Analyzing large-scale people movement data

Akanksha Nichrelay
University of Virginia
an9sx@virginia.edu

Arjun Malhotra
University of Virginia
am2cj@virginia.edu

ABSTRACT

Big Data analysis is a hot and highly valuable skill to have in today's world. The data keeps flowing whether it is clickstream data from large websites or sensor data from a massive Internet of Things application, it is becoming increasingly important to be able to study these data and do data processing and data analytics to make sense of data. With the increasing use of mobile phones to track geo-locations of users, the people movement data can be used to find patterns in human activities and perform user modeling. In this project we analyze people movement data of Richmond, Virginia and predict hotspots in the routes with most human movement and turn it into a profitable model for marketing and advertising. Hot-spot is a time interval in particular region with most human activity during the day. By dividing the region into meaningful logical regions and analyzing the people movement in these clusters over time, we can predict which areas can serve as the most favourable and profitable locations for various marketing and advertising applications. Here in our project we focus on the smart bill board application and base our analysis on this but our idea can be extended to many other applications.

CCS CONCEPTS

• **Big data** → *Data Mining*; • **Unsupervised algorithm**; • **Machine Learning** → *Clustering*;

KEYWORDS

Data Mining, Cloud computing, Machine Learning Clustering, Unsupervised algorithm

ACM Reference Format:
Akanksha Nichrelay and Arjun Malhotra. 2018. Analyzing large-scale people movement data. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

The objective of our project is to analyze the large scale people movement data set, and predict the routes with most human movements. This may be useful for the Marketing domain, to put an apt advertisement on a smart billboard. This analysis will also be useful for other domains too. Like where in a route could one open a to-go coffee shop or a drive through restaurant. The study and understanding of the human motion is very important in various other numerous activities: it can be used in urban planning activities, to plan new roads based on the traffic, and to predict the spread of virus and diseases, just to name a few applications. This data can also be used by cab companies for analyzing where should they deploy shared rides versus personal rides. In recent years, the prevalence of positioning methods and technologies, such as the global positioning system, cellular radio tower geo-positioning, and WiFi positioning systems, has driven efforts to collect human mobility data With a surge in the mobile devices, a lot of movements can be tracked these days. Which gives account to various human mobility patterns that reflect many aspects of life [3]. As an example, daily commute patterns. This indirectly promotes the development of location-based services and applications. Our intent is to provide one such service. In this project, we survey and assess different approaches and models that analyze and learn human mobility patterns using mainly machine learning methods.

With the advances in machine learning, we can use the people movement data in a more scientific way to determine prevalent patterns. Application of machine learning allows use to logically break down the problem into meaningful pieces and analyze the regions of Richmond city in a given time frame. We plan to use K-means algorithm and time-series regression in order to analyze our data and find people movement hot-spots in a given region in a particular time window. A question here arises "Why do we use K-means or clustering ?" To put it in simple words, we would want the billboards to be a certain distance from each other, this distance must not be very small, as then the billboards would be right next to each other which is not good for a business model. At the same time the billboards must not be very far from each other too, as that would mean that billboards would have little impact. Hence we employ clustering/K-means, with the right number of clusters we try to maximize inter-cluster distance at the same time minimize intra-cluster distance. We analyze the clusters keeping different regions in mind. Regions with more human population density would have smaller regions as clusters. Since we had Time series and Geo-location data we divided whole data into meaningful regions/clusters. We then analyze the population density in each cluster region over time by dividing the time into 10 minute time-windows. We further apply some simple data modelling to predict the human density at a given time t by using the previously known data for previous time-stamps.

2 RELATED WORK

There have been instances where a lot of human movement data has been studied before but is has been done to find applications that can be categorized into three classes: user modeling, place modeling, and trajectory modeling, each class with its own characteristics [4]. While we focus on making our model more profitable, such that it could be used across various other domains. There are many challenges in trajectory modeling that can occur and thus special paradigms related to trajectory modeling have been creating in the

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

field of data mining in order to properly handle the trajectory data [5].

Another work that has been studied before is "Large scale movement analysis from WiFi based location data" which does lay a little emphasis on the relationship between a human and a place but is in it's entirety based on Wifi Data [3]. Which means not much mobility with the humans.

We are focusing on building our model more suited for advertisement and smart billboards as we think this is a hot topic and can profit with human movement patterns when combined with other user related and location related details [1].

3 PROPOSED METHOD

Modelling human motion has attracted the attention of the scientific community during the past few years. We plan to use the real Richmond city GPS data for our analysis. We did not receive the whole data set due to NDA and legality issues. We were provided a sample of the data set of about 8 million records. Our analysis and proposed methodology so far is thus based on this sample data and may change as per the original data set. Our sample data set mainly consists of User Identifier, Latitude and Longitude coordinates and time stamp of capturing of the location. There are some other geo-location and user device based features which might help in our analysis but for the scope of this project, we focus on the above mentioned features only. Upon analyzing our data, we further found that the data given to us is not continuous and is spread across the month of January for year 2019. We merged the two files from 3rd - 4th January and 4th- 5th January as our training data. Then to get meaningful test data, we chose 11th-12th January data as it falls on same day of the week as our training data - both 11th and 12th January and 4th and 5th January are Friday and Saturday respectively for the year 2019. We analyze the time-series pattern for the city of Richmond using these training data and do simple data modelling using training data to make predictions of times-series pattern on test data.

The data set is not labelled and thus we plan to use unsupervised machine learning models to analyze the data. The people movement data set is unique in a way that it consists of geo-location features as well as the time-series features which makes it challenging to solve using regular regression techniques that we can use on any other sequential data. This problem thus can be thought of as two sub problems - divide the given area into optimal number of logical regions and then then break up time into a appropriate sized window in which we can analyze our regions for people movement patterns and hot-spots.

We want to break the whole city of Richmond into logical regions so that we can analyze the human movement pattern in that region over the time-period. One can ask that there is no need for K-means clustering and we can just divide the regions using latitude and longitude. But our idea is that instead of drawing random boundaries, we want to form meaningful comparable clusters so that we can quantitatively say which cluster is serving as a hot spot at a given time. Further, we want to keep these clusters at a certain inter-cluster distance so that they are not too close from each other as it would not make sense to put billboards too close to each other and yet not too far as to not make the best use of space and turn down profit. Thus, we will use the latitude and longitude features to find out the optimal number of clusters using K-means algorithm. In order to define clusters, we decided to keep the inter cluster radius as 10 miles since we decided that could be the optimal distance to place smart billboards or advertisements. We then choose the best k number of clusters by analyzing how many clusters are at a greater distance of 10 miles from each other and the minimum inter-cluster distance between any two clusters. We would want higher number of clusters to be within 10 miles distance while keeping the minimum inter-cluster distance large.

We then analyze the population density in each cluster region over time by dividing the time into 10 minute time-windows. We further apply some simple data modelling to predict the human density at a given time t by using the previously known data for previous time-stamps. The time-series data is a special case of sequential data since we need to predict what will the outcome at a time t+1 given the data upto time t. This makes it thus a unique kind of regression problem where it needs to consider the pattern or knowledge of previous data. There are many methods to simplify analysis of time-series data such as Fourier transformations which we plan to use in order find the time variations of patterns such as peak time etc. We construct three types of features Fourier transformations, ratio features and previous known values to predict the human movement patterns in each cluster. We apply simple data modelling like Simple Moving Averages, Weighted Moving Averages and Exponentially Weighted Moving Averages to determine which of these features are more useful in making predictions. We evaluate these models using Mean Square Error and Mean Absolute Percentage Error to determine which model gives the best result.

So to summarize, we will use K-means algorithm to predict the routes with most human movements hot-spots at certain time of a day. We plan to analyze the hot-spots of people movement in a time bin or window size of five minutes as we think this should give us a good estimate of dwell time of people in a certain given area. In essence we are trying to predict the number of people in a particular cluster in a given time window. This will be determined once we have our geo-location clusters and time-series features in place. We currently plan to first work with the provided data of January 2019 and build a simple model to get us started. We will make predictions on a test data to find population density patterns in every region.

4 EXPERIMENT

As we mentioned in the previous section, the data set is not labelled and thus we plan to use unsupervised machine learning models to analyze the data. Mainly, we will use K-means algorithm to divide Richmond area into logical regions and then analyze and predict the density with most human movements hot-spots at certain time of a day. We will first clean our data as there are many unnecessary columns as the GPS data can be very noisy [2].

Upon analyzing our data, we further found that the data given to us is not continuous and is spread across the month of January for year 2019. Our data is divided into five data files which consists of geo-locations readings on 3rd- 4th January, 4th- 5th January, 11th-12th January, 22nd- 23rd January and 28th- 30th January. As this data is sparse in terms of date-time, we focus on the most continuous

data we can extract. We merged the two files from 3rd - 4th January and 4th- 5th January as our training data. Then to get meaningful test data, we chose 11th-12th January data as it falls on same day of the week as our training data - both 11th and 12th January and 4th and 5th January are Friday and Saturday respectively for the year 2019. We did that in order to make meaningful predictions on continuity of time-series data. We analyze the time-series pattern for the city of Richmond using these training data and do simple data modelling using training data to make predictions on people density at a given time t using previous known data of previous time-stamps.

We begin with cleaning our data and dropping the unnecessary columns that had no data. We also check for possible outliers in our data set. We found the actual range of latitude and longitude of Richmond city in order to weed out any data points that may lie outside of this range. We begin our analysis by taking initial 4GB of data which has over five million records. As stated above, we proceed with a training data of over two million records and a test data of over one million records. We then try to find the area of Richmond that is covered by this sample data. To break our the whole city of Richmond into logical regions we will use the latitude and longitude features to find out the optimal number of clusters using K-means algorithm. For this we cluster all the latitude and longitude in training data using K-means algorithm and try to vary the number of data points a cluster can contain. We are able to find that the sample data only covers a certain county / section of the Richmond city as shown in fig 1.

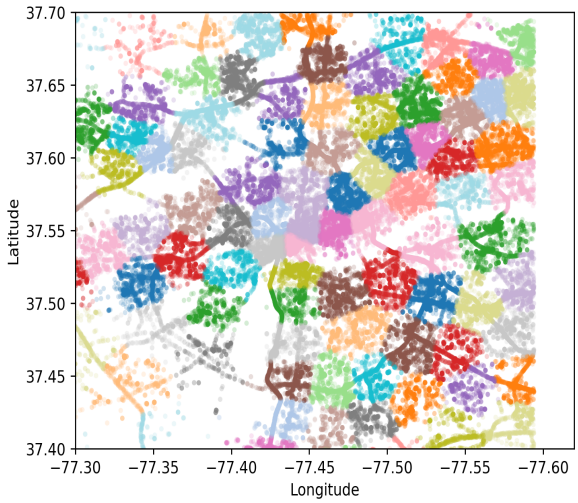

Figure 1: The data set covers the entire area of Richmond city. Clusters of k =90 to show the area covered by the data set.

Here, in order to define clusters, we decided to keep the inter cluster radius as 10 miles since we decided that could be the optimal distance to place smart billboards or advertisements. We try to find the optimal number of clusters by varying the cluster sizes in a range of 10-100. The above plot shows the clustering when the cluster size is 90. In this case, we calculate the average number of clusters within the vicinity (i.e. inter cluster-distance < 10) which is 51.0 and the average number of clusters outside the vicinity (i.e. intercluster-distance > 10) which is 39.0. We found that k=50 gives us the best ratio of the number of clusters closer and farther than a inter cluster distance of 10 miles. This analysis also gave us an idea of how the data is spread, this particular clustering is a good estimate. Figure 2 shows the best number of cluster with k=50 and the graph figure 3 shows the plot of cluster centers across Richmond city. We then assign the cluster center label to each data point in our data frame for further reference. We can see the most dense cluster are present at the core of Richmond city where the big businesses are present while the outer suburbs are sparsely populated.

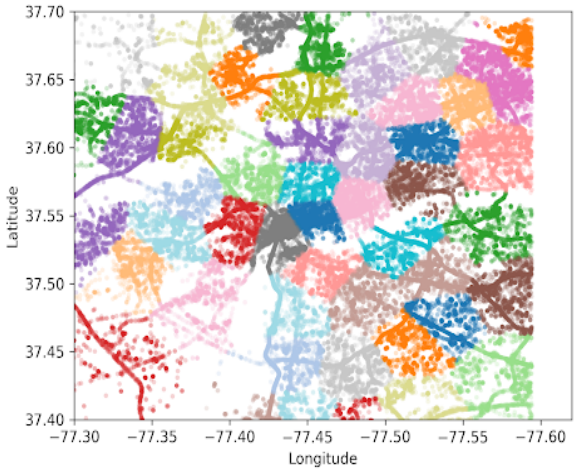

Figure 2: Best clustering found for k=50 using K-means on training data set.

Next, we divided the time into 10 minute window by converting it into unix timestamps. We took care of the Time Zone differences and converted the timestamps to EST from UTC as unix timestamps convert data into UTC. For our data we found that the training data can be divided into 419 time bins. Then we assign and store the time bins to our data in the data frame. Next we calculate the the number of people in a cluster in each time bin by putting the advertiser id, cluster id and time-bin id and grouping by cluster id and time-bin id. This gives us the count of people in each time bin for each cluster. However, upon trying to plot the data, we realize that we have missing data for certain time-bins for each cluster. Since we are predicting the population density for a time bin $t + 1$ based on the previous time bins t, there were time bins which had zero population density. Which meant that we couldn't have predicted the next time -bins population based on these 0 population density time bins. To get around this we employed smoothing. We performed smoothing by checking number of time bins with zero population density in each cluster, there were quite a lot of them. We broke the smoothing scenario into 3 cases and then filled the missing values with the average values.

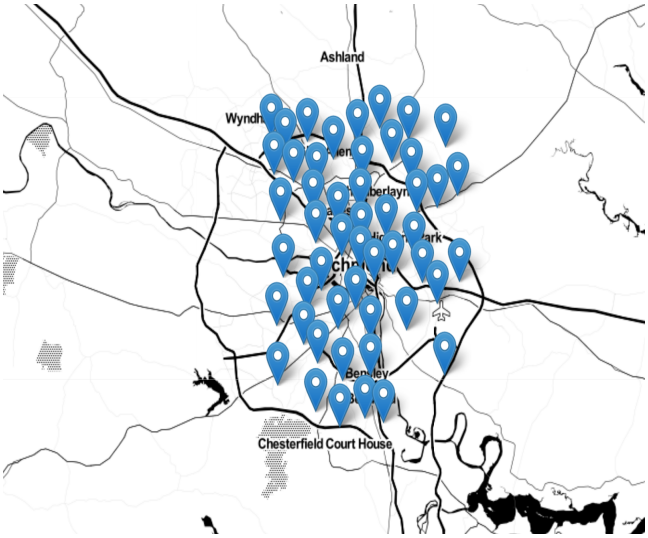

Figure 3: Best clustering found for k=50 using K-means on training data set.

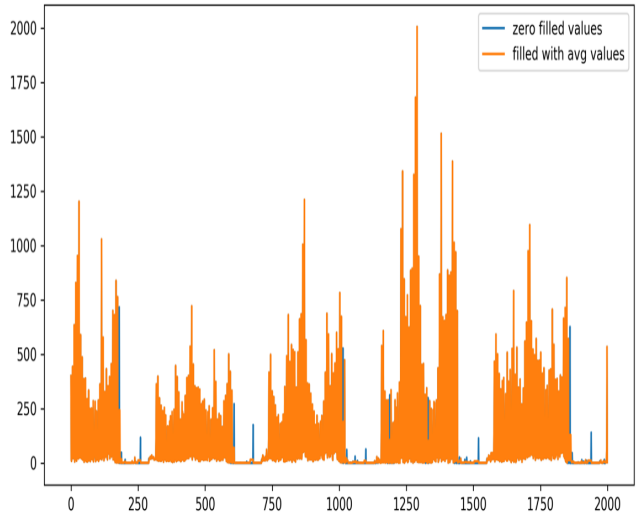

Figure 4: The result of smoothing the time-series data by filing the gaps gives us continous wave form of the data. The result above is shown for a particular time-interval of a cluster

(1) Case 1: Values missing at the start
Ex1: _ _ _ x =>ceil(x/4), ceil(x/4), ceil(x/4), ceil(x/4)
Ex2: _ _ x => ceil(x/3), ceil(x/3), ceil(x/3)

(2) Case 2:Values missing in middle
Ex1: x _ _ y => ceil((x+y)/4), ceil((x+y)/4), ceil((x+y)/4), ceil((x+y)/4)
Ex2: x _ _ _ y => ceil((x+y)/5), ceil((x+y)/5), ceil((x+y)/5), ceil((x+y)/5), ceil((x+y)/5)

(3) Case 3: Values missing at the end
Ex1: x _ _ _ => ceil(x/4), ceil(x/4), ceil(x/4), ceil(x/4)
Ex2: x _ => ceil(x/2), ceil(x/2)

As mentioned above smoothing was basically the average of the values that were present. Which explains the denominators of the above equations. The smoothed result gave us continuous data as shown in figure 4.

As a next step, we worked with the time-series data and processed it so that we can analyze the hots-pots at a particular time. We looked at how to work with times-series data and how to apply Fourier transformations in order to better identify time peaks and patterns of our data set. We plotted the graphs on test and train data to identify daily patterns. Figure 5 to 8 shows some examples of people movement in clusters of training data.

Figure 9 and 10 shows the corresponding cluster in the test data. As we can visualize, the pattern of movement is similar for similar days of the week.

Next we extract time related features that we think can help in predicting the human population density hot-spots using the known data so far. What we mean by that is that we can not only use the training data of time bins observed so far but also we can use the test data we have observed up to time t-1 if we are trying to predict human movement at a particular time bin t. We use this idea to extract three features: Fourier Transformations, Ratio features and Previous known values feature. Firstly, we tried

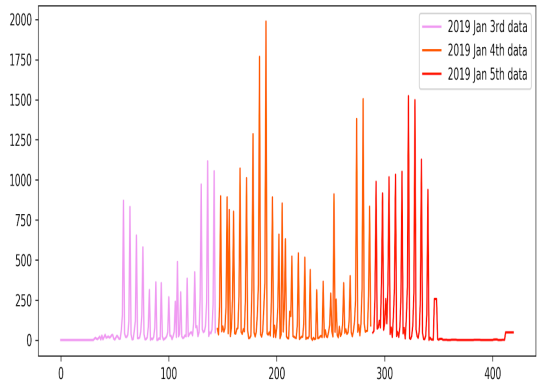

Figure 5: train cluster 1

to apply Fourier Transformations on the time series graphs that we plotted for each cluster. We plan to use the Amplitude and Frequency to help analyze the daily patterns observed. Figure 11 shows the Fourier transformation features observed. We can see that there is a peak for daily pattern and the other peak corresponds to the busy times of the day, one during morning and the other again we see a peak time during the evening when people are heading home from work. We believe that the noise in the graph is present because we lack continuous data. Despite that the peaks are quite dominant and we can use these peaks of daily patterns as features to predict hot-spots.

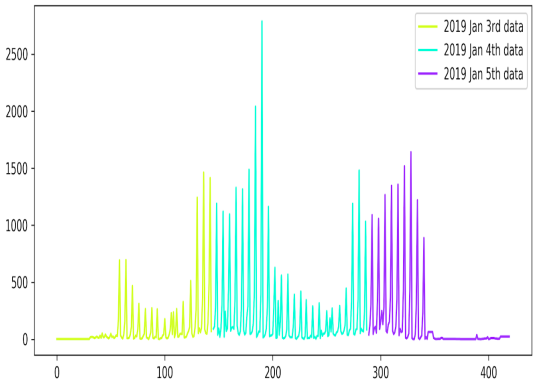

Figure 6: train cluster 2

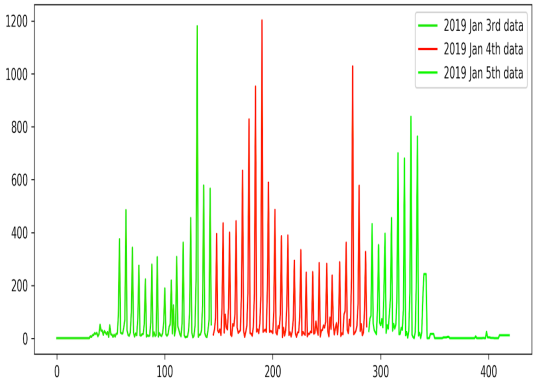

Figure 7: train cluster 3

Next, we created two more related features. We can use Ratios of time bins using below equation:

$$R_t = \frac{P_t^{train}}{P_t^{test}}$$

where P_t is the value of people density at time t. We can also use previous value of test set itself to determine the next value. We apply data modelling and use simple algorithms to determine which feature is best to forecast population density in a region. We use Simple Moving Averages, Weighted Moving Averages and Exponential Weighted Moving Averages and apply to each feature individually to assess which feature is more useful for our analysis.

4.1 Simple Moving Averages

The First Model used is the Simple Moving Averages Model which uses the previous n values in order to predict the next value. Using Ratio Values:

$$R_t = (R_{t1} + R_{t2} + R_{t3}....R_{tn})/n$$

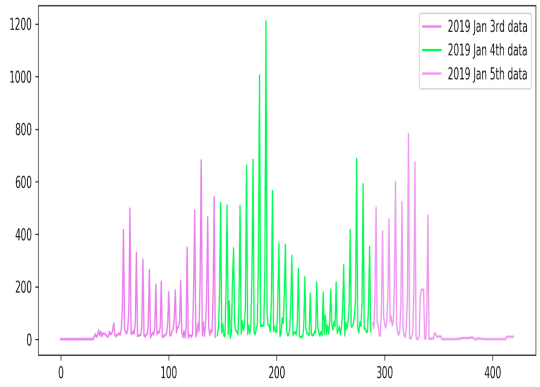

Figure 8: train cluster 4. The result of smoothing the time-series data by filing the gaps gives us continuous wave form of the data. The result above is shown time-interval of some clusters

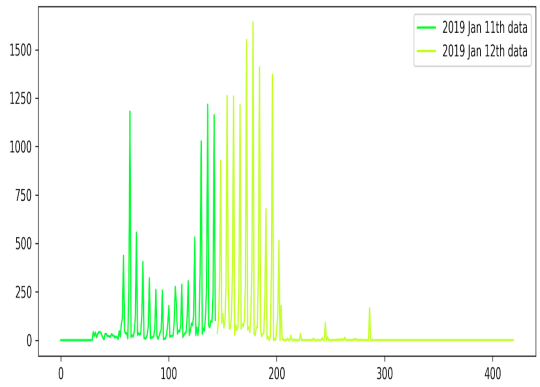

Figure 9: test cluster 1

For the above the Hyper-parameter is the window-size (n) which is tuned manually and it is found that the window-size of 3 is optimal for getting the best results using Moving Averages using previous Ratio values therefore we get $R_t = (R_{t1} + R_{t2} + R_{t3})/3$. Next we use the Moving averages of the test values itself to predict the future value using

$$P_t = (P_{t1} + P_{t2} + P_{t3}....P_{tn})/n$$

For the above the Hyper-parameter is the window-size (n) which is tuned manually and it is found that the window-size of 1 is optimal for getting the best results using Moving Averages using previous test values therefore we get $P_t = P_{t1}$

4.2 Weighted Moving Averages

The Moving Avergaes Model used gave equal importance to all the values in the window used, but we know intuitively that the

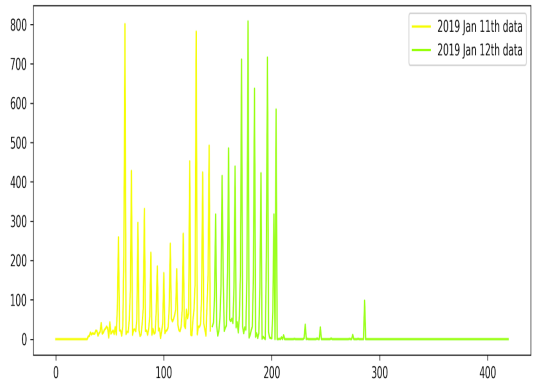

Figure 10: test cluster 2. The same days of the week show similar pattern. Please note that the result above corresponds to 4th and 5th January as these are the same day of the week.

future is more likely to be similar to the latest values and less similar to the older values. Weighted Averages converts this analogy into a mathematical relationship giving the highest weight while computing the averages to the latest previous value and decreasing weights to the subsequent older ones. Weighted Moving Averages using Ratio Values -

$$R_t = (NR_{t1} + (N1)R_{t2} + (N2)R_{t3}....1R_{tn})/(N(N+1)/2$$

For the above the Hyper-parameter is the window-size (n) which is tuned manually and it is found that the window-size of 5 is optimal for getting the best results using Weighted Moving Averages using previous Ratio values therefore we get

$t = (5Rt1 + 4Rt2 + 3Rt3 + 2Rt4 + Rt5)/15$

Weighted Moving Averages using Previous test Values -

$$Pt = (NPt1 + (N1)Pt2 + (N2)Pt3....1Ptn)/(N(N+1)/2)$$

For the above the Hyper-parameter is the window-size (n) which is tuned manually and it is found that the window-size of 2 is optimal for getting the best results using Weighted Moving Averages using previous test values therefore we get

$$Pt = (2Pt1 + Pt2)/3$$

.

4.3 Exponential Weighted Moving Averages

Through weighted averaged we have satisfied the analogy of giving higher weights to the latest value and decreasing weights to the subsequent ones but we still do not know which is the correct weighting scheme as there are infinitely many possibilities in which we can assign weights in a non-increasing order and tune the the hypeparameter window-size. To simplify this process we use Exponential Moving Averages which is a more logical way towards assigning weights and at the same time also using an optimal window-size.

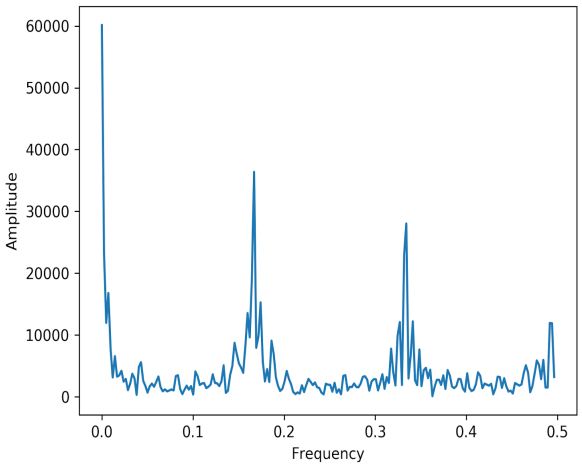

Figure 11: Fourier Transformation Features. There is a peak for daily pattern and the other peak corresponds to the busy times of the day, one during morning and the other again we see a peak time during the evening when people are heading home from work.

In exponential moving averages we use a single hyper-parameter alpha (α) which is a value between 0 1 and based on the value of the hyper-parameter alpha the weights and the window sizes are configured. For eg. If $\alpha = 0.9$ then the number of days on which the value of the current iteration is based is $1/(1\alpha) = 10$ i.e. we consider values 10 days prior before we predict the value for the current iteration. Also the weights are assigned using $2/(N+1) = 0.18$,where N = number of prior values being considered, hence from this it is implied that the first or latest value is assigned a weight of 0.18 which keeps exponentially decreasing for the subsequent values.

$$R_t^{'} = \alpha * R_{t-1} + (1-\alpha) * R_{t-1}^{'}$$

and

$$P_t^{'} = \alpha * P_{t-1} + (1-\alpha) * P_{t-1}^{'}$$

We finally evaluate these results using MAPE and MSE in the next section.

5 PERFORMANCE ANALYSIS

We will evaluate the performance of the our model we will use both visualization and statistical methods. We try to first find the best number of clusters by using inter cluster distance. We then visualize these clusters in heat-map plots on a geo-coded map of Richmond city and use human annotations to judge the correctness of our predictions. Here we are in a way trying to predict the number of people in a cluster in a particular time window. We would want to reduce the difference between the actual number of people and the predicted number of people. We also use Mean Absolute percentage error and Mean Squared error as our key

performance indicator when evaluating our prediction models. We use these measures to find the percentage error in our prediction of the number of people in a particular cluster. We aim to minimize the mean squared percentage error across all the time windows to get a good prediction score. We plan to target percentage error and not the actual numbers since this makes more sense to talk about percentages in terms of business perspective and provides a better insight in the performance of our model. As the businesses are more interested to know the percentage fluctuation in the prediction of the number of the people in a given area in a given time.

As noted above, we have chosen our error metric for comparison between models as MAPE (Mean Absolute Percentage Error) so that we can know that on an average how good is our model with predictions and MSE (Mean Squared Error) is also used so that we have a clearer understanding as to how well our forecasting model performs with outliers so that we make sure that there is not much of a error margin between our prediction and the actual value. Figure 12 shows our final results. For the scale of our data, we find that Ratio features perform better than Previous value features.

final.PNG

Error Metric Matrix (Forecasting Methods) - MAPE & MSE

Moving Averages (Ratios) -	MAPE: 1.11499	MSE: 222312.32457
Moving Averages (Previous Values) -	MAPE: 1.3601	MSE: 42144.06067
Weighted Moving Averages (Ratios) -	MAPE: 1.20318	MSE: 239628.71267
Weighted Moving Averages (Previous Values) -	MAPE: 1.41722	MSE: 35850.06319
Exponential Moving Averages (Ratios) -	MAPE: 1.15711	MSE: 293570.36686
Exponential Moving Averages (Previous Values) -	MAPE: 1.37604	MSE: 34542.71481

Figure 12: Final results for our models. We see that the ration features perform better than the previous value features.

6 CONCLUSION

To summarize, in this project we analyzed the Richmond city people movement data and build a model to predict human movement profitable hot-spots on the basis of time bins. By analyzing the hot-spots of people movement in a ten minute time-window, we trained our model that will be able to bring in profits to the small and big companies in smart business and marketing decisions to target the right locations. We predicted the number of people in a region for a given time, which accomplished our task. The result of the the analysis can then be used by variety of domains.

As for our results. Simple Moving Averages gives a good prediction of the population density with Ratio features. Same days of week follow same pattern in a given location. Companies that want to display their advertisement during the time-bins with more population density in the predictions should be charged more.

7 FUTURE WORK

AS for our future work we could Use more data volume to get data continuity and robust predictions. Also this data set could be combined with various other data sets to have great practical uses. Also future work could be to apply more sophisticated machine learning models like deep learning models to have better predictions on the peoples movement.

REFERENCES

[1] KAYLA FERGUSON. 2017. The Latest Trend in Data Collection — Smart Billboards That Look Back at You. (2017).
[2] Deepak Ganesan. 2017. Chapter 8: GPS Clustering and Analytics. http://web.cs.wpi.edu/~emmanuel/courses/cs528/F18/slides/papers/deepak_ganesan_GPS_clustering.pdf
[3] F. Meneses and A. Moreira. 2012. Large scale movement analysis from WiFi based location data. In *2012 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. 1–9. https://doi.org/10.1109/IPIN.2012.6418885
[4] Eran Toch, Boaz Lerner, Eyal Ben Zion, and Irad Ben-Gal. 2018. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems* (03 2018). https://doi.org/10.1007/s10115-018-1186-x
[5] Microsoft Research YU ZHENG. 2015. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology* 6, 38–46. https://doi.org/10.1145/1345448.1345455