CYPLAN 257

Daily trip distribution analysis of Oslo bicycle-sharing system using PCA and K-means clustering

Ola Tranum Arnegaard
UC Berkeley, BGA Exchange Fall 2019
Norwegian University of Science and Technology
olatar@stud.ntnu.no

This paper serves as an introductory analysis of the trip dynamics in Oslo bicycle-sharing system (BSS), *Oslo Bysykkel,* using PCA and K-means clustering on the weekday trip data of 2019, as well as some demographic information of the city. The results show that Oslo BSS is heavily influenced by a work/school commuter pattern.

Introduction

Information about the trip distribution dynamics may provide valuable insights towards solving common problems in a bicycle sharing system (BSS), e.g. station location planning and the redistribution problem. Extensive research is available in the domain of BSS's – many of which providing advanced predictive algorithms[1]. However, varying demography and geography are reasons why local insight is important before applying predictive algorithms. This paper will provide such an introductory analysis into understanding the local the trip distribution patterns in Oslo BSS. A substantial part of the paper will be dedicated to exploring how the work/school-commuter patterns may influence the overall trip distributions.

The analysis is based on the 2019 weekday trip data in Oslo BSS. An overview of the format is provided in figure 1. In "Household Energy Consumption Segmentation Using Hourly Data", Kwac et al. present a way to represent a large set of electricity consumption data as a small set of electricity consumption patterns through a method of dimensionality reduction[2]. The same concept will be practiced in this paper; however, through a different approach. The methods that will be applied in this paper are Principle Component Analysis and K-means clustering. Subsequently, all pickup and drop-off patterns across

[1] Dimitrios Papanikolaou, "Reconstructing, Visualizing, and Simulating Dynamics of Mobility on Demand Systems for Scenario Analysis, part 4

[2] Jungsuk Kwac, June Flora and Ram Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data", page 424, part D

Oslo will be represented by a handful of distributions, similarly to Kwac et al. Consequently, it simplifies how to conceptually understand and compare the trip distribution patterns in Oslo.

Trip Feature	**Format**
Start time	Date and time
End time	Date and time
Start location	Coordinates — *WGS 84*
End location	Coordinates — *WGS 84*
Duration	Seconds

Figure 1: the trip data from Oslo BSS, *Oslo bysykkel*.
The data is made accessible through *Norsk lisens for offentlige data (NLOD) 2.0* [Norwegian license for public data 2.0].

Pre-processing

The data are formatted as individual trips; therefore, some pre-processing is needed to get the data on the desired format. First, weekday pickups and drop-offs at each station are separated into two different datasets. This is because we wish to compare these distributions later. Then, all trips are binned to the closest hour of the day, creating a daily histogram of trips for each station with 24 bins. Subsequently, all daily histograms per station for weekdays in 2019 are combined to a single histogram by taking the mean of all trips per hour of the day. Finally, all distributions are normalized such that we have a single representative trip distribution for pickups and drop-offs for each station, see figure 2. Naturally, some stations are less busy than others, and the normalization will eliminate that information of the data. However, this is not an added aspect of this analysis, although it may provide valuable insights as well. Like Kwac et al.[3], the pre-processing pipeline may be reduced into (1) mean over all days for each hour for a given stations and (2) normalization of the distribution for a given station:

$$M^{station}[h] \frac{\sum_{day=1}^{total\ days} t_{day}^{station}[h]}{\boldsymbol{total\ days}} \quad (1)$$

$$N^{station}[h] \frac{\sum_{day=1}^{total\ days} t_{day}^{station}[h]}{\boldsymbol{total\ days}} \quad (2)$$

[3] Jungsuk Kwac, June Flora and Ram Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data", page 421, equation (1)

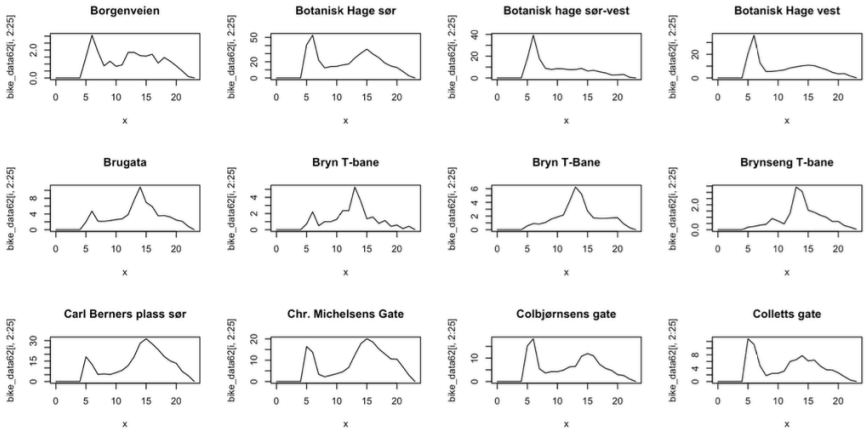

Figure 2: Daily pickup distributions for some stations in Oslo, 2019.
Y-axis: number of trips (normalized). X-axis: time of the day.

Principle Component Analysis and K-means clustering

PCA is applied to the trip distributions for both pickups and drop-offs. As seen from the Pareto plot in figure 3, 90.4% of the total variance is explained with only 4 principle components (PC′s) for bicycle pickups. Similarly, 4 PC′s for the drop-off distributions explains 91.4%.

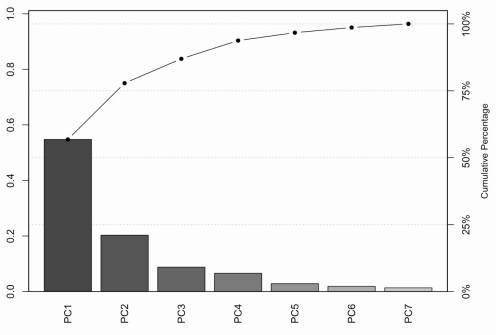

Figure 3: Pareto histogram of the variance explained
of the pickups distributions with a given number of components.

Subsequently, the K-means algorithm is applied. The optimal number of clusters is not clear-cut, but there are several methods that may be helpful. An attempt at the Elbow method gives no useful result as no defined elbows are apparent in figure four. Consequently, the parameter is selected by visualization of the reconstruction of the centres of the clusters, seen in figure five. With four clusters, the first three reconstructions are distinct; however, the fourth is similar

to the third. Therefore, three clusters is the optimal choice. Note that the mentioned figures are produced with the pickup data; however, the results are similar for the drop-offs.1

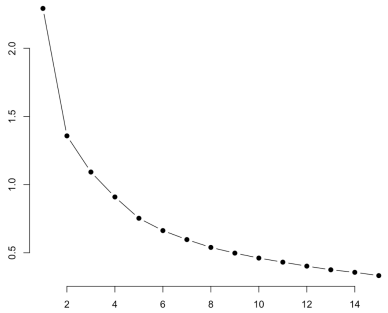

Figure 4: Elbow method for K-means of PC′s from pickup distributions.
X-axis: number of clusters.

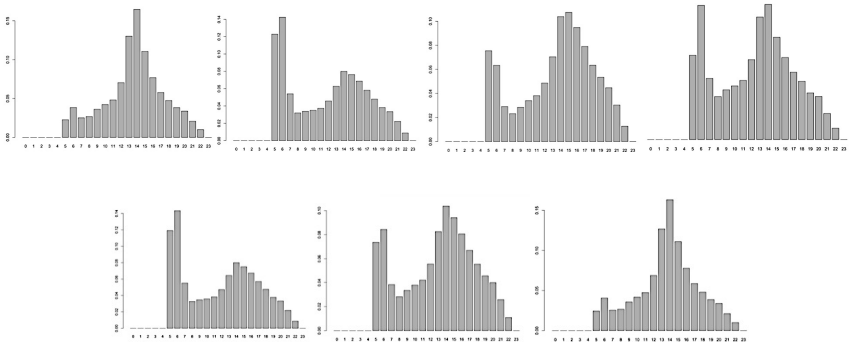

Figure 5: Reconstruction of centers from each cluster.
Above: 4 clusters. Below: 3 clusters. Y-axis: number of trips (normalized).
X-axis: time of the day.

Results

Pickups

Figure 6 shows the three reconstructed centers from the K-means clustering and their corresponding clusters. *Cluster 3* (C3) corresponds well with the residential areas in Oslo[4]. Contrary to C3, C1 appear frequently in three central areas: the business districts in Skøyen and Karl Johan, as well as the University of Oslo (UiO). C2 is located in all areas, however slightly more in residential

[4] *Residential density* and *Business density* in Oslo: https://kart.ssb.no/

regions. Conclusively, pickups distributions in residential areas suffer from high peaks in the morning, with slightly less pickups throughout the day. In contrast, the business areas and UiO experience most pickups in the evening. Intuitively, this is close to what some would hypothesize; that workers and students pick up bicycles in the morning and after they are done for the day – a system influenced by a work/school commuter pattern.

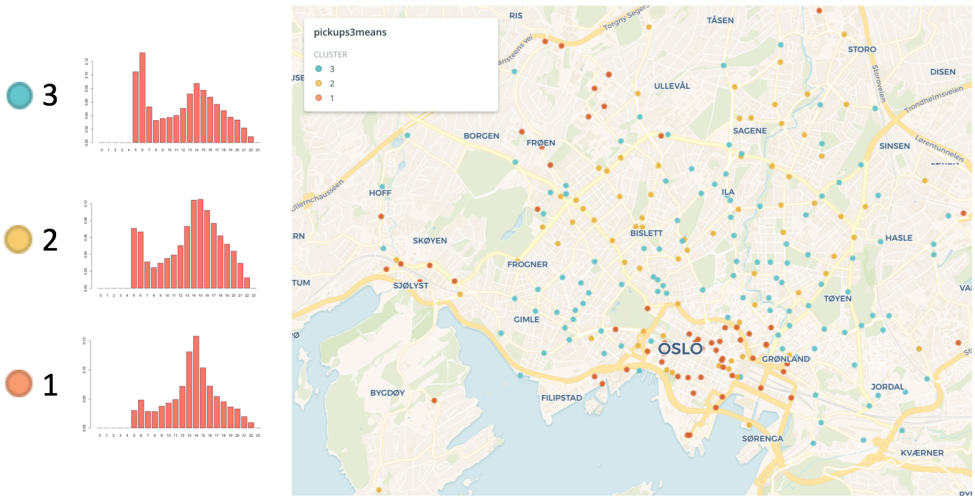

Figure 6: Reconstructed centers and map of the stations in Oslo and their corresponding clusters. 1: Bimodal, with peak of pickups in the morning 5am-8am, lower in the evening 1pm-4pm. 2: Bimodal, with higher peak of pickups in the evening than in the morning, however not a large gap. 3: Unimodal, with a peak in the evening. The map shows which of the given clusters each station corresponds to. Areas: *Ullevål-Frøen* area corresponds to University of Oslo, and capital letters *Oslo* and *Skøyen* correspond to the two major business areas in Oslo. Y-axis: number of trips (normalized). X-axis: time of the day.

Drop-offs

Figure 7 illustrates that the drop-off distribution reconstructions follow the same patterns as the pickup clusters, but with even more distinct centers. C3 corresponds with residential areas, similar to C3 and C2 in pickups. C2 suffer from most drop-offs in the morning, and corresponds with business areas as well as UiO, like areas from C1 in pickups. C1 also corresponds well with business areas; however, also close to the park *Vigelandsparken* in *Frogner*.

Intuitively, this shows that people travel to the central areas in Oslo as well as the famous park both in the morning and in the evening. However, some areas in central Oslo as well as UiO are only morning destinations. By intuition these seems to be those areas of central Oslo that are heavily influenced by offices and contain fewer social points of interest, e.g. stores and restaurants.

Finally, C3 shows that large residential areas suffer from most drop-offs in the evening, strengthening the theory from the pickups section, that bicycle stations in Oslo are heavily influenced by a work/school commuter pattern.

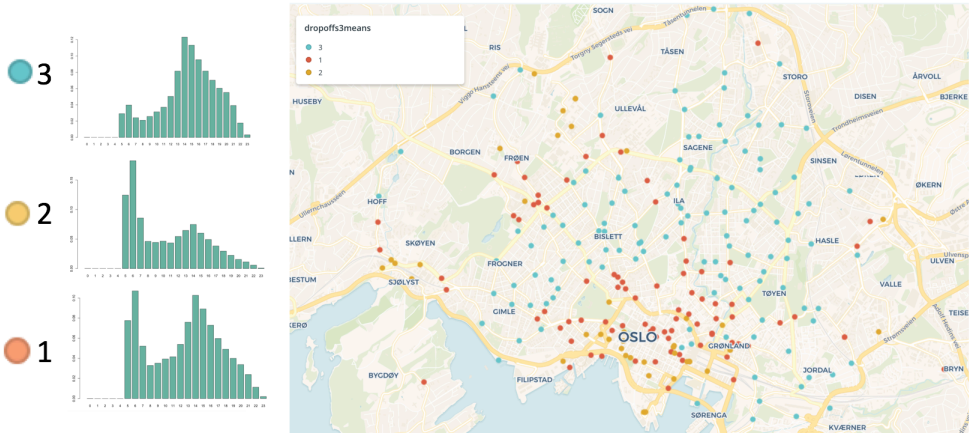

Figure 7: Reconstructed centers and map of the stations in Oslo and their corresponding clusters. 1: Almost unimodal, with prominent peak of drop-offs in the evening 1pm-4pm. 2: Barely bimodal, with prominent peak of drop-offs in the morning. 3: Strongly bimodal, with equal peaks of drop-offs in the morning and evening. The map shows which of the given clusters each station corresponds to. Areas: *Ullevål-Frøen* area corresponds to the University of Oslo, *Frøen-Frogner* corresponds to the park *Frongerparken*, and capital letters *Oslo* and *Skøyen* correspond to the two major business areas in Oslo. Y-axis: number of trips (normalized). X-axis: time of the day.

Comparison

After separately analyzing the pickups and drop-off patterns, a natural next step is to compare them. In this case, one interesting comparison is to look at distributions that are not complementary to each other, thus paramount to the rebalancing problem. First, the combination that results in empty stations, and secondly the combination that may cause overloading. Figure 8 shows stations that suffer from one of the two combinations. The clusters are spatially quite well separated, with C3 in residential areas and C1 around central business areas, as well as UiO. Intuitively, stations in residential areas are prone to scenarios with less or none bicycles during the day because of the peaks in the morning and minimal returns until the evening. Similarly, C1 suffer from the same problem, just the other way, i.e. overloaded during the day. Note that the map only show stations with one of the given combinations and will therefore not illustrate all stations. This is because some stations experience more even load throughout the day than others, and may therefore operate closer to autonomously.

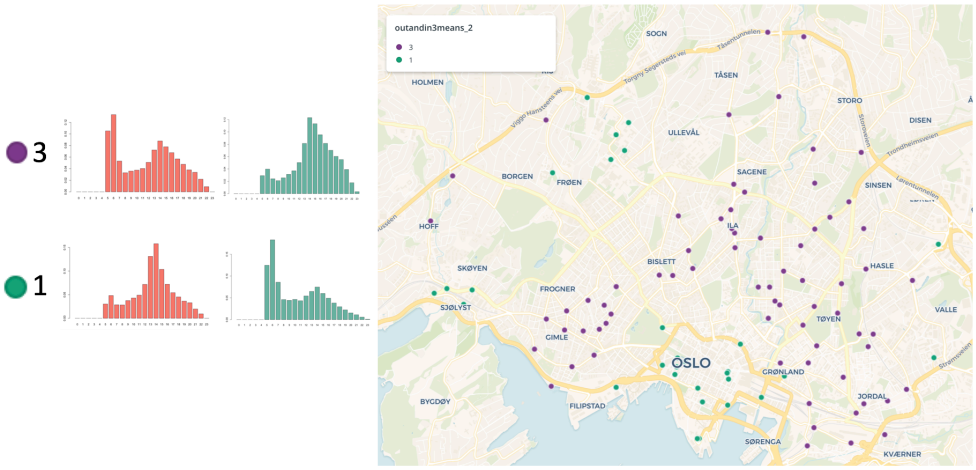

Figure 8: Left: pickup distributions. Right: drop-off distributions. The map shows those stations that suffer from either combination 3 or 1. Cluster 3: combination that may lead to empty stations during the day. Cluster 4: combination that may lead to full stations during the day.

Conclusion

The results from the introductory analysis may be considered as close to what most would hypothesize. However, it should be noted that not all BSS's produce similar results; therefore, introductory analyses such as this one may still be insightful.

A comprehensive conclusion may be ambitious and should perhaps be avoided, as detailed interpretations may not be hugely accurate because of the generality in this study. However, it may be concluded that Oslo BSS is heavily influenced by a work/school commuter pattern. This is prominent for most of Oslo city, and especially in the two central business areas and at the University of Oslo. The study was rather general: only three reconstructions to represent all the 248 different stations, and only some social and spatial knowledge [4]. Therefore, it is interesting that also *Vigelandsparken* was prominent, and apparently, a popular attraction throughout the year. This conclusion is made on the fact that the trip distributions in this study are created by means over all weekdays in 2019. Therefore, one may assume that weeks with dynamics that may be considered as outliers, such as vacations, should roughly be diluted by the typical week-to-week dynamics. Therefore, it is safe to assume that *Vigelandsparken* is more than a vacation attraction.

Future Work

This project has served as an introductory analysis the demand patterns in Oslo BSS. Future work will naturally go more into details. However, before delving into complicated predictive algorithms, it may also be wise to do a similar study on how vacations and weather may influence the trip distributions - two additional major factors in BSS dynamics. A similar analysis may easily be constructed: normalize weekly trip distributions instead of stations and see which cluster each week correspond with. A similar approach may be created for weather conditions. Conclusively, these three measures will give a strong indication of the dynamics in Oslo BSS. After the general analyses, more detailed predictive algorithms may be constructed. Even though the introductory studies are general, it will give some indicators about what those algorithms may output.

References

[1] Dimitrios Papanikolaou, "Reconstructing, Visualizing, and Simulating Dynamics of Mobility on Demand Systems for Scenario Analysis, part 4

[2] Jungsuk Kwac, June Flora and Ram Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data", page 424, part D

[3] Jungsuk Kwac, June Flora and Ram Rajagopal, "Household Energy Consumption Segmentation Using Hourly Data", page 421, equation (1)

[4] *Residential density* — *Population density* — *Business density* in Oslo: https://kart.ssb.no/