Utilizing EP-Means to Cluster Distributions of Airline Flight Data

# **Brandon Sams Winter 2021 https://brandonsams.github.io/**

# Which Domain?

What domain is this data going to come from? Please list 10 references (with a brief annotation) to use to make sense of what you’re doing with these data.

EP-Means, or Empirical Probability Means, is a variant of K-Means that attempts to cluster not individual data points, but collections of distributions. Flight data, in my opinion, is an excellent publicly available dataset that can be massaged to create distributions that can then be clustered using EP-Means.

*2015 flight delays and cancellations*. (n.d.). Retrieved February 14, 2021, from <https://kaggle.com/usdot/flight-delays>

This dataset offers a large download of historical data, broken down by flight. (600MB) I plan to start the analysis with this dataset, as it is very comprehensive and appears complete. Data is from 2015.

*Airline on-time performance statistics—Dataset by dot*. (n.d.). Data.World. Retrieved February 14, 2021, from <https://data.world/dot/airline-on-time-performance-statistics>

This dataset is very similar to the above dataset, with the exception that the data is from 2018. Perhaps this will give us a look into how clustering may have changed over time.

*Aviation data & statistics*. (n.d.). [Template]. Retrieved February 14, 2021, from <https://www.faa.gov/data_research/aviation_data_statistics/>

This is where the FAA hosts a large amount of statistical data for research, directly related to the airline industry.

chilamkurthy, K. (2020, October 23). *Wasserstein distance, contraction mapping, and modern rl theory.* Medium. <https://towardsdatascience.com/wasserstein-distance-contraction-mapping-and-modern-rl-theory-93ef740ae867>

EP-Means relies on the ability to find the “distance” between two distributions. This article goes into a specific way of measuring distance called the “Wasserstein Distance”, and connects this metric to modern Reinforcement Learning.

*Clustering probability distributions—Methods & metrics?* (n.d.). Cross Validated. Retrieved February 14, 2021, from <https://stats.stackexchange.com/questions/13186/clustering-probability-distributions-methods-metrics>

This stackexchange discussion shows some alternatives to EP-Means clustering.

*Epmeans: Ep-means algorithm for clustering empirical distributions in maotai: tools for matrix algebra, optimization and inference*. (n.d.). Retrieved February 14, 2021, from <https://rdrr.io/cran/maotai/man/epmeans.html>

This is the documentation page for a specific R package that is used for computing EP Means.

Henderson, K., Gallagher, B., & Eliassi-Rad, T. (2015). EP-MEANS: An efficient nonparametric clustering of empirical probability distributions. *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, 893–900. <https://doi.org/10.1145/2695664.2695860>

This is the landmark paper where I originally discovered EP Means. It highlights the efficiencies of this technique in regards to other distribution clustering techniques, and shows how one might implement this technique.

Olive, X., Strohmeier, M., & Lübbe, J. (2021). *Crowdsourced air traffic data from The OpenSky Network 2020* [Data set]. Zenodo. <https://doi.org/10.5281/ZENODO.3737101>

Another possible source of air traffic data. Data is from 2020. Data contains firstseen and lastseen properties, so duration can be computed.

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, *6*(1), 405–431. <https://doi.org/10.1146/annurev-statistics-030718-104938>

Detailed review of Wasserstein distances (possibly showing multiple types?)

Ye, J., Wu, P., Wang, J. Z., & Li, J. (2017). Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, *65*(9), 2317–2332. <https://doi.org/10.1109/TSP.2017.2659647>

Shows how Wasserstein distance can be used to cluster distributions, but no explicit mention of EP means appears to be present in this paper.

# Which Data?

What is the dataset you’ll be examining? Please provide a codebook if there is one or a link to the dataset as well as a detailed description.

I will be examining individual flight data. I will start by looking at this collection of data from 2015:

<https://www.kaggle.com/usdot/flight-delays>

Per the description:

*“The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.”*

This indicates that there may be additional datasets available from other years. This one, however, is a good starting place.

# Research Questions? Benefits? Why analyze these data?

How are you proposing to analyze this dataset? This is about your approach. Here, you’ll be proposing your research questions as well as justifications for why you’d offer these data in this way.

Specifically, I am interested to see if there exist ways to cluster airline data. Specifically, I am interested in looking at flight durations, but by filtering by airline and/or by route. There is a distribution of possible travel times between each pair of locations, for example, and those distributions can likely be clustered. I would like to know what the best number of clusters is for a given collection of distributions, and why.

This data deserves to be analyzed, because of the tremendous economic impact that the airline industry holds. Also, if there is a natural clustering of airlines, when picking a flight, that clustered information would/could become useful in that decision making process.

# What Method?

What methods will you be using? What will those methods provide in terms of analysis? How is this useful?

I plan on using the method of EP-Means to cluster the distributions of flight durations. If this clustering was filtered to have each distribution be representative of a single airline, this clustering could identify other airlines that are competitors. It could also highlight gaps in the industry where some airlines may be ignoring possible demand.

# Potential Issues?

What challenges do you anticipate having? What could cause this project to go off schedule?

I anticipate that I will have at least some issues with data cleanup. There is plenty of data out there on flights, but I have no idea how much is incorrect or needs to be filtered. I also was able to find data from 2015, 2018, and 2020. The data is all from different sources, so the datasets may not be comparable.

I also anticipate that the method of EP Means will be fairly difficult to implement. There is a stage where a “centroid”-ish distribution needs constructed, which may end up being difficult.

# Concluding Remarks

Tie it all together. Think of this section as your final report’s abstract.

The duration of a flight is not a single value. Sure, there is an estimated duration, but the true duration will come from somewhere in a distribution of possible durations. By examining historical flight-by-flight data, this paper aims to cluster different routes and different airlines together by looking at the probability distributions, and how those probability distributions compare. The method of EP-Means (Empirical Probability Means) is used to cluster these collections of distributions. Distributions are compared by using the Wasserstein Distance, which is sometimes referred to as the Earth Mover’s Distance. Analysis also takes place that examines the proper number of clusters to combine the distributions into.