Statistical methods for detecting groups of patterns in daily step count activity profiles

Authors Elijah S.Meyer, Tan Tran, and Mark Greenwood

This is a copy of an article that originally appeared in Skyline, the Big Sky Undergraduate Journal in 2016.

Meyer, Elijah S.; Tran, Tan; and Greenwood, Mark (2016) "Statistical methods for detecting groups of patterns in daily step count activity profiles," *Skyline - The Big Sky Undergraduate Journal*: Vol. 4: Iss. 1, Article 6.

Skyline - The Big Sky Undergraduate Journal

Volume 4 | Issue 1 Article 6

2016

Statistical methods for detecting groups of patterns in daily step count activity profiles

Elijah S. Meyer
Montana State University-Bozeman, elijah.meyer@montana.edu

Tan Tran
Montana State University, tan.tran@montana.edu

Mark Greenwood
Montana State University, greenwood@montana.edu

Recommended Citation
Meyer, Elijah S.; Tran, Tan; and Greenwood, Mark (2016) "Statistical methods for detecting groups of patterns in daily step count activity profiles," *Skyline - The Big Sky Undergraduate Journal*: Vol. 4: Iss. 1, Article 6.

Section 1: Introduction

The growth of wearable technology is apparent. A total of 2.7 million wearable bands were shipped worldwide in the first quarter of 2014 (Sullivan, 2014). Of this number, Fitbit, a company focused on developing and manufacturing compact, wireless wearable technology devices accounted for half of the shipments alone. A Fitbit device can track steps, distance, calories burned, floors climbed, active minutes, and sleep patterns and provides regular users with daily totals or totals over 15-minute intervals throughout the day. The motivation of staying fit coupled with the convenience of collecting the data through a simple wrist band is fueling this industry. The utility of activity tracking for improved health outcomes is left for other researchers. Here we focus on whether the focus on daily summaries is making dissimilar days look similar, and whether we can group days in such a way that balances total activity as well as timing of that activity.

Daily total step counts have become a major focus of activity tracking with the Fitbit, rewarding the user with vibration upon completion of 10,000 steps in a day. Daily totals were the entirety of information available with analog pedometers and, while informative, may not tell the full story. The studies on daily step counting are still relatively limited. The physical activity of adolescents has been examined in Norway (Kolle et al., 2010), England (Riddoch et al., 2004), and Scotland (McCrorie et al., 2015) using step counts to compare the daily activity of children across seasons. Although these authors looked into hourly step counts as well as daily step counts, the length of study was limited to only several consecutive days in a season and the study focused on finding the relationships between the activities with seasons, sex, time of day, etc., and not daily activity pattern itself. Patel et al. (2015) suggest four principles to effectively promote health behavior using wearable devices. Along with affordability, wearing it, and having it record accurate information, they suggest that information be provided to the user in a way "that can be understood, motivates, that motivates action, and that sustains that motivation toward improved health". This project attempts to add to the tools for understanding activity patterns.

The coveted daily step count goal, often considered to be 10,000 steps, can be achieved in many different ways, from a short period of high activity to a consistent, low activity level over a longer period of time, as demonstrated in Figure 1. These two different scenarios could meet the same total goal, but potentially could have very different health impacts. While the total count is always close at hand to the users, fitness tracking device interfaces are

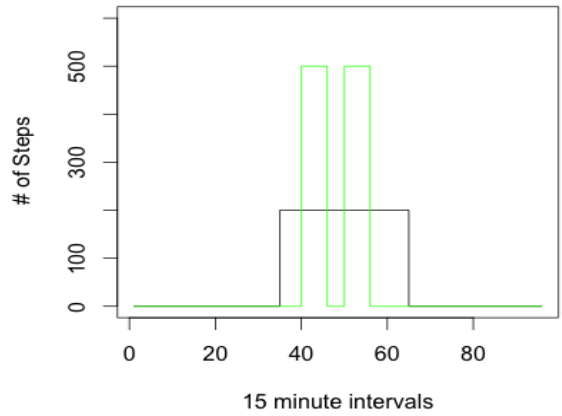

Figure 1. Plot of two days with similar total counts
and different activity patterns.

encouraging users to explore their information at finer time scales. But this results in an overwhelming amount of information for the typical user to process. The identification of types of daily patterns that respect both total activity and levels of activity by time of day may provide a useful summarization of the high-time resolution information. For example, discovering that yesterday was a "high activity early in the day" type of day might help the user reflect on their patterns of behaviors in contrast to a day that was classified into a group that had lower activity over a longer period of time or a day where high activity did not occur until the evening. These groups may also provide more useful information for exploring health outcomes than just focusing on the daily totals, but that is not explored here. We discuss methods for comparing daily activity profiles to respect total daily activity, and develop a new way of comparing daily step count profiles that allows for more flexible comparisons of days based on the initiation of the day than previously considered. The Appendix contains some recommendations and sample code for scraping and analyzing high temporal resolution Fitbit data using R (R Core Team, 2016), making these methods available to a highly motivated reader with their own Fitbit tracking device. As such, the methods developed here could help Fitbit users to understand their activity patterns.

Section 2: Data

The Fitbit has a very complex algorithm that it uses to define a step from the accelerometry data recorded by a wrist band. “The algorithm is designed to look for motion patterns most indicative of people walking. One condition for a motion pattern to be recognized as a step is the motion itself must be large enough. The algorithm implements this by setting a threshold. If a motion and its subsequent acceleration measurement data meet the threshold, the motion will be counted as a step. If that threshold is not met, the algorithm won’t count the motion as a step” (Fitbit, 2015). The step counts are provided to regular users as totals over fifteen minute intervals throughout the day. One author (Meyer), a 22 year old male, undergraduate student, wore a Fitbit on his wrist between July 28th and September 5th, 2015. Using functions developed in the R package FitbitScraper (Nissen, 2015), we accessed the 96 15-minute interval step counts for each of the days to create the dataset as well as extracting the daily total step counts. Other information is also provided by the activity tracker, such as stairs climbed, but we did not focus on this information; these methods could be applied to other similarly collected activity information. We also added Season and Day of week information to the data set based on the type of day, namely, summer weekday, summer weekend, school weekday and school weekend, to categorize each day specific to Meyer's responsibilities. To avoid misrepresentation by days that may have technological or human errors, such as, the battery ran out, or the device was taken off for an extensive portion of a day, we removed any days that were made up of a majority of 0 step counts. Specifically, there are 96 15-minute intervals in a day and days with nonzero counts for less than 10% (9 detected activity observations) of the total 96 intervals were removed from the data set. Based on this rule, six days were removed from the dataset and two more days were removed to provide complete observations between 6 a.m. and 6 p.m., leaving N=61 days for the subsequent analysis. The measured days had total counts between 284 and 31,641 steps. The 61 observations by time of day are displayed in Figure 2 with different colors illustrating each day. With so many days considered it is hard to extract much information from the display, a task that clustering the days can make easier.

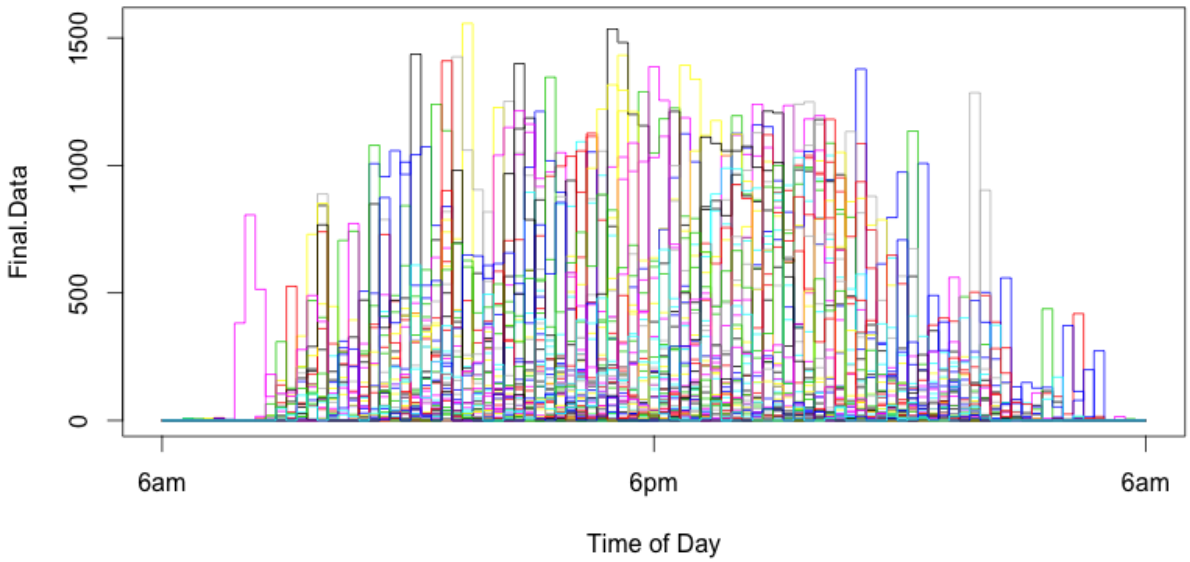

Figure 2. Plot of N=61 step count profiles starting at 6 a.m.

Section 3: Methods

In this section, we develop methods to provide groups in daily step count profiles, starting with a new way of comparing different days that provides a dissimilarity matrix among all pairs of profiles that can be used to perform hierarchical cluster analysis. Hierarchical clustering and related techniques are then discussed for identifying and summarizing groups of days. We used the statistical software R (R Core Team, 2016) for all work, with a selection of the R code used provided in the Appendix. The data set is available from the Montana State University Scholarworks repository (Meyer et al., 2016; http://scholarworks.montana.edu/xmlui/handle/1/9897).

Dissimilarities among pairs of days:

There are a wide variety of different measures for comparing multivariate responses, starting with the most commonly used Euclidean distance, which calculates the straight line distance between the two sets of points in multivariate space. With X_{it} defined as the count on day i at time t, the Euclidean distance between day i and day i' over times t=1,…,96 is

$$A_{ii'} = \sqrt{\sum_{t=1}^{96}(X_{it} - X_{i't})^2}. \quad (1)$$

It is commonly used because of its ease of understanding, easily understood properties, and ties to variance measures that drive most univariate statistical methods (like ANOVA and regression models). Disadvantages of Euclidean distance are that it is sensitive to outliers and the units of the distances with multivariate responses are hard to interpret since they involve the square root of the sum of the squared differences over all the variables.

With count data such as is present in the step counts over 15 minutes or an entire day, the Manhattan or City-Block distance is an attractive alternative distance metric to consider. It is based on the sum of the absolute value of the differences in the multivariate responses, calculated as

$$B_{ii'} = \sum_{t=1}^{96} | X_{it} - X_{i't}|. \quad (2)$$

As such, the units of differences across days are in the total absolute difference in steps. This provides two advantages: first, with being based on absolute values instead of the sum of squared differences, it is less impacted by outlying counts (like a large total count interval or day) and second, it has a more natural scale of interpretation in units of total difference in steps. The difference between this measure and just comparing the difference in the total steps for two days is that it accumulates differences based on each 15-minute interval. So two days that have the same total count could be reported as being very different on the City-Block metric if those counts occurred at different times of day. Only when the counts are similar at all times of day will the days have small differences.

This metric compares days at high time resolution which can provide more interesting comparisons than focusing just on comparisons of the overall daily totals. However, a person could have similar daily profiles that have slightly different starting and ending points and this metric would suggest those days were very different. To allow for this, we have developed a novel dissimilarity measure that compares pairs of days while allowing for potential shifts in the timing of different days. Days are allowed to shift forward and backward up to 2 hours, checking each step of 15 minutes for the best alignment of the two days as the minimum distance of those considered,

$$C_{ii'} = min_{j \in [-8,\ldots,8]}\{\sum_{t=1}^{96} |X_{i[t+j]} - X_{i't}|\}, \quad (3)$$

where j=1,…,8 is the 15-minute shift of the i^{th} step-count profile.

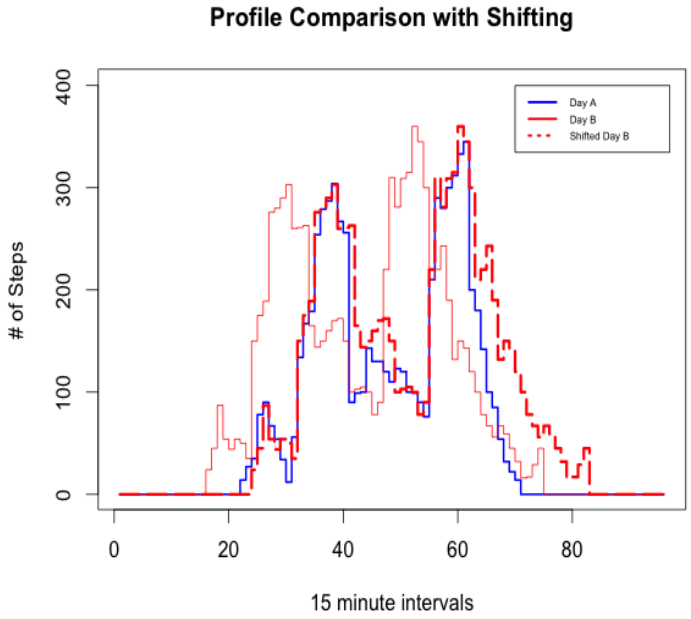

Figure 3. Plot of two generated days with optimal shift to closer align profiles.

This shifting creates missing observations in any pairwise comparison except when j=0 because there are fewer than 96 times to compare between the two days. To make the search across j values fair we need to adjust for the varying numbers of times being compared. The idea of comparing observations in the presence of some missing values was discussed as part of Gower's dissimilarity (Gower, 1971) and extended by Greenwood et al. (2011) to weigh the comparison of time profiles based on the amount of time or number of observations available for the pairwise comparison. In this situation, we divide the total absolute difference in steps by the number of time points being compared and select the j-value that produces the smallest result in terms of the total count difference per number of times compared. For j=0, this is just $C_{ii'}$ divided by 96. Specifically, the measure used for clustering below is

$$D_{ii'} = min_{j \in [-8,\ldots,8]} \left\{ \frac{\sum_{t=1}^{96} | X_{i[t+j]} - X_{i't} |}{96 - |j|} \right\}. \quad (4)$$

The search for an optimized alignment is illustrated with simulated daily profiles in Figure 3 which shows that an adjustment in the starting/ending time of the day can provide a much clearer alignment.

The pairwise alignment of daily profiles is a simple example of a technique called curve registration that is discussed in detail in Ramsay and Silverman (2002); it is a technique for aligning curves either based on characteristics of the curves or to match a template profile. Curve registration, when successful, can provide more clear-cut comparisons of profiles by aligning them optimally in time. Due to these changes, the $D_{ii'}$ measure is no longer a distance metric and can be called a "semi-metric" that has some, but not all, of the metric properties. Specifically, it could violate the identity of indiscernibles and triangle inequalities that are required for a dissimilarity to define a metric space (Ferraty and Vieux, 2006). With the pairwise registration of curves to optimally align each pair, it is possible (although this never happened) for two days that are different to be aligned and provide $D_{ii'}$ of 0. Additionally, with different shifts chosen as optimal for different pairwise comparisons, the triangle inequality can be violated. These violations do not impact our ability to cluster observations but can impact other uses of dissimilarity matrices so are important to note.

Because information is lost at the edges of the days in the sliding process, it is best to lose information where the least activity is present. For these observations, this occurred most frequently around 6 a.m. By using this as the definition of the end of one day and the beginning of the next day, as opposed to midnight, the edges of the days compared resided in this most inactive period for this subject. Figure 2 showed each day plotted with the adjustment already incorporated. Because of this shift and some days with almost no activity in this span of days, a total of 61 days had complete information from 6 a.m. to 6 p.m. and were used in the analysis.

It is important to note that the two hour shift is arbitrary, but provides a seemingly reasonable variation in shifts without creating too much change in the total number of intervals compared. Also, choosing to split days at 6 a.m. instead of other times was based on the activity profiles of this subject and would be different for others. Given these subjective choices, the semi-metric dissimilarity matrix is calculated across the N=61 days for use in cluster analysis.

Cluster analysis:

Cluster analysis is an exploratory technique designed to expose grouping in a data set (Everitt and Hothorn, 2011). The objects assigned to a specific cluster are more similar to each other and dissimilar from objects in other clusters. Ward's hierarchical cluster analysis (see Murtagh and Legendre, 2014, for a description) was chosen because it often provides reasonable cluster solutions. Ward's method minimizes the total within cluster variability and maximizes between cluster variability. Starting with clusters that are each individual day, Ward's method

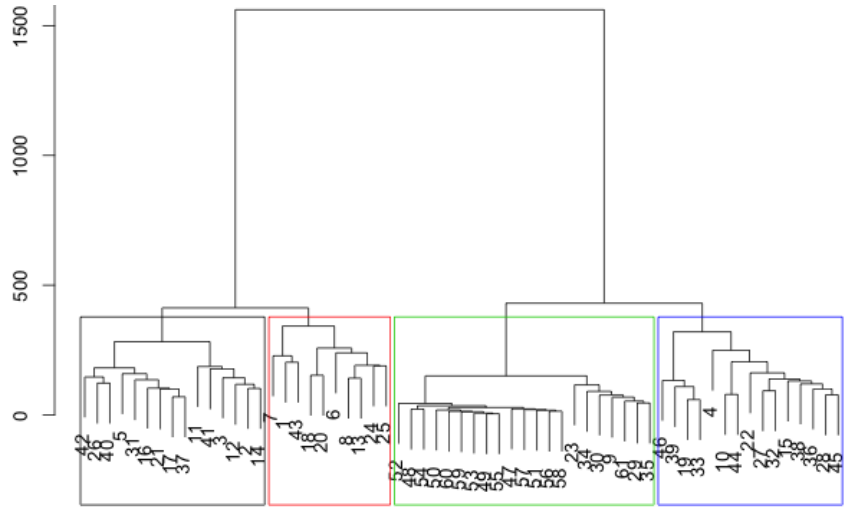

Figure 4. Dendrogram of sliding metric cluster solution for 61 days with a four cluster solution illustrated by boxes.

merges or agglomerates clusters together based on minimized pooled variability until all the objects are formed into one cluster. At each step, two clusters are joined together that result in the least increase in the pooled within-group variability. Once we obtain the cluster solution, we need to choose a specific number of clusters. We use the relative change in heights in the dendrogram that are displayed from the results of Ward's hierarchical cluster analysis. Relatively large changes in height in the dendrogram suggest large differences in the groups and small changes in height suggest clusters that should not be split apart.

Once the cluster solution is obtained based on the dendrogram results, the clusters are explored based on the results for each cluster in terms of the original daily step count profiles. In many cluster analysis situations, the mean responses by cluster are also used to represent each cluster. With the pairwise alignment of curves, the means would misrepresent the results for each cluster. Instead, we elected to use the cluster medoids (Kaufman and Rousseeuw, 1990) which are defined as the observation that has the shortest average distance to all the observations in each cluster. This central or representative observation is used to aid in characterizing the suite of observations that were grouped together. We also compared the resulting cluster groups on overall information about each day, specifically the daily total counts and the "type" of day – whether the day was a school day, weekend during the school semester, summer weekday, or summer weekend.

Section 4: Results

We chose to split the data set into the four clusters based on the relative heights in the dendrogram (Figure 4). If more clusters had been considered, some of the clusters presented would have had very small sample sizes and are also displayed being joined relatively closely to their neighbors. The first source of understanding a cluster solution is to explore the responses in the clusters, which are displayed with cluster medoids in Figure 5. Each cluster seems to demonstrate some pretty clear general themes in activity although the timing of some of the activity varies within each cluster. Cluster 1 (Figure 5a) contains days that tended to have relatively consistent, moderate activity levels through the entire day. Cluster 2 (Figure 5b) had higher intensity but only later in the day. Cluster 3 (Figure 5c) had higher intensity but earlier in the day with maybe lower peaks than in cluster 2 (late day peaks). Cluster 4 (Figure 5d) contained generally minimal activity with just occasional activity spikes.

Adding information on the types of the days in each cluster to these previous results, Table 1 contains the count of each specific type of day in each cluster. Cluster 1 was primarily associated with summer weekdays which contained moderate activity throughout the day. This sort of activity was associated with a summer job that kept Meyer on his feet continuously. Cluster 2 had a high proportion of summer and school year weekdays in it, suggesting that either could correspond to high activity later in the day. Cluster 3 also had a high proportion of weekdays, either from the school year or summer, providing a single peak in activity like in Cluster 2 but earlier in the day. Cluster 4, that contained generally limited activity, is almost exclusively associated with days during the school year. This suggests a general shift to lower activity associated

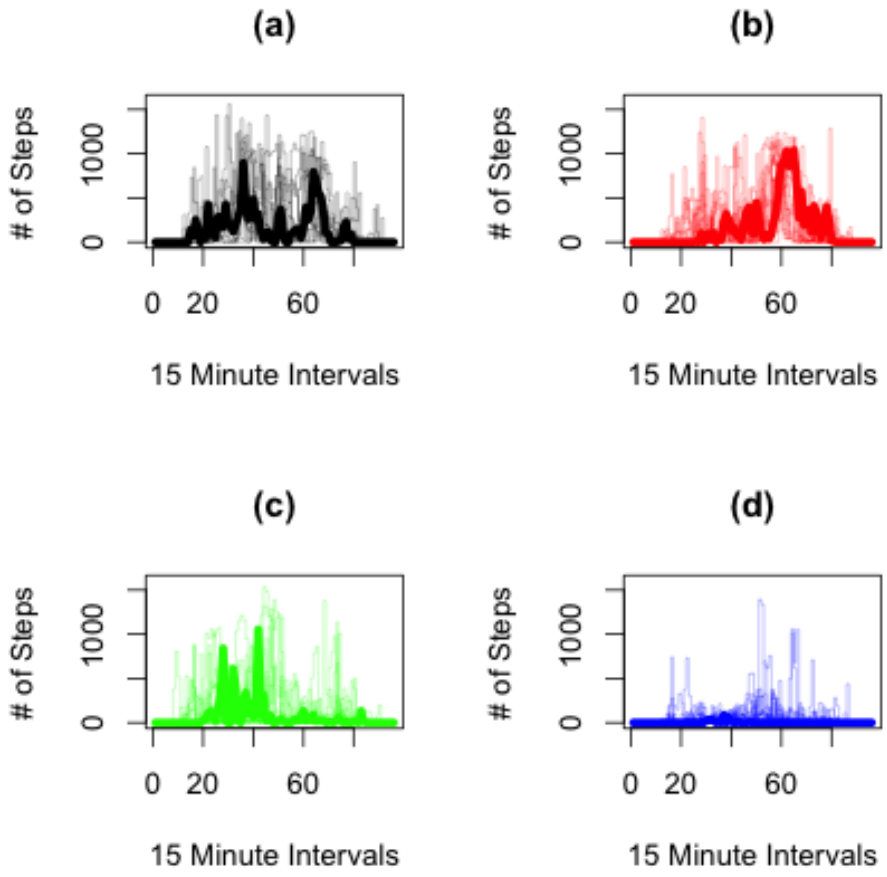

Figure 5. Plot of daily step count profiles by clusters from sliding dissimilarity with cluster medoids in bold in each panel.

with the change from summer activities to more sedentary days with mental activities pre-occupying his time. This information could be used to help alter fitness plans with the intent to be more active during the school year to maintain fitness levels obtained during the freer summer period.

It is also useful to compare the resulting daily totals for the different clusters. To support the utility of the novel dissimilarity measure that we proposed, we also compared the results with performing similar four-group hierarchical cluster analyses using our proposed semi-metric to regular Manhattan distance (no sliding) and with just clustering the daily totals. Beanplots (Kampstra, 2008) in Figure 6 display the total counts (tick marks), means of groups (wide line) and nonparametric density curves (shaded area) for the four clusters from each cluster analysis. These plots allow the reader to compare the individual observations, means, and shapes of the distributions both within and across cluster solutions.

The total counts for the four clusters described above are shown in Figure 6a. Using our proposed semi-metric, Cluster 1 has the highest step count days and the highest mean total count, the total counts overlap with results from other clusters. This suggests that Cluster 1 is more than just about accumulating a large count through activity within the day but also is related to consistency of activity. Cluster 2 also has fairly high average total step counts but these were generated in more limited times of activity later in the day and so often have lower total counts than Cluster 1. The same holds true for Cluster 3 but with even lower mean counts. This suggests that the activities that occurred in the morning may not have generated as many steps as typical afternoon activities. Cluster 4 has clearly lower total counts and a medoid that contained little activity.

Comparing the cluster solution using our semi-metric with the other two results displayed in Figure 6 provides further information about the information actually used in our method to form the clusters. Using Manhattan distance without sliding provides three clusters with average daily totals that are very similar (Figure 6b). The fourth cluster contains less activity but mixes in some fairly active days. Because of the rigid comparisons of times in this analysis, two days with the same profiles but a 15-minute timing offset could end up in different groups. The wide variety of total counts in each cluster suggest that timing is dominating these results, not the amount of activity. This is contrasted with just clustering using the univariate total counts (Figure 6c) which provides four clusters that are generated just based on total count information. These results suggest that our method is balancing timing and total activity to generate clusters of daily profiles that are based on both aspects of the information.

Table 1. Counts of type of days in each of the clusters using the sliding dissimilarity.

Cluster	School Day	School Weekend	Summer Week	Summer Weekend
1	1	1	7	1
2	5	1	6	3
3	8	2	4	1
4	13	6	1	1

Section 5: Discussion

The purpose of this study was to show that analysis of high resolution data for a given Fitbit user has the capability of detecting interesting groups of activity profiles that just using information from the standard daily summary counts fails to do. The results of this study showed four clusters of days seem to be present with very different characteristics of total and timing of activity in each group. The analysis methods developed seem to provide more information on when activities were performed within each day while still retaining aspects of overall activity. Focusing only on the daily totals means that all of the information about when the steps were taken is ignored. With this higher resolution analysis, one could set more specific individual goals related to timing of activities. Short, high intensity periods versus constant activity and having higher activity in the morning versus later in the day may yield different health benefits and these sorts of differences can be monitored using the proposed methods.

The limitations of this study are that the results apply directly to the user from which the data set was constructed and the time period that was studied. Any data set from a different Fitbit user could be analyzed, which might yield similar generic categories of days, but might produce very different types of groups of days and potentially a different number of suggested clusters. Additionally, a different user might be more active at the times used as the "edges" of days here and require a different choice. And a different amount of sliding might be found to be reasonable for a different subject. This does not consider all the impacts on clustering that can occur by using different clustering algorithms and, especially, when choosing to explore cluster solutions of other sizes. So, while the presented methods provide really interesting results here, the variety of required subjective aspects of the analysis make it hard to envision these methods being easily employed automatically for the regular Fitbit user. The highly interested reader is encouraged to use their own Fitbit and the provided R code to scrape and perform a similar analysis of their activity profiles.

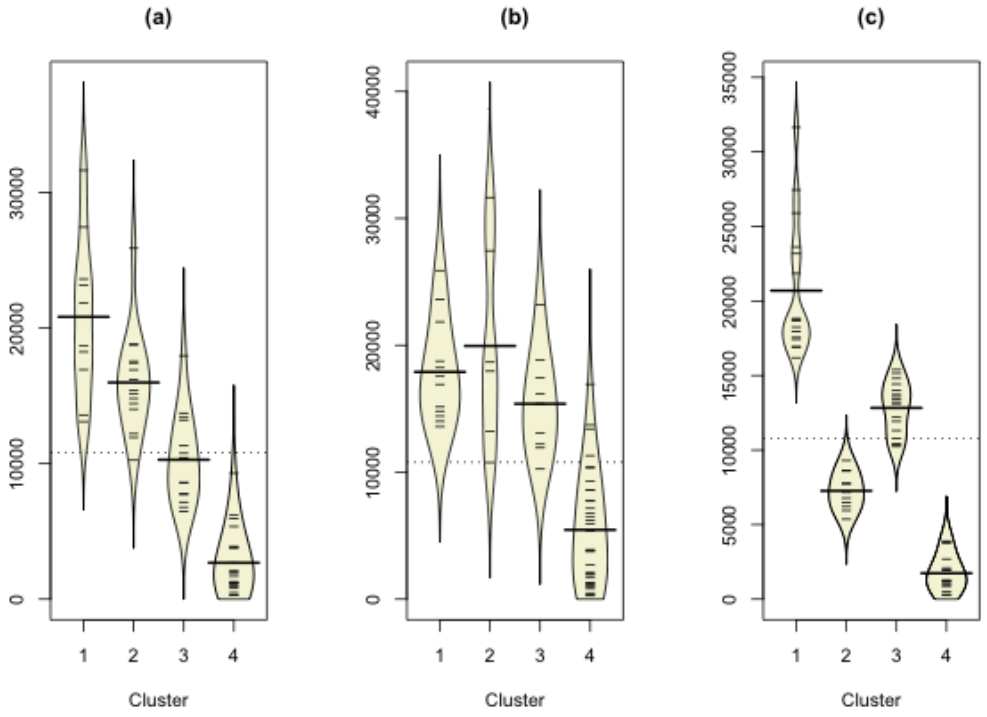

Figure 6. Beanplots of daily total step counts for Ward's four cluster solutions using (a) sliding dissimilarity, (b) Manhattan distance on profiles, and (c) Manhattan distance on univariate total counts.

Section 6: References

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R* (1st ed.). Springer.

Ferraty, F., & Vieu, P. (2006). *Nonparametric Functional Data Analysis - Theory and Practice*. *Springer*. New York: Springer.

Fitbit. (2015). Help article: How does my tracker count steps? Retrieved December 9, 2015, from http://help.fitbit.com/articles/en_US/Help_article/How-does-my-tracker-count-steps

Kampstra, P. (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software, Code Snippets 28*(1). 1-9.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis* (1st ed.). Wiley-Interscience.

Kolle, E., Steene-Johannessen, J., Andersen, L. B., & Anderssen, S. A. (2010). Objectively assessed physical activity and aerobic fitness in a population-based sample of Norwegian 9- and 15-year-olds. *Scandinavian Journal of Medicine & Science in Sports*, *20*(1), e41–7.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties, *Biometrics, 27*, 857–874.

Greenwood, M. C., Sojda, R. S., Sharp, J. L., Peck, R. G., & Rosenberry, D. O. (2011). Multi-scale clustering of functional data with application to hydraulic gradients in wetlands. *Journal of Data Science, 9,* 399-426.

Meyer, E., Greenwood, M., and T. Tran (2016). Daily Step Count Profile Data for 61 Days [dataset]. *Montana State University ScholarWorks*. http://doi.org/10.15788/M2BC7C

McCrorie, P. R. W., Duncan, E., Granat, M. H., & Stansfield, B. W. (2015). Seasonal variation in the distribution of daily stepping in 11-13 year old school children. *International Journal of Exercise Science*, *8*(4), 5.

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? Journal of Classification, 31(3), 274–295.

Nissen, C. (2016). fitbitScraper: Scrapes Data from Fitbit. Retrieved from https://cran.r- project.org/web/packages/fitbitScraper/index.html

Patel, M. S., Asch, D. A., & Volpp, K. G. (2015). Wearable Devices as Facilitators, Not Drivers, of Health Behavior Change. *JAMA*, *313*(5), 459–460.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B.W. (2002). *Applied Functional Data Analysis*. Springer, New York.

Riddoch, C. J., Bo Andersen, L., Wedderkopp, N., Harro, M., Klasson-Heggebø, L., Sardinha, L. B., Ekelund, U. (2004). Physical activity levels and patterns

of 9- and 15-yr-old European children. *Medicine and Science in Sports and Exercise*, *36*(1), 86–92.

Sullivan, M. (2014). Fitbit is crushing it in wearables; one out of two devices sold bears its name. Retrieved from http://venturebeat.com/2014/05/22/fitbit-is-crushing-it-in-wearables-one-out-of-two-devices-sold-bears-its-name/

Section 7: Appendix: R code

```
require(fitbitScraper)
#A user must specify their email address and password associated with his or her
Fitbit
#account in the form "login(email="#########@####.###", password=######)".
Name this something
#of your choosing using the "<-" entry. The code
#"get_intraday_data(name_of_login_and_password, what="steps" , date ="
#will allow the user to pull out 15 minute interval step count totals for the
provided dates.

#Sets up a sequence of user specific dates that he or she would like to pull high
resolution data from.
#Dates from #### through #### of data
#date <-  as.Date("####/##/##",format = "%Y/%m/%d") #Enter dates in the form
year, month, day
#date2 <-  as.Date("####/##/##",format = "%Y/%m/%d")
#s <- seq(date,to = date2, by='days')
#s <- as.character(s)

#DD <- NULL #Set up an empty Vector
#for (i in 1:#) #Want to pull out # number of dates
#{
#DD[[i]] <- cbind((get_intraday_data(name_of_login_and_password,
what="steps" , date = s[i]))[,c(1,2)])
#}
#name_your_data_frame# <- data.frame(DD)

#Function to create odd list to eliminate time from the data set

#Odd_Func <- function(p) {
#P is the number of days
#Odd_Seq <- seq(1,p*2,2)
#return(Odd_Seq)
```

```
#}

#Odd.mat <- as.matrix(Odd_Func(#)) Where # is the number of days he or she would like to pull data from
#eliminates time from his or her data set by pulling out every odd row
#data# <- #name_of_data_frame#[,-c(Odd.mat)]

#Meyer_2016 is the dataset used for analysis and is available using the following code from the Montana State University Scholarworks repository.

Meyer_2016 <-
read.csv("http://scholarworks.montana.edu/xmlui/bitstream/handle/1/9897/Meyer_2016.csv?sequence=1&isAllowed=y",header=T)
Final.Data <- t(Meyer_2016[,-c(97:100)])

require(zoo)

#Function to calculate distance with sliding between all days
Pair.dist4 <- function(t,s, lag= 8 ) {
  B8s <- lag(zoo(s),-lag:lag , na.pad=T)
  k <- data.frame(abs(t-B8s))
  keepdist <- colSums(!is.na(k))
  keepd2 <- (colSums(k , na.rm=T))
  dists<-(keepd2/keepdist)
  mind<-min(dists)
  obstemp <- keepdist[mind==dists]
  if (length(obstemp)>1) {
    obstemp = sort(obstemp , decreasing = T)[1]
  }
  return(c(Dist=mind,Obs=obstemp,Lag=names(obstemp)))
}

#Distance Matrix with lag version
#Specify your data set

dataset <- Final.Data
N<-ncol(dataset) #number of days
Ydist7 <- matrix(0,N,N)
for (j1 in 1:(N-1)){
  for (i1 in (j1+1):N){
```

```
    Ydist7[j1,i1]<- as.numeric(Pair.dist4(dataset[,j1],dataset[,i1] , lag=8)[[1]])
#distance
  }
}

Ydist7<-as.dist(t(Ydist7))

require(cluster)
require(fpc)

hc7 <- hclust(Ydist7 , "ward.D")

#k = 4 cluster solution
ward.clust <- cutree(hc7 , k = 4)

ID<-1:N

#Medoid Day Representative Function
#Ward.clust == # indicates the cluster memberships of your cluster analysis.
#The result will be the medoid, representative day for the specified #cluster

med <- function(members,Dist){
  if(length(members)==1){return(members)}
  else{
    if(length(members)==0){return(0)}
    dists<-apply(Dist[members,members],1,sum)
    medoid<-members[which(dists==min(dists))]
    return(medoid[1])
  }
}

med(members = ID[ward.clust==1],Dist = as.matrix(Ydist7))
med(members = ID[ward.clust==2],Dist = as.matrix(Ydist7))
med(members = ID[ward.clust==3],Dist = as.matrix(Ydist7))
med(members = ID[ward.clust==4],Dist = as.matrix(Ydist7))

#Example code to create plots of clusters with medoid highlighted #representative
day for each cluster
par(mfrow=c(2,2))
#Defining each medoid
M1 <- Final.Data[,13]
```

```
M2 <- Final.Data[,21]
M3 <- Final.Data[,28]
M4 <- Final.Data[,55]
matplot(x = 1:96 ,Final.Data[,ward.clust==1], ylim=c(0,1600), type = "s",col =
ward.clust[ward.clust==1], lty = 1 , main = "A" , xlab = "15 Minute Intervals" ,
ylab = "# of Steps")
matlines(M1, x = 1:96 , type = "l" , col = "black" , lwd = 5)
matplot(x = 1:96 ,Final.Data[,ward.clust==2], ylim=c(0,1600), type = "s",col =
ward.clust[ward.clust==2] , lty = 1 , main = "B", xlab = "15 Minute Intervals" ,
ylab = "# of Steps")
matlines(M2, x = 1:96 , type = "l" , col = "red" , lwd = 5)
matplot(x = 1:96 ,Final.Data[,ward.clust==3], ylim=c(0,1600),type = "s",col =
ward.clust[ward.clust==3] , lty = 1 , main = "C", xlab = "15 Minute Intervals" ,
ylab = "# of Steps")
matlines(M3, x = 1:96 , type = "l" , col = "green" , lwd = 5)
matplot(x = 1:96 ,Final.Data[,ward.clust==4], ylim=c(0,1600), type = "s",col =
ward.clust[ward.clust==4] , lty = 1 , main = "D", xlab = "15 Minute Intervals" ,
ylab = "# of Steps")
matlines(M4, x = 1:96 , type = "l" , col = "blue" , lwd = 5)
```