ANALYSIS OF QUUPPA TRACKING DEVICE DATA AT THE SMITH LEARNING THEATRE

Gary Nguyen

Abstract: In this short research project, I applied three clustering algorithms, namely k-means clustering, Gaussian Mixture Model (GMM) with expectation maximization, and mean-shift algorithm to analyze the Quuppa tracking location data in order to understand group membership and dynamics in an educational setting. On a higher level, I hope to understand how location-based data is used to further smart education and learning. Clustering is valuable in understanding how participants interact, form groups and change group memberships over time. The dataset I used is provided by EdLab, Teachers College, containing movements of 26 distinct sensors 19 seconds. I also provided briefly the advantages and disadvantages of each clustering algorithm.

1. INTRODUCTION

With the development of Internet of Things and tracking device, the amount of personal data has increased rapidly in the past decade, giving rise to opportunities to model human movements and behaviors. One such application is the analysis of group dynamics using tracking system data in learning spaces. These analyses are beneficial to both educators, event organizers and participants alike. Event organizers and educators can incorporate the insights from location analytics to design highly effective activities and event formats, potentially revolutionizing how humans learn and interact with each other [4]. As a result, this research topic contributes directly to smart education.

The main goal of this research is to see how clustering changes with different clustering algorithms, and which algorithms are good candidates to analyze bigger datasets. While I want to understand more about how group dynamics changes with time, 19 second is a very short timespan for any meaningful change to take place. In the scope of this research, I analyzed the groupings of participants of an event at The Smith Learning Theatre in a single-second snapshot. The snapshot is chosen so that the highest number of distinct Quuppa trackers were in use.

2. LITERATURE REVIEW

Introduction: smart education and location system

Analyzing location-based data has long been employed to understand smart education. Smart education is defined as the prominent paradigm in global education with the aim of improving learner's quality of lifelong learning. The focus of smart education is contextual, personalized and seamless learning [6], all of which are supported by the Quuppa tracking system data analysis. According to NMC [7], IoT, wearable technology and location-based data analysis enhances the advancement of "contextual and seamless learning" by synchronizing people and objects. "Wearable technology can integrate the location information, exercise log, social media interaction and visual reality tools into the learning." [7] To support that purpose, according to Bartels, smart computing is the latest cycle in tech disruption, which fuels the rise of smart learning. It blends all aspects such as software, hardware, network, smart devices, sensors, big data and intelligent system to materialize innovative applications in education. "All these technologies can effectively support learning to happen in different situations. Above all, the advancement of computing technologies leads smart computing to a new dimension and improves the ways of learning". [8]

Smart education is usually included under the scope of smart cities. In these innovative cities, citizens have the opportunities to learn anywhere and anytime, which produces a high volume of behavioral and location-based data for analysis. How to integrate these data in smart cities and, as a result, build data-centric smart education remains a formidable challenge to educators. "The interconnected and interoperable learning service and experience between smart education system and other systems of smart city" [9] is the area focus of the future.

Smart education and location-based data in the context of Smart Cities

Smart education is one of the core parts of developing smart cities and communities. For example, in smart cities, tracking systems and their data can be safely utilized to optimize academic research [10]. In the past, to be able to evaluate the effectiveness of an educational activity or to conduct a research, educators had to either ask participants to fill out information by hand, or observed the participants diligently themselves. Tracking system such as Quuppa makes this process easier, especially when there are multiple research participants involved. In addition, educational institutions, such as Teachers College at Columbia, can easily assess their allocation or resources and their success level more easily.

This data analysis can give rise to important educational insights on a societal level. The Quuppa system and the tracking sensors, as smart devices, can enhance education effectiveness and productivity, as well as can enhance support for lifelong learning [10] for a smart city or community. As researches into the advancement of smart cities are attracting international attention, analysis of these cities' citizens, their group dynamics, how they move and learn, and how that changes over time, can reveal insights into how to build a community conducive to life-long education. This will in turn gives rise to the emergence of new teaching and learning tools to deliver and acquired knowledge [10]. Also these analyses can be used by administrators and the government to identify fields of study that need structural changes, to "create communities of practice and standardize the presentation of knowledge", to "observe educational shortages" [11] and revise the curriculum to benefit everyone involved.

Some applications of location-based sensors and services may include:

1. Provide park visitors with personalized immersive and educational experience based on where they are in the park, using GPS system. [12]
2. Wanaka et al. used users' location data to estimate users' preference and personalized educational news recommendation. They found that the best method for news article recommendation is the method labeling location log data by Bayesian model "location hierarchical Dirichlet process" (LocHDP) [13].
3. Masaki Murata analyzed the visitors' stay duration at different booths through wifi system data from at a large expo to gain insights into which factors affect a visitor's interest in a certain booth. [14]

Challenges

Accompanying the myriad potential uses of this location data are great challenges that need to be addressed to avoid dangerous precedence. "These relate to available big data tools, real-time analytics, accuracy, representation, cost, and accessibility." [15] These issues, if not resolved, would undermine the use and performance of location-based data, and big data in general, in an educational setting:

1. Analytics architecture: with the growing velocity and complexity of location data, there needs to be effective data analytics architectures and pipelines to process, analyze and visualize the data for real-time analysis and applications. [15]
2. Data Quality: data quality is prone to both technological and human errors, and tend to deteriorate over time. According to Nuami et al [10], "sensors data collected through a third party without a centralized control could have been produced by sensors that are faulty, wrongly calibrated, or beyond their lifetime." It is essential to frequently update quality of data through algorithms, maintain the quality of sensors and update usage policies to ensure the performance of the location data [16].
3. Privacy concern: location data, like all other personal data, needs to be secured and protected. As the data travel through many types of network systems, it is exposed to security breaches [17]. To make the matter more complex, "most big data technologies today, including Cassandra and Hadoop, suffer from a lack of sufficient security." [18] Proper policies and regulations on consent and terms of uses should be put in placed and properly enforced to protect participants' personal data and protect the integrity of the analysis for the greater good.

The above challenges, coupled with the problem of increasing costs and lack of explainability, should be addressed very early in the research process to avoid the costly repairment and purchase of additional software or hardware.

To sum up, analysis of location-based data can provide many benefits to education and learning. One of such application is analyzing participants' clusters and group dynamics, which I am doing in this research.

3. DATA PREPROCESSING AND ANALYSIS

Data Preprocessing

The dataset contains 96776 location data points from 26 distinct trackers in 19 seconds. After converting the log files into a json dictionary, I took the X coordinate, Y coordinate and the recorded time, as well as the trackers' id and their names, to create a data frame for analysis. I dropped the data points where 'zones' is null (there are nine such instances) because I suspect these were technological errors. The overall density in 2D & 3D space is visualized below:

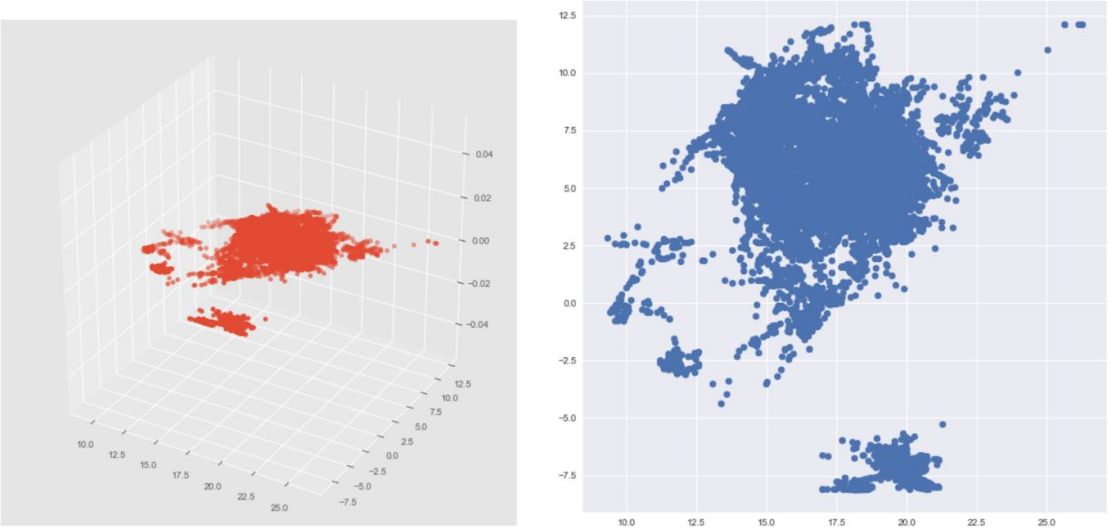


Figure 1: Overall density of Quuppa trackers

There appears to be 2-4 clusters. This estimation is necessary because two of the algorithms I will be presenting involves setting a prespecified number of clusters.

The individual trackers scatter plots are visualized below. From both Figure 1 and Figure 2, it appears that there are two main clusters: the one in the top part (Y coordinates higher than 0) and the one in the bottom right corner. The participants in the upper part also has the tendency to move around much more often compared with the one on the bottom. Without any a priori knowledge, I can hypothesize where the "activity centers" (e.g. exhibits, presentations etc.) are according to the density of the data points. Seven trackers concentrating around the bottom right corner can be either acquaintances or sharing a strong interest, as they tend to stay in a narrow area and do not move much over time.

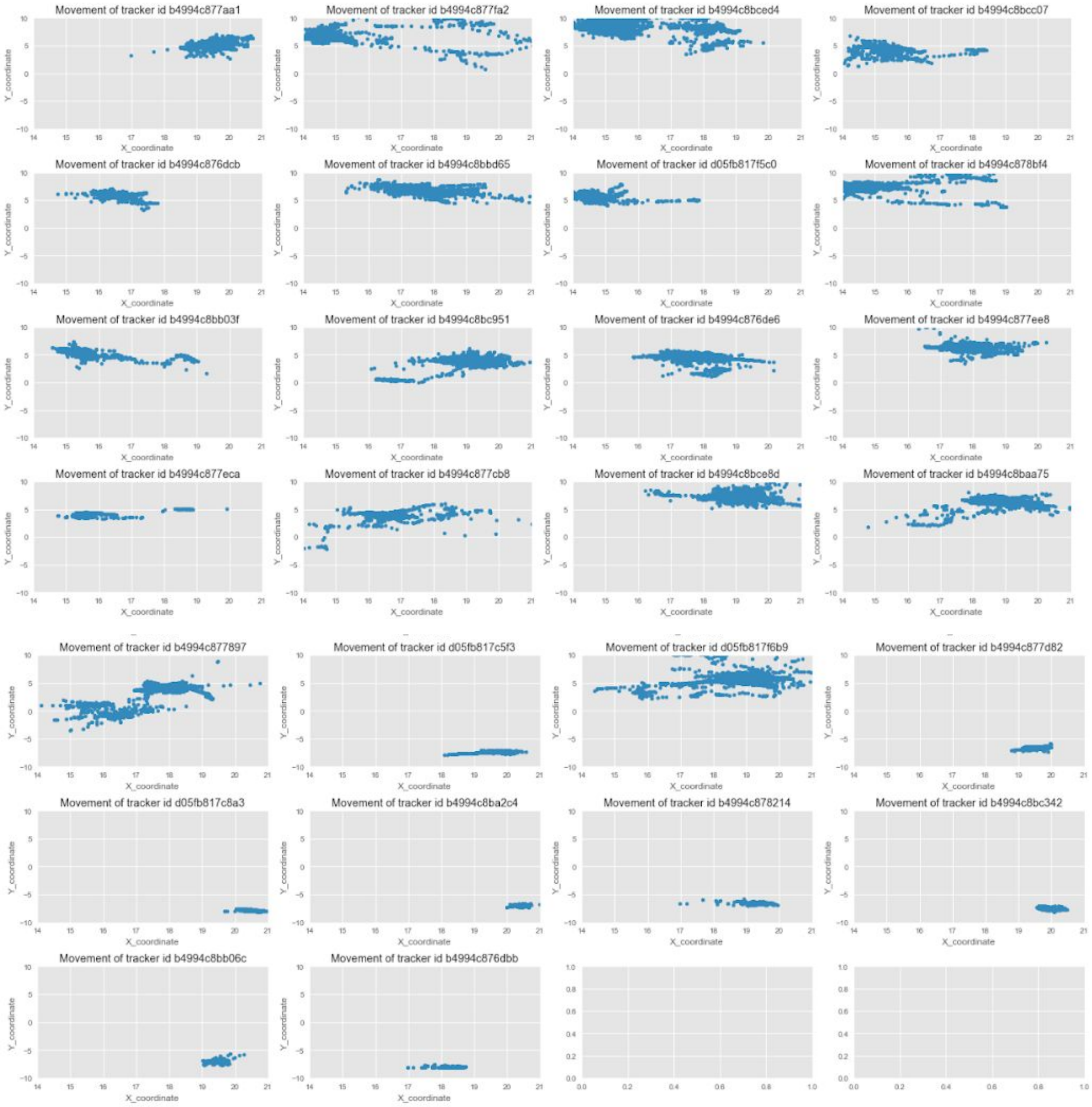


Figure 2: Movements of individual Quuppa trackers

Methods of Analysis

For this project, I used three main methods to analyze group dynamics: k-mean clustering, GMM with expectation maximization and mean-shift algorithm.

K-means clustering algorithm

K-mean clustering is the baseline clustering algorithm where group membership is dependent on precomputed distances to group centers and group centers are shifted around until convergence. It is one of the fastest clustering algorithms, yet it has major disadvantages: one has to pre-specify the numbers of clusters; in addition, it is limited to circle-shape clusters. Finally, because of the random initialization, k-mean clustering can result in different clustering every time.

Gaussian Mixture Model with Expectation Maximization

Gaussian Mixture Model utilized probability distributions to detect clusters. GMM can tackle the disadvantages of k-means algorithm as it has two parameters to "describe" the shape of the clusters: the mean and the variance of the Gaussian distributions. Each cluster is described by a single Gaussian distribution. The expectation maximization (EM) algorithm is then used to find the parameters of each Gaussian distribution [5]. The steps of GMM with EM are described as follows:

1. The algorithm begins by specifying the number of clusters and randomly initializing the Gaussian distribution parameters for each cluster.
2. We then compute the probability that each data point belongs to a particular cluster. The closer a point is to the cluster's center, the higher probability of it belonging to that cluster is.
3. Based on these computed probabilities, we calculate a new parameter set for the Gaussian distributions that maximizes the probabilities of data points within the clusters. One can compute these new parameters using a weighted sum of the data point positions, with the weights being the probabilities of the data point of a cluster.
4. Steps 2 and 3 are repeated until convergence.

GMM has two major advantages. First, GMM is more flexible in terms of covariance than k-means thanks to the standard deviation parameter. This means that the cluster can take on, for example, an ellipse shape instead of being restricted to circles. Secondly, since GMM uses probabilities, they can have multiple clusters per data point (i.e. GMM supports mixed membership). Therefore, if one data point is in the overlapping portion of two clusters, we can simply define it according to the probabilities.

Mean-shift Clustering Algorithm

Mean-shift is a non-parametric clustering technique. This algorithm is a sliding-window-based algorithm with the goal of locating the center point of each cluster. In each iteration, the algorithm updates center-point candidates of each cluster to be the mean of the points within the sliding windows. These candidate windows were also processed in a way that prevents near duplicates, forming the final set of center points and their corresponding groups [1]. The main advantage of mean-shift algorithm is that one doesn't have to specify the number of clusters in advance for the algorithm to work. However, one should specify the bandwidth for the algorithm [1]. The first disadvantage is that the selection of the window size is non-trivial. Secondly, it doesn't perform well in the presence of noise. Instead of finding an option for treatment of noise for mean-shift, one can use the DBSCAN algorithm, which can detect outliers as noises [5].

DBSCAN (this method is not used in the scope of this research, but it appears in "Future Work" section below)

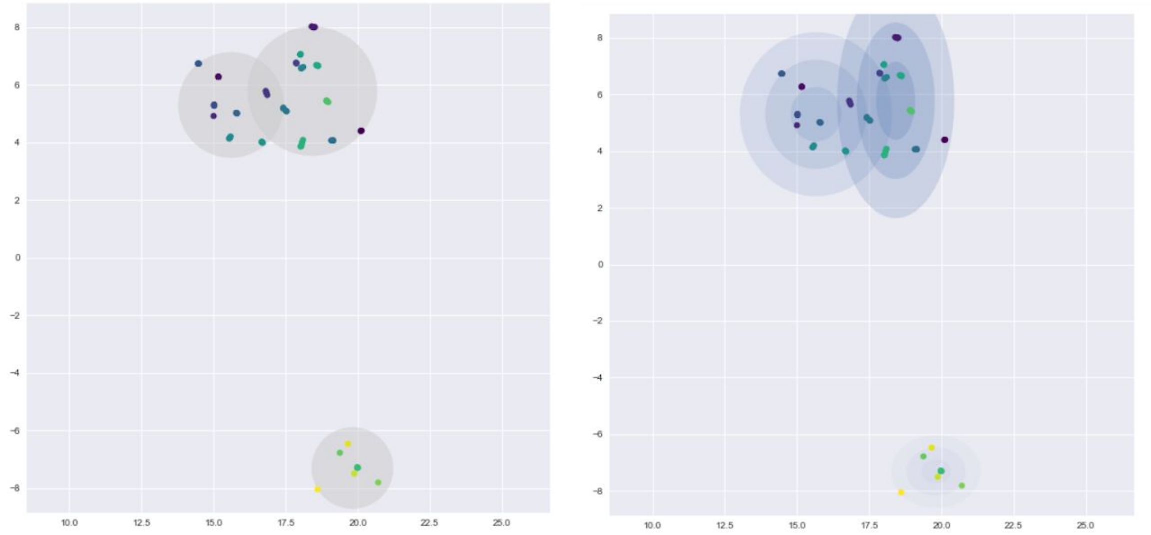
This is a density clustering algorithm that detects clusters based on density [5]. "Density clustering algorithms overcome the shortcoming that clustering algorithms based on distance can only group the spherical clusters. They can group clusters in any shape. The typical density clustering algorithms include DBSCAN (Ester et al. 1996) and OPTICS (Ankerst et al. 1999) algorithm." [3]

Results

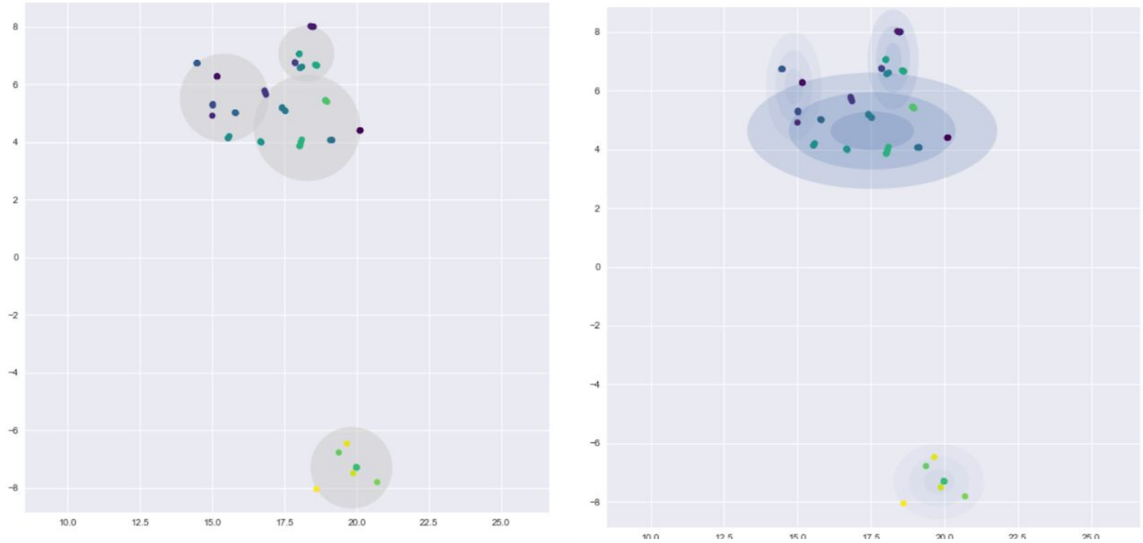
Because of the lack of a "true" group labels, a meaningful performance comparison among algorithms is difficult. Therefore, the main focus of this section is to compare the clustering result by different algorithms. I visualized the clusters detected by different algorithms below using scikit-learn implementation [1] and references from "Python Data Science Handbook" [5]. Each data point represents a tracker in a millisecond. Each color represent a distinct tracker. Since there are only 1-4 tracking signals per millisecond, each snapshot below represent a single second.

K-means and GMM Comparison with increasing number of predefined clusters:

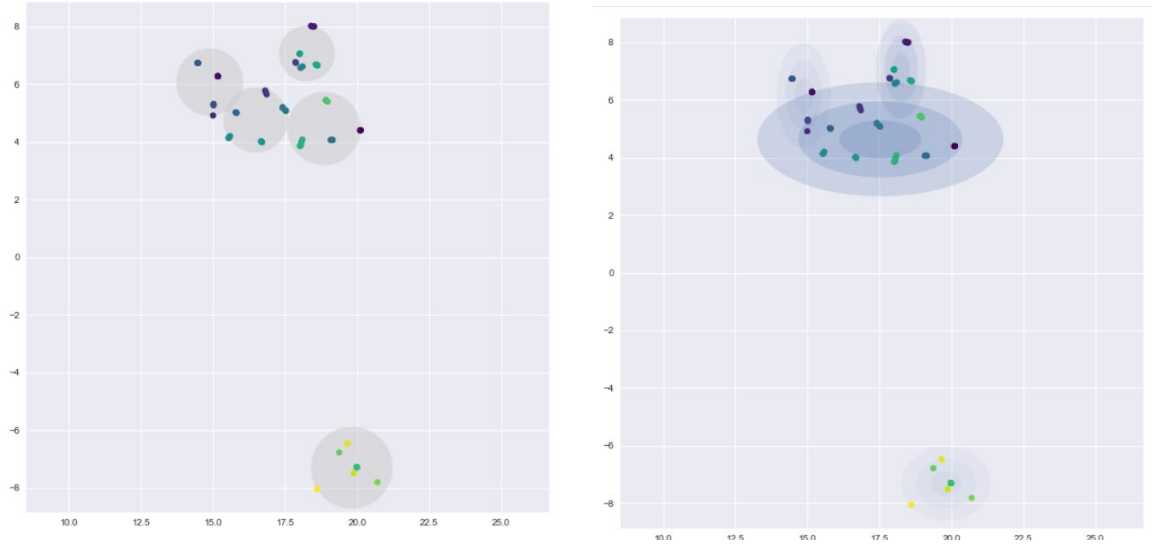
N_clusters = 3

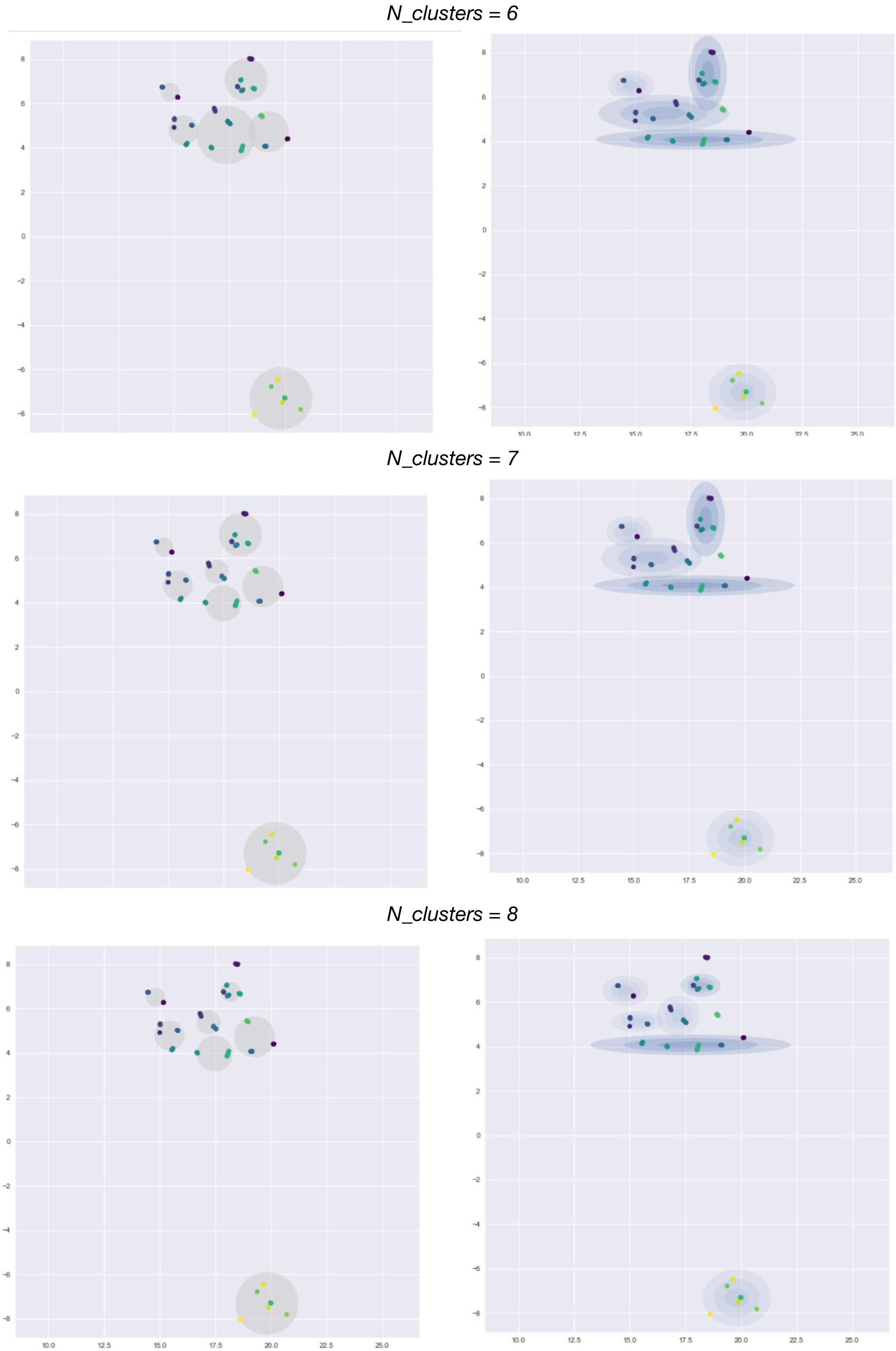


N_clusters = 4



N_clusters = 5





N_clusters = 6

N_clusters = 7

N_clusters = 8

Figure 3: Comparisons between k-means clustering and GMM as n_cluster increases

It is easy to observe that when the number of prespecified clusters increases, the clustering patterns differ more and more between k-means algorithm and GMM. This difference persists even with different starting points for k-means, because k-means assumes equal importance in all directions while GMM assumes that data points are Gaussian distributed. Therefore, with expectation-maximization algorithm, the clusters detected by GMMs come in different ecliptic shapes. In addition, naturally, as the number of cluster increases, mixed memberships become rarer for GMM. In this sense, GMM still has more advantages compared to k-means and we can easily compute and observe the probabilities of a data point belonging to a cluster.

Mean-shift Clustering Algorithm: As mentioned in the above section, we do not have to specify the number of cluster in advance. In addition, depending on the choice of bandwidth used, the KDE surface and end clustering result will be different [5]. It is observed from Figure 5 that the number of clusters decreases as bandwidth increases. Here I choose the bandwidth = 1.6 corresponding to 4 clusters. The result of mean-shift algorithm is in line with GMM and k-means, given the number of cluster is 4.

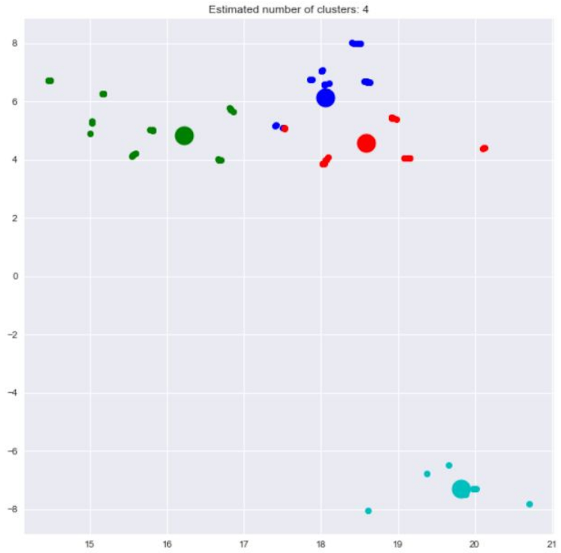
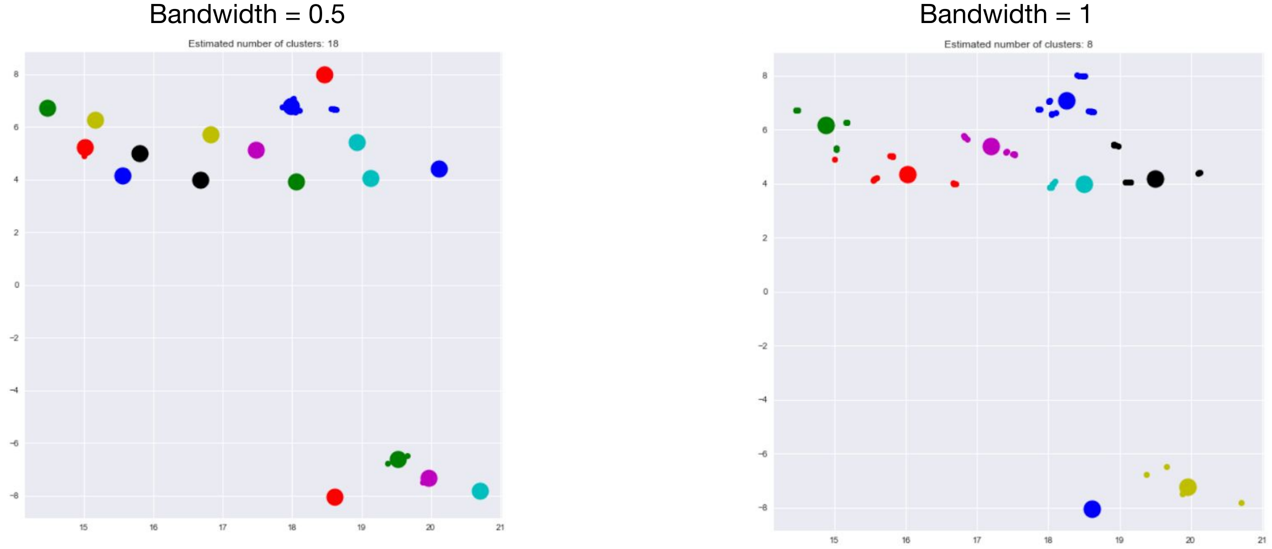
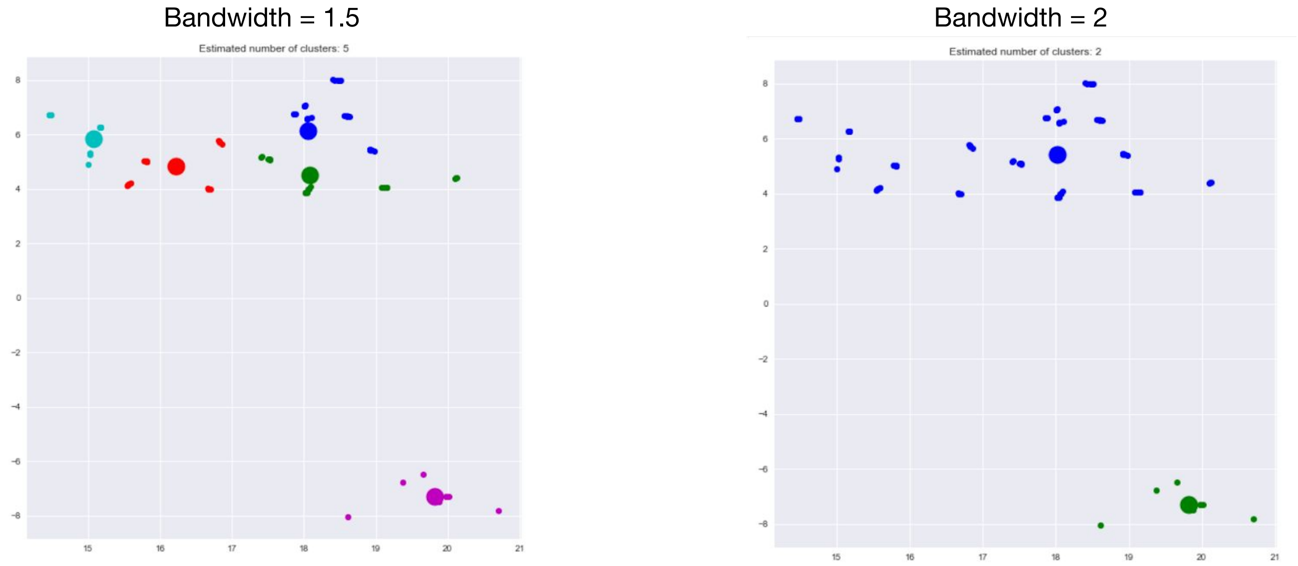


Figure 4: Mean-shift clustering


Bandwidth = 0.5

Bandwidth = 1


Bandwidth = 1.5

Bandwidth = 2

Figure 5: Mean-shift clustering with different bandwidth

4. DISCUSSION

Impact

After taking all factors into considerations, Mean-shift and GMM are best candidate models for analyze tracking data by Quuppa. There is a trade-off: while GMM can work with mixed memberships and has a probability component, it needs the number of cluster as human input, which is more prone to errors. Mean-shift algorithm can detect clusters without a priori knowledge, given a bandwidth specified by industry knowledge. However, mean-shift has a tendency of taking noises into its clustering. In both case, the chosen clustering algorithm will give sufficiently good insights into the group membership of each participants in a discrete snapshot in time. Educators can use this information to foster collaborations in their events, or to design the layout of the venue to achieve desired group outcomes. Another application is to design activities to break up pre-formed groups and facilitate cross-group collaborations to benefits everyone involved [4]. Facilitators can be assigned, for example, to improve group dynamics and lend support to participants that need to join a group.

The time span for this dataset is only 19 seconds. Therefore, while some participants move around a lot, there are not many meaningful shifts in group membership. However, with a larger datasets, the clustering algorithm can help answer questions such as “How long will a group remain its size or popularity?” and “How often will groups exchange members?” These are core to the educational experiences of participants, which educators should pay attention to.

Future Work

The method described in this research assumes that the different snapshots are independent of each other (i.e. each snapshot is discrete), and therefore group membership in one snapshot doesn’t affect the others. This is not true in practical situations where group membership may change depend on preceding group events. Therefore, a natural next step is to apply Dynamic Mixture Models (DMM) to analyze the change in group dynamics through time. DMM can be applied to time series and can explain similarities between movements of participant pairs, which cannot be explained by a single snapshot clustering [2].

Another similar approach is to apply time-series clustering using Dynamic Time Warping (DTW) and DBSCAN. Specifically, I want to suggest an approach where we calculate the “distance” between each pair of participants using fastdtw (a dynamic time warping implementation in Python). Then we can use DBSCAN, a density-based clustering algorithm, to detect cluster(s) of participants that have similar movements [3]. A code outline is as follows:

```
def fastdtw_distance(X, Y):
    import numpy as np
    from scipy.spatial.distance import euclidean
    from fastdtw import fastdtw
    distance, path = fastdtw(x, y, dist=euclidean)
    return distance

def create_pairwise_distance(X):
   # X is the movement data
   # create a matrix of pairwise distances between all elements in your X_data

X_distance = create_pairwise_dist(X_data)

dbscan = DBSCAN(eps=1.3, metric=fastdtw_distance)
dbscan.fit(X_distance)
```

Both approaches will be helpful for educational events and activities where there participants have to move around a lot. In such environment, understanding the underlying group dynamics and member relationships over time will assist in better activity design to improve interactions between group of people. In addition, this can be valuable in event planning to maximize space efficiency.

REFERENCES

1. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011
2. X. Wei, J. Sun, and X. Wang. Dynamic mixture models for multiple time series. Proceedings of the 20th international joint conference on Artificial intelligence, (Dmm): 2909–2914, 2007
3. A. Sawas, A. Abuolaim, M. Afifi and M. Papagelis, "Tensor Methods for Group Pattern Discovery of Pedestrian Trajectories," 2018 19th IEEE International Conference on Mobile Data Management (MDM), Aalborg, 2018, pp. 76-85.
4. Cluster Analysis of Real Time Location Data - An Application of Gaussian Mixture Models. Proceedings of the 10th International Conference on Educational Data Mining, 2017.
5. J. VanderPlas. Python Data Science Handbook. Chapter 5. Machine Learning. O'Reilly Media, November 2016.
6. Zhu et al., A research framework of smart education. Smart Learning Environment. Springer Open, March 2016.
7. New Media Consortium, The NMC Horizon Report: 2015 Higher Education Edition, 2015, pp. 1–50
8. A. H. Bartels, Smart computing drives the new era of IT growth. Forrester Inc. 2009.
9. R.G. Hollands, Will the real smart city please stand up? Intelligent, progressive or entrepreneurial? City 12(3), 303–320. 2008.
10. Nuaimi et al., Applications of big data to smart cities. Journal of Internet Services and Applications. Springer Open, December 2015.
11. Marsh O, Maurov-Horvat L, Stevenson O. Big Data and Education: What’s the Big Idea?. UCL Policy Briefing. 2014.
12. Shu, Chang. Location based Educational mobile application design and implementation . Electronic Thesis or Dissertation. Kent State University, 2017. OhioLINK Electronic Theses and Dissertations Center. 25 Feb 2019.
13. Shinnosuke Wanaka, Kota Tsubouchi. Location History Knows What You Like: Estimation of User Preference from Daily Location Movement [C]. Proceedings of International Conference, 2016. 8–13
14. Masaki Maruta, Yuta Sano, Kohei Yamaguchi, et al. Visitor Behavior Analysis Based on Large-Scale Wi-Fi Location Data[C]. Proceedings of Iiai International Congress on Advanced Applied Informatics, 2015. 55–60.
15. Fan W, Bifet A. Mining big data: current status, and forecast to the future. ACM SIGKDD Explor Newsl. 2013;14(2):1–5.

16. Bertot JC, Choi H. Big data and e-government: issues, policies, and recommendations. In Proceedings of the 14th Annual International Conference on Digital Government Research. ACM; 2013. pp. 1–10.
17. Khan M, Uddin MF, Gupta N. Seven V's of Big Data understanding Big Data to extract value. In American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the IEEE; 2014. pp. 1–5.
18. Kim GH, Trimi S, Chung JH. Big-data applications in the government sector. Commun ACM. 2014;57(3):78–85.