Statistical Engineering Approach to Improve the Realism of Computer-Simulated Experiments with Aircraft Trajectory Clustering

Sara R. Wilson and Kurt A. Swieringa

National Aeronautics and Space Administration

Langley Research Center

Hampton, VA

Robert D. Leonard, Evan Freitag, and David J. Edwards

Virginia Commonwealth University

Richmond, VA

Fall Technical Conference

October 5, 2018

Outline

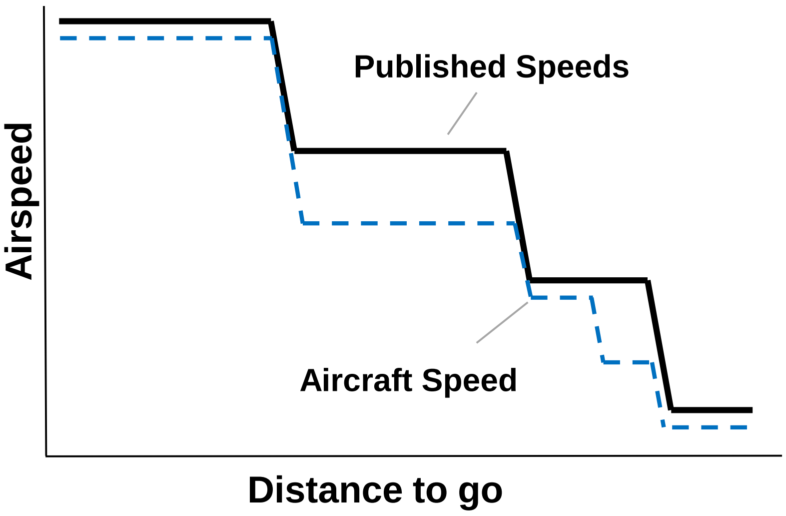
- Background and Motivating Example
- Cluster Analysis
- Challenges
- Trajectory Clustering Methodology
- Application of Trajectory Clustering
- Conclusions

NASA Interval Management Technology

- **NASA is developing and demonstrating technologies that enable the use of efficient aircraft arrival procedures during high demand operations**
- **Interval Management is a technology that enables an aircraft equipped with NASA's spacing algorithm and avionics to achieve precise spacing behind a lead aircraft**
- **Fast-time simulation was planned to evaluate recent modifications to the spacing algorithm**

NASA Interval Management Technology

- **A variety of realistic aircraft trajectories were needed for this fast-time simulation**
- **Previous fast-time simulations used trajectories created by subject matter experts**
- **These trajectories did not represent the variability expected during actual operations**
- **Data collected during a recent human-in-the-loop experiment provided an opportunity to increase the level of realism of the fast-time simulation by incorporating a greater variety of and more accurate trajectories**



Statistical Engineering

- **Development of a methodology to simulate more realistic aircraft trajectories required the use of statistical engineering**

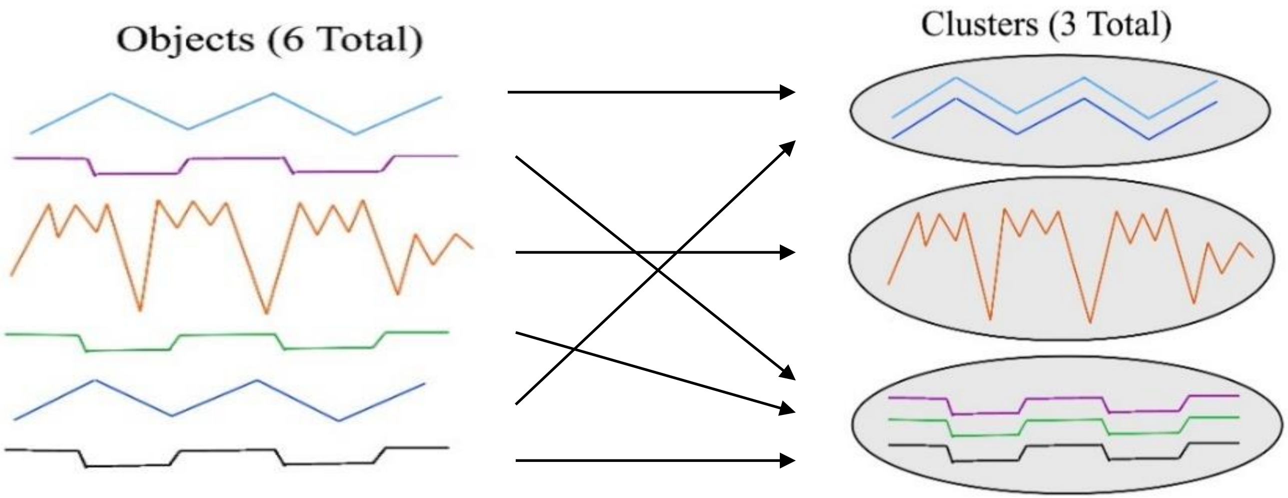
- **Goal of implementing a statistical engineering approach was to develop a methodology that could be utilized in future datasets and simulations**

- **Statisticians and aerospace engineers collaborated throughout the process to develop a methodological framework that is repeatable and scalable**

- **Cluster analysis was adapted and extended to identify patterns in recorded aircraft trajectories and results were incorporated into the fast-time simulation**

Cluster Analysis

- **Goal of cluster analysis is to group similar objects together to discover natural underlying behavior**
- **Clustering is an unsupervised learning approach, which means that the number of groups and group membership are not known *a priori***
- **Number of groups and how items are assigned to each group is based on how one defines similarity with regards to the data items**



Cluster Analysis

- Cluster analysis involves two steps:

  1) Quantify the sameness between objects using a similarity measure

  2) Place objects that are considered similar into groups using a clustering algorithm

- Similarity measure is a function that quantifies the degree of uniformity or sameness between a pair of objects

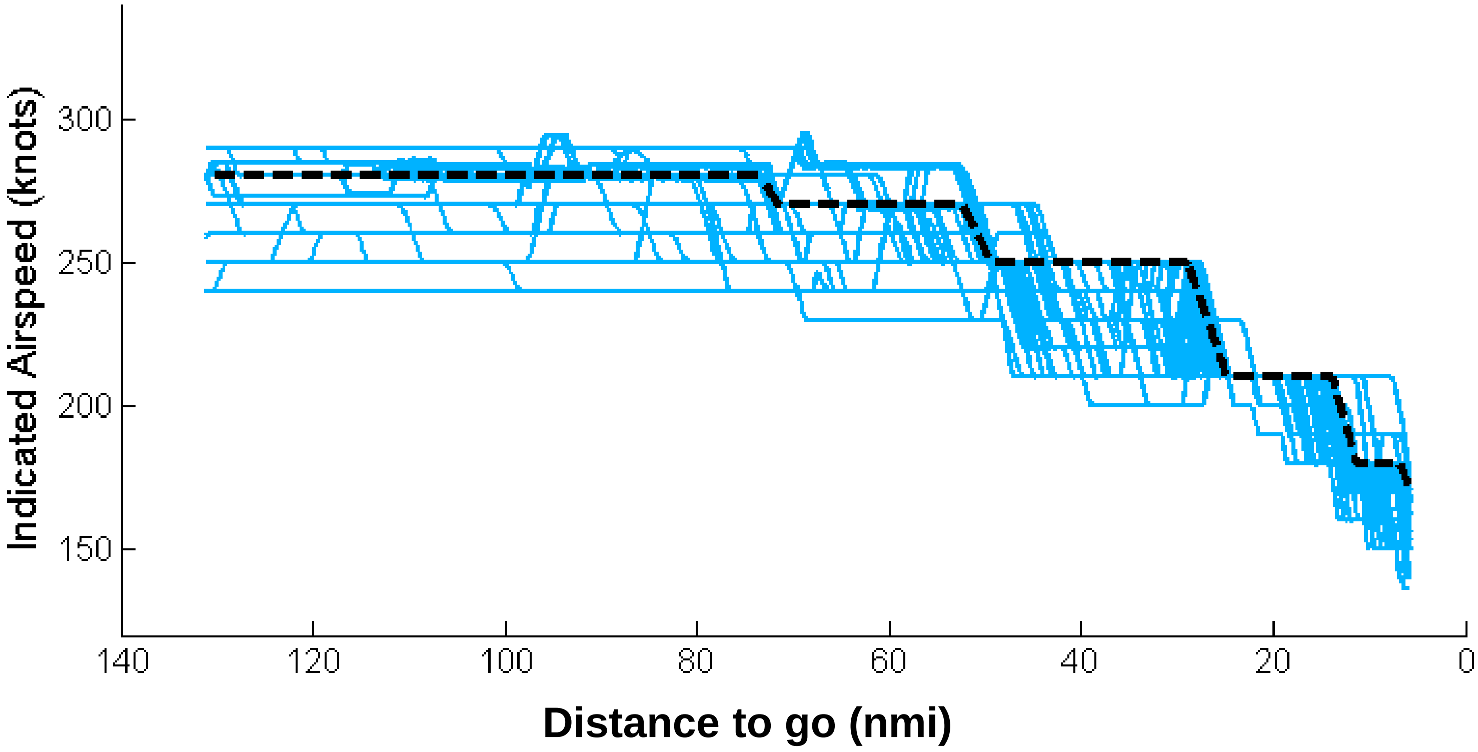
- Clustering algorithm is a procedure capable of placing similar objects together into groups

Challenges

- Formulating an acceptable similarity measure for aircraft trajectories and what constituted an acceptable cluster is challenging
- Would like the sequences of data points that discretize the trajectories to represent the paths of the original trajectories, but in practice these sequences are less then ideal
- Identifying similarities between trajectories is challenging in the presence of:
  - Missing data points and interruptions in trajectories
  - Trajectories of different lengths
  - Trajectories with different starting locations
- Unequal numbers of observations per trajectory add complexity

Challenges

- **Large amount of variability makes it challenging to visually distinguish between underlying groups and identify outliers**



Trajectory Clustering Methodology

- **The trajectory clustering methodology developed incorporates**

  1) **Dynamic Time Warping to form similarity measures**

  2) **k-Means algorithm applied to the similarity measures to determine the clusters**

- **The resulting clusters are used to identify patterns in aircraft trajectory data**

Dynamic Time Warping

- Similarity measure is a distance function that quantifies the degree of sameness between two objects

- Dynamic Time Warping was used because it provides a similarity measure that is
  - Definable for sequences of differing lengths
  - Robust to short interruptions
  - Insensitive to high variability in our trajectory dataset

Dynamic Time Warping

1. **For trajectories $Q = \{q_i\}_{i=1}^{n}$ and $S = \{s_j\}_{j=1}^{m}$, generate the matrix D with $n \times m$ entries $d(q_i, s_j)$.**

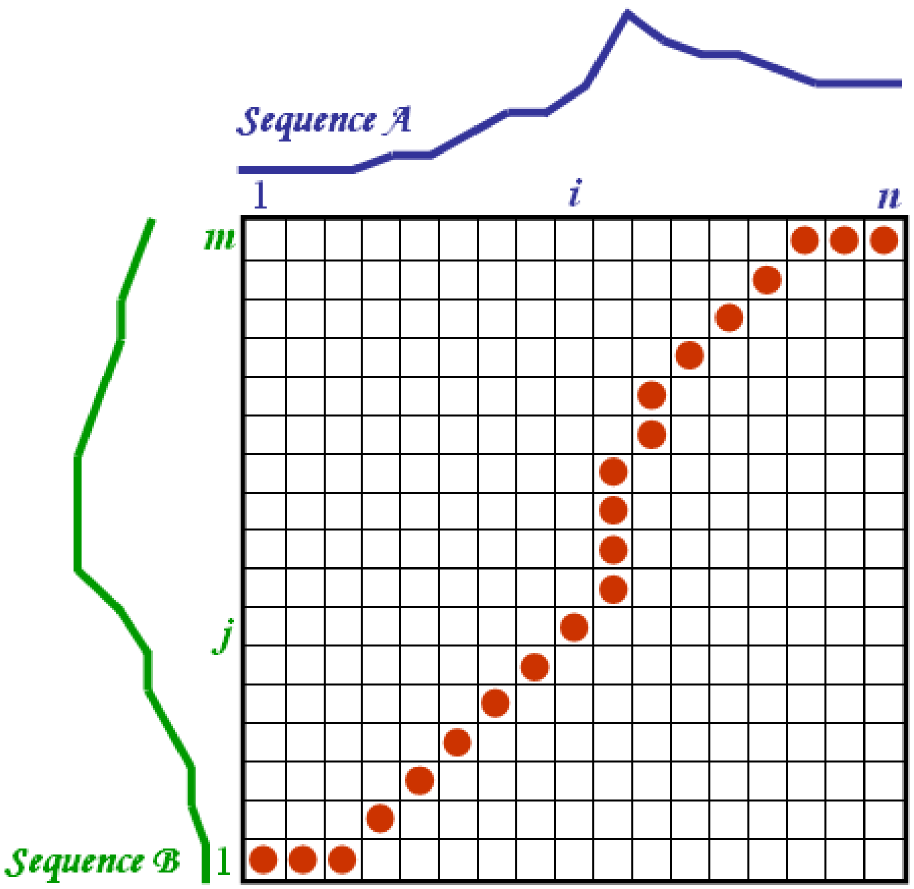
2. **Calculate the entries of the cumulative distance matrix Γ.**

   a. **Set $\gamma(1,1) \equiv d(q_1, s_1)$ and $\gamma(0,0) = \gamma(0, j) = \gamma(i, 0) \equiv 0$.**

   b. **Calculate the remaining entries of Γ using the elements of D and $\gamma(i,j) = d(q_i, s_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}$**

3. **Retain $\gamma(n, m)$ as the optimal solution to $DTW(Q, S) = min \sum_{p=1}^{P} w_{(i,j),p}$ where $W = \{w_{(i,j),p}\}_{p=1}^{P}$ is a warping path through the matrix D. This is the Dynamic Time Warping similarity measure.**

Dynamic Time Warping



Three requirements of Dynamic Time Warping:

1. **Respective endpoints have to map to each other**

2. **Every point has to be mapped to some other point on the other sequence**

3. **Mappings are not allowed to cross each other**

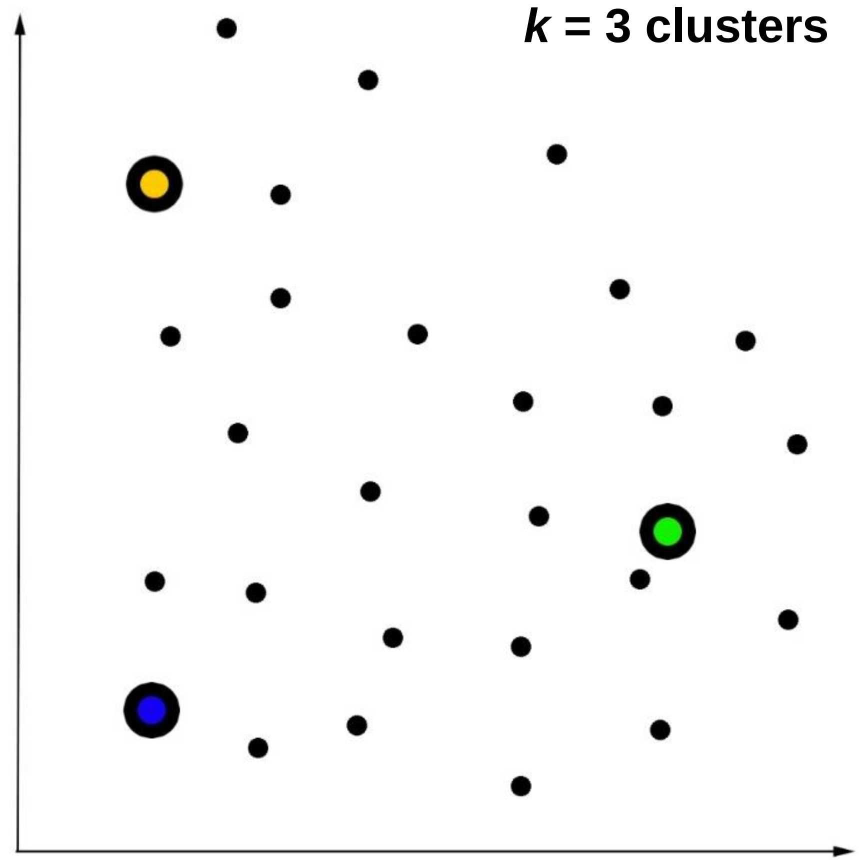
k-Means Clustering Algorithm

<u>Step 1:</u> Randomly choose *k* points from the dataset as the initial centroids.

<u>Step 2:</u> For each centroid, a cluster is formed by the set of objects that are closer to that centroid than to any other centroid.

<u>Step 3:</u> Calculate the within cluster means.

<u>Step 4:</u> Repeat Steps 2 and 3 until the process converges.



k-Means Clustering Algorithm

- **k-Means algorithm finds a user-specified *k* number of clusters for a given set of objects**
- **Determining a “best” choice of *k* can be determined through trial and error or using a systematic procedure such as the Gap Statistic**
- **For a range of possible values of *k*, the Gap Statistic calculates the expected number of clusters for each *k* using a comparison distribution and compares the differences of within cluster variation of these results to those when randomly sampling from the data**
- **A plot of these differences, or Gap Statistics, over the range of *k* can be visually inspected to help determine an appropriate choice**

Integration of Dynamic Time Warping and k-Means

- **Apply Dynamic Time Warping to the trajectory dataset to compute the similarity measures**

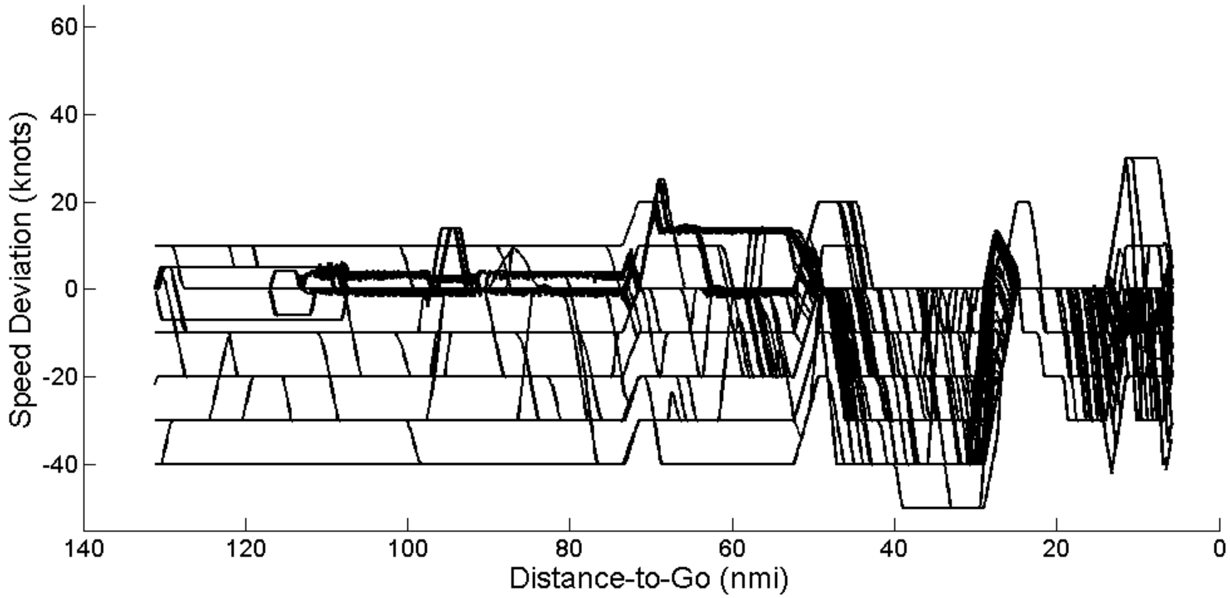
- **Select a value for *k* using the Gap Statistic applied to the new dataset of similarity measures**

- **Apply k-Means algorithm to the dataset of similarity measures**

- **The *k* clusters of Dynamic Time Warping similarity measures represent the clusters of trajectories**

Application of Trajectory Clustering

- **Dataset containing 164 trajectories collected during a human-in-the-loop experiment**
- **Lengths of trajectories ranged from approximately 600 to 1200 data points**
- **Patterns appear to exist but cannot visually determine which trajectories are similar**



Application of Trajectory Clustering

- **Interested in incorporating both distance and speed deviation so pairwise Euclidean distance was used to calculate the Dynamic Time Warping similarity measures**

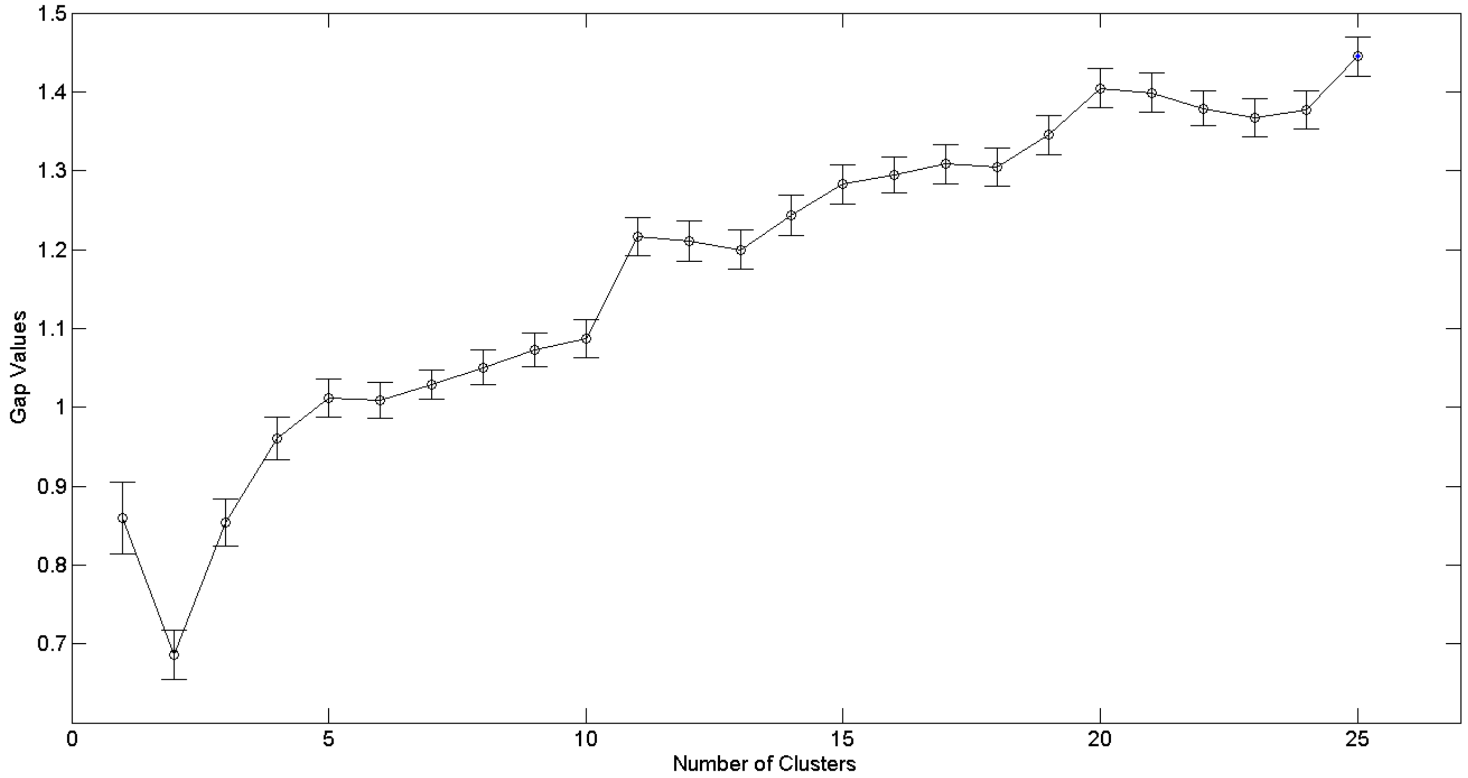
$$d\big((a,c),(b,d)\big) = \sqrt{(a-b)^2 + (c-d)^2}$$

- **Data were normalized to scale between zero and one to permit equal influence of each variable when calculating the similarity measures**

- **Calculations were carried out using MATLAB®**

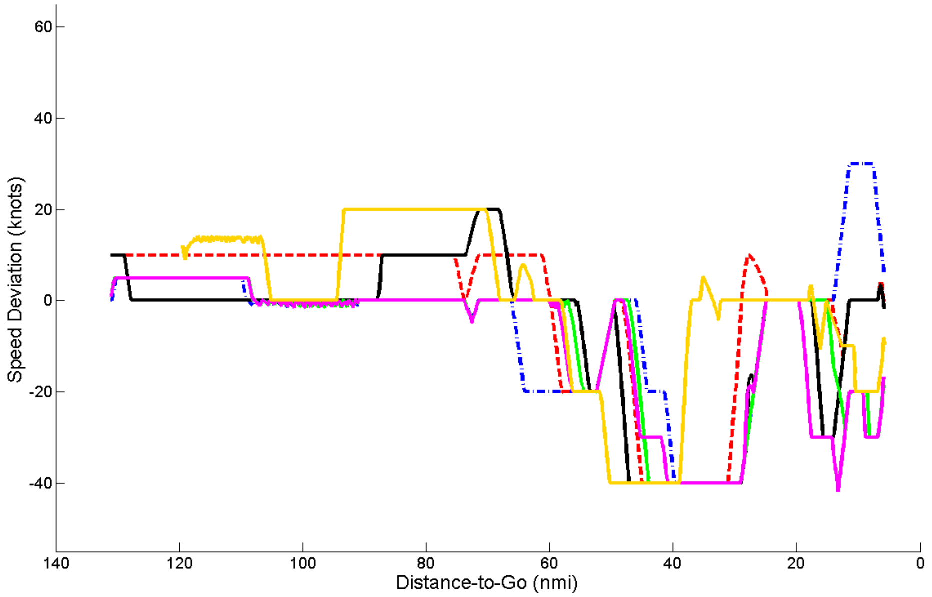
Application of Trajectory Clustering

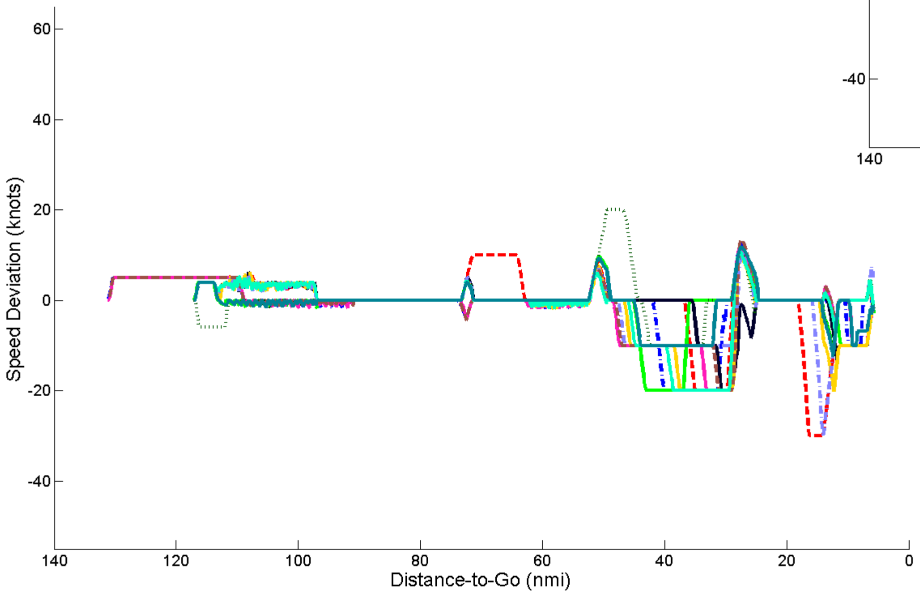
- Gap Statistic was implemented to determine an appropriate value of *k* for the k-Means algorithm
- Uniform comparison distribution was used
- *k* = 20 suggests a good starting place



Application of Trajectory Clustering

k-Means algorithm was applied with *k* = 20





Application of Trajectory Clustering

- **Within the general patterns of behavior, researchers were able to discern useful within cluster variability**
- **Of the 20 clusters, one contained a single trajectory and one contained two trajectories, indicating these as potential outliers**
- **Resulting clusters were determined by subject matter experts to identify realistic patterns in the data**
- **A sample of trajectories from each of the remaining 18 clusters could be selected and incorporated into the fast-time simulation to provide a more realistic variety of aircraft trajectories**

Conclusions

- **Successfully developed and implemented a methodology to uncover patterns within trajectory datasets**
- **Dynamic Time Warping similarity measure performed well despite challenges encountered**
- **Achieved a deeper understanding of the underlying behavior for aircraft trajectories**
- **Increased the level of realism and therefore the knowledge gained from the fast-time computer simulation**
- **Statistical engineering approach was a collaborative effort within a multi-disciplinary team**
- **Methodology developed can be applied in future simulations to investigate other air traffic management technologies or adapted for other types of problems involving trajectories**

References

- Arthur, D., and S. Vassilvitskii. 2007. K-means++: The Advantages of Careful Seeding. SODA '07: Proc. Of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms: 1027-1035.
- Gariel, M. A.N. Srivastava, and E. Feron. 2011. Trajectory clustering and an application to airspace monitoring. *IEEE Transactions on Intelligent Transportation Systems* 12 (4): 1511-1524.
- Sankoff, D., and J. B. Kruskal. 1983. Time warp, string edits, and macromolecules. Reading, MA: Addison-Wesley.
- Tibshirani, R., G. Walther, and T. Hastie. 2001. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2):411-423.