certain individuals with a particular walking pattern drop out of the game sooner than others, it is helpful for the researchers to identify them and introduce interventions. Understanding walking patterns on an individual level can also inspire game designs that further motivate the users.

Data

This study utilizes data from 73 office workers, who wore Fitbit in a consecutive period of 80 days on average. Since all of them participated in the MapTrek game and previous studies suggested that the game changes users’ walking behaviors, so we filter out the days in which the users are already in the game. Therefore, we started out with the raw data that has 4863 days of minute-level step counts from these 73 individuals. Further data cleaning process reduced the sample to 2769 days from 72 individuals, in which we will explain in the following sections.

Throughout the study, we define each day as from 5am to 5am of the next morning. We also treat each day as the basic unit to work with, regardless of which individual it belongs to.

Method

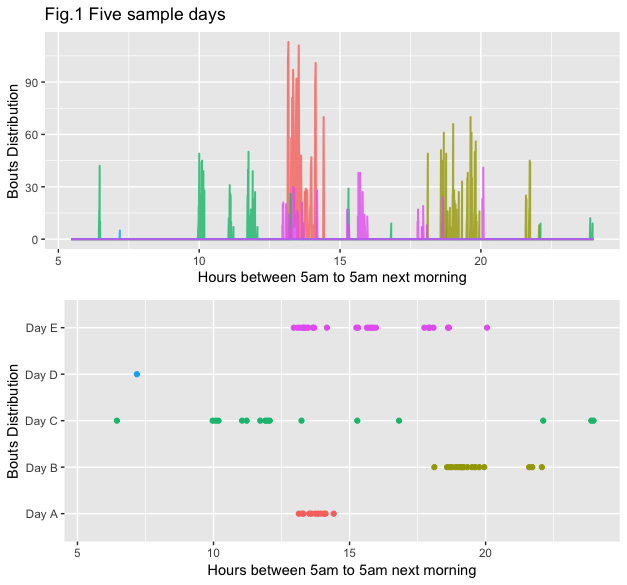
The analysis uses one of the unsupervised learning algorithms, k-means clustering, to recognize the major walking patterns from users on a day-to-day level.

K-means clustering algorithm starts with **n** random points (centroids) generated in the space, whereas **n** is specified by the user as the number of desired clusters. Then each data point is assigned to a cluster based on which centroid is the closest to it in space. After assigning every data point to one of the **n** centroids, the algorithm takes the average positions of all points assigned to the same centroid as the new centroid. Then again the algorithm distributes every data point to its closest centroid, and the process repeats until the centroid positions become stable.

It is particularly appropriate to apply k-means clustering to this problem because we are not asking any particular close-ended question such as how does one’s step counts per day relate to his or her compliance in the game. Instead, we hope that by treating step counts throughout the day as a time series, patterns will emerge on its own by the resultant clusters.

**Step 1: Identify usable data**

A general look at the original data reveals that in some days users wear their Fitbits only for a few hours. Fig.1 shows a sample of 5 person-days. Though Day A in red has a huge spike in step counts between 1pm ~ 2pm, but the user has no step counts on that day in all other time period, which suggests that it is likely that the user only wore Fitbit for that specific period but took it off for the rest of the day. Similarly, Day E (pink), Day D (blue), Day B (Yellow) also suggest that the users probably did not wear their Fitbits for the whole day. In comparison, Day C (green)’s step counts, though relatively in small amount, develops throughout the day, suggesting the user is constantly wearing the Fitbit. It is not useful then to use data on days like A, B, D and E for clustering as they do not fully reflect the actual walking habits.



To distinguish the usable days from the unusable ones, we sampled 150 person-days from all the *weekdays* of the original data, since we expect weekday walking behavior to be less erratic. Then we manually identified the days in which we are confident that users wore their Fitbits all day. For example, in Fig.1, only Day C appears as a convincing case that the user wore their Fitbits throughout the day.

To automate the process to the whole weekday dataset of 4863 cases, we trained a random forest model with cross validation on the 150 encoded sample with the indicator of whether one wore Fitbit for the whole day as the outcome.

From the minute-level step count data, we extracted these features to train the random forest:

* Number of bouts for the day
* Mean between-bouts time
* S.d. between-bouts time
* Duration between the first and the last bout of the day
* Whether the day has bouts in the early morning (before 9 am)
* Whether the day has bouts in the late evening (after 9 pm)

\*Note: the definition of a bout is a continuous period of walking without stop.

Applying this random forest model with an area under ROC curve around 0.95 (which essentially measures the accuracy of the classifier), we selected 2890 usable weekdays from the 3462 weekdays.

**Step 2: K-Means Clustering using step counts**

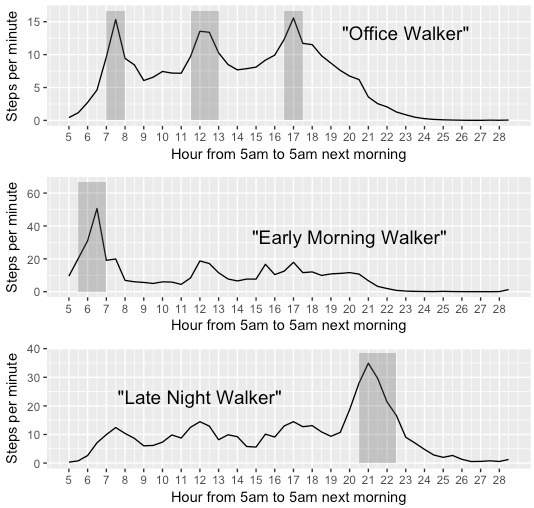
Now with the 2890 usable person-day data, we aggregated the minute-level step counts into every 30-minute step counts to keep down the dimensions for clustering without compromising too much of the details.

Then we can supply the factors of interest to the clustering algorithm, which include the time series itself, i.e. every 30-minute step counts throughout the day, as well as the extracted features:

* Mean step count
* S.d. of step count
* Number of steps taken between 5am ~ 7am
* Number of steps taken between 9pm ~ 2am
* Mean sedentary time
* S.d. of sedentary time

In one of the primary analysis, the clustering algorithm in fact singled out the person days of a particular user with a Subject ID of 13. It turned out that the user has normal step counts during the day, but has very high-intensity and consistent step counts throughout the night period, which seems like a device error which does not reflect the user’s actual walking patterns. Therefore, every data related to that user is removed from the clustering process.

After taking out the outlier subject data, we have 2857 days from 72 individuals. Using k-means algorithm to categorize these days into 3 clusters actually gave us very interesting insights:



Note: The hour 24 ~ 28 corresponds to 0 ~ 5 am of the next morning.

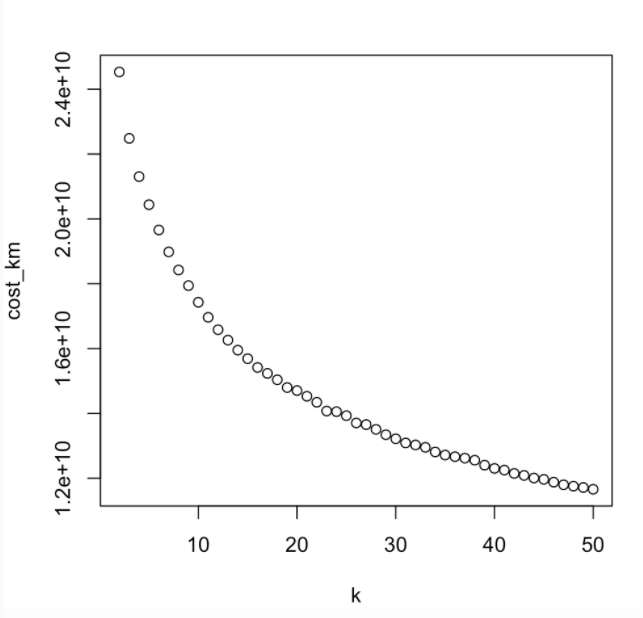
We see distinct characteristics from each of the three groups here:

* Group 1: the “Office walkers”. The 2541 person days in this group exhibit routine office walking behavior where the step counts spike during commuting periods (7am ~ 8am and 4pm ~ 5pm), and lunch break (noon).
* Group 2: the “Early morning walkers”. The 203 person days in this group still exhibit the step counts spikes in the commuting and lunch break period. Yet on top of these spikes, we notice a significant peak in the early morning around 6am ~ 7am. In this period, the walking speed reached 50 steps per minute, which is higher than a normal walking speed. Thus we can infer the users in these days very likely made a commitment to walk for exercise in the early morning.
* Group 3: the “Late night walkers”. Similarly, the 113 person days in this group show significant step counts increase between 9pm ~ 10pm, reaching 34 steps per minute. We can draw similar conclusion that the users in these days show habits of exercising in the late evening.

Though patterns already emerge from the three clusters, the next step is to determine what the optimal number of clusters is.

**Step 3: determine the optimal number of clusters**

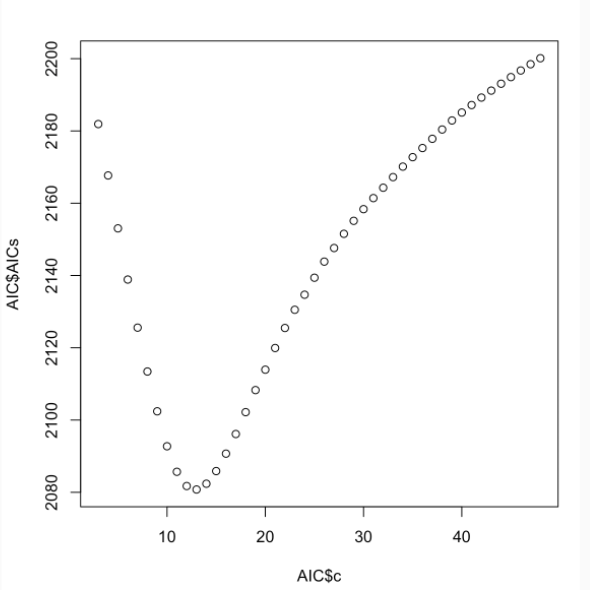
To determine the optimal number of clusters, we applied the clustering algorithms with **n** ranging between 2 and 50. We observe the plot of the **cost** of clustering against the number of clusters as Fig.2. We measure the cost of the clustering by looking at the total sum of squares within each cluster. The ideal situation is when each data point coincides with their respective centroids and the total sum of squares within each clusters will be zero.



In convention, the optimal number of clusters is determined by an “elbow” – a significant drop in cost in Fig.3.

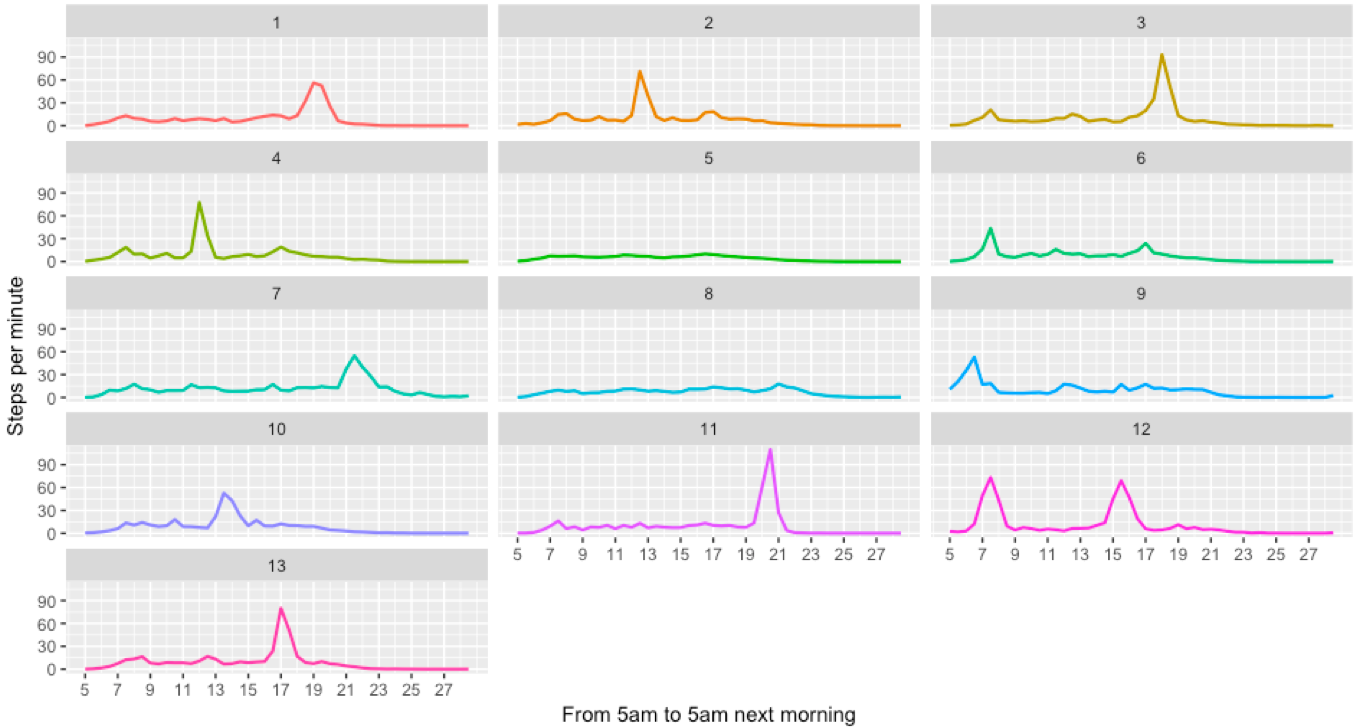
However, as you can see, the plot does not suggest an obvious elbow. Therefore a way to computationally determine the elbow is to fit two straight lines from both sides of the slope. Then the intersection of the two best-fit straight lines will be the elbow.

We can use AIC as a metric to assess the fitness of the two linear regression lines–the lower the AIC, the more the straight lines fit the points in Fig.3. The following plot shows AIC against the number of clusters. When the number of clusters is 13, AIC reaches its minimum. Thus 13 is the most likely to be the elbow–the optimal number of clusters.



Result

K-Means Clustering algorithm gives us the 13 clusters as follows:



The 13 clusters give us much more granular categories. The table below shows the number of days categorized into each cluster.

| **Cluster** | **Number of Days** | **Proportion** |
| --- | --- | --- |
| 1 | 106 | 0.0366782 |
| 2 | 135 | 0.0467128 |
| 3 | 86 | 0.0297578 |
| 4 | 177 | 0.0612457 |
| 5 | 1302 | 0.4505190 |
| 6 | 332 | 0.1148789 |
| 7 | 37 | 0.0128028 |
| 8 | 170 | 0.0588235 |
| 9 | 171 | 0.0591696 |
| 10 | 126 | 0.0435986 |
| 11 | 63 | 0.0217993 |
| 12 | 48 | 0.0166090 |
| 13 | 104 | 0.0359862 |

One of the most noticeable differences between clusters is at which time point the peak step counts take place, or if there is a peak in step counts at all. Though the peak happens in different hours, one can in fact see the shadow of the 3 clusters we observed in Step 2:

* There are days when users went for early morning jogs between 6am ~ 8am such as those in cluster 6 and 9.
* Noon jogs: cluster 2 and 4.
* Afternoon jogs: cluster 3, 10 and 13.
* Evening jogs: cluster 7 and 11.
* Twice jogs in a day: cluster 12.
* No jogs: cluster 5 and 8. Note: Cluster 5 has almost half of the sample, indicating that the users did not make a commitment for walking-related exercises half of the days.

In comparison with the 3-cluster model, the 13-cluster categorization gives us more details of users’ specific walking habits. By looking at the ratio of between-cluster sum of squares by the total sum of squares, we can have an idea of how much of the variance is explained by the model. In this case, the model with 13 clusters significantly improved the ratio from 0.163 in the 3-cluster model to 0.395.

However, depending on the future steps of the study, the 3-cluster model of the days may be sufficient for us to create further categorization of the users, while the 13-cluster model may not provide meaningfully more information. Nonetheless, the 13-cluster model, which shows the importance of the peak points in the clustering process, helped us make more sense of the 3 main categories in the previous model.

Future directions

* We are interested in using the categorization of each weekday to help categorize the walking patterns on a person level.
  + For example, a user that mostly has “early morning walker” days can be categorized as “early morning walker”.
  + We can also take into account the day in which the walking pattern occurred. Does the user switch between “early morning walker” and “office walker” as the week goes on? Is the user’s walking pattern consistent on a weekly basis?
* We would also hope to carry out similar analysis on the weekend days. Though the Step 1 of extracting out usable data will be messier, as we cannot conclude that one did not wear their Fitbit if the step counts start at 2pm on a Sunday, since people’s schedule on weekends are indeed more flexible. It might result in fewer data for clustering, but there might be promising results.

### Part B: Internship Reflection

### Learning Outcome

“The key learning outcomes I want to achieve through this internship is to learn both the application of statistical modeling techniques to real-life datasets and learn to draw on statistical theories to support such application.” – written in my SLIP.

Firstly, working on the summer projects is a process that inherently addresses the first goal on a regular basis. I applied a variety of statistical modeling and machine learning techniques, such as logistic regression, general linear models, k-means clustering, time-series analysis, random forest models to the real-life datasets.

To address the second goal, I spent a significant amount of time learning the theories behind the algorithms. To gain an in-depth understanding, I would go through the mathematical formula in the theories unless it gets far complicated. The Machine Learning course by Stanford University on Coursera was very helpful in giving a general knowledge map of various components in machine learning field. Applying the algorithms introduced in the course to the project I am working on, and troubleshooting along the way, gave me the confidence to learn new and related concepts and apply them in new settings. By intentionally designating time to strengthen my theoretical understandings, I notice that I am able to debug more modeling issues on my own because now I know if the outputs of a model look reasonable; I am also more mindful when choosing a modeling tool over the other given the type of data and the type of research questions.

### Technical Skills

Coding in R on a daily basis helped me become significantly more familiar with the commonly used packages in R such as “dplyr”, “readr”, “tidyverse”, “caret”, etc. Sometimes it takes a little bit more familarity of the knowledge to elevate one’s ability to understand the bigger picture. I feel the 9 weeks’ practice prepared me well to explore other new tools and not feel impossible.

### Professional Development

Going through the various stages of research under the guidance of the professionals like Jacob is an invaluable experience. By observing both his professional practice and analytical perspective, I have come to recognize some of the most distinct characteristics of capable data scientists:

* **Curiosity**. A capable data scientist always wonder and ask what happens if we solve the problem in another way and they make sure they will find an answer to that question.
* **Critical**. They hardly buy into even the most convincing narratives without taking into account what is not shown, and what could be the alternative explanation.
* **Creative**. Jacob is extremely creative in his way of approaching the problems. He would suggest building a model to automate data cleaning, applying tools across fields (such as this unconventional project applying clustering to time series), and always encouraging me to look up new algorithms that even he is unfamiliar with.

In fact, the last characteristic connects to my own personality very well. Always fascinated by new ideas, I envisioned one of my potential careers be in the field of digital design. However, now I realize that I can think creatively in this analytical field too. Thus data science has become a professional direction I will seriously consider.

### Future actions

With this direction in mind, I have identified a few areas I want to work on in my senior year:

* Analytical skills to unpack complex issues. I have observed how we can unravel one complex research question into multiple segments that are more workable. However, I still find the process of unpacking challenging.
* Independent learning skills and stamina. Data science is a fast-emerging field, where information updates on a day-to-day basis. To keep up with the field, I not only need to be able to process the new information, but also do so fast enough. To do so, reading up data science blog post or project on a regular basis can help me stay informed of the new methods and applications.
* Critical thinking skills in my own work. During my summer internship, I presented twice at the weekly meetings of professors and graduate students from Biostatistics, Econometrics in the School of Pharmacy. Though I had prepared well for both presentations, my professional audience ended up highlighting issues that I failed to pay attention to. It signifies that I have plenty of blind spots in my thinking that I should always keep questioning my results. Since the first presentation, I have been made aware of this and started working on it but it is indeed a long-term process to cultivate a critical thinking habit.