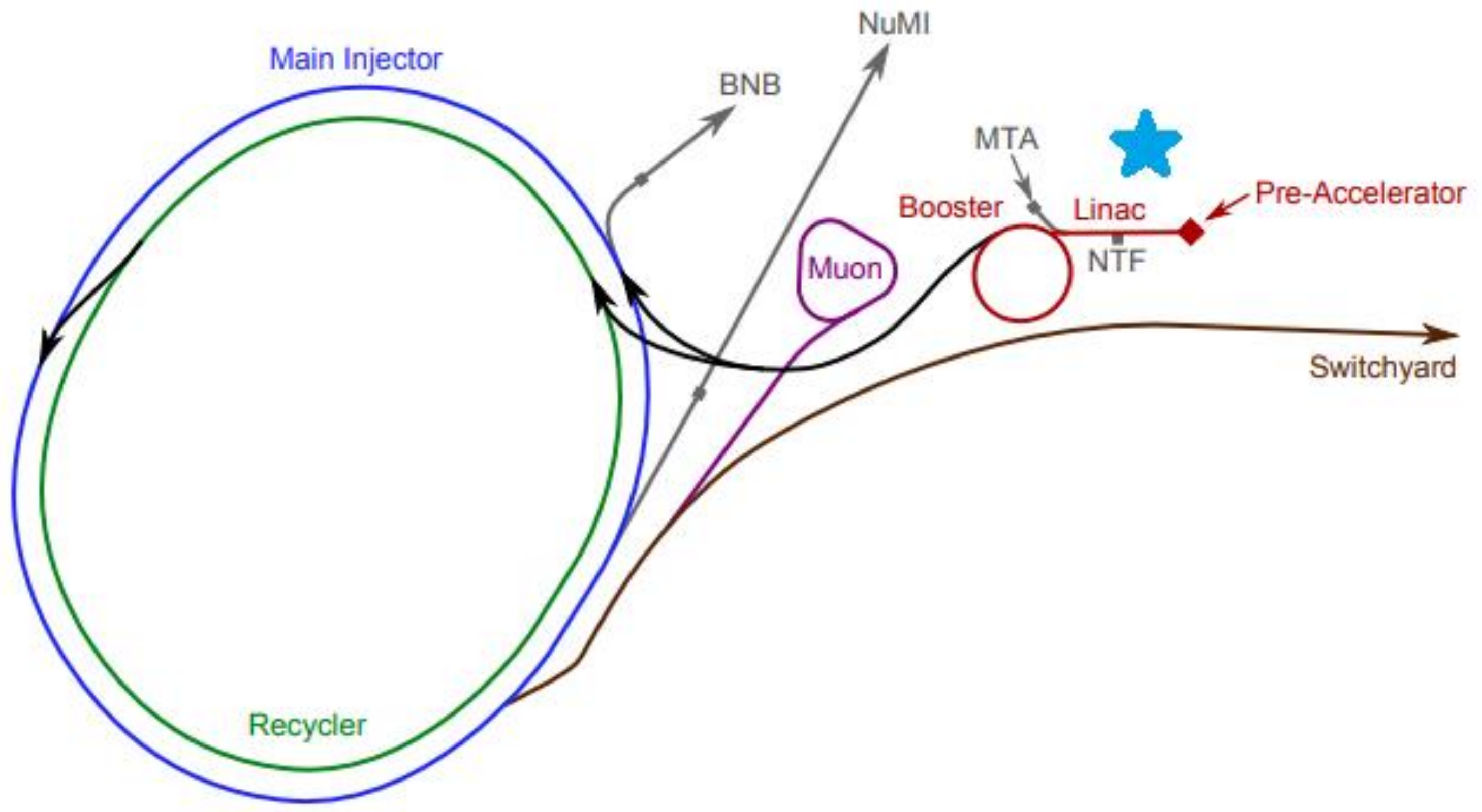
FERMILAB-POSTER-22-176-AD-STUDENT

Study of clustering methods of Linac outages: K-Means vs G.M.M

Aleksandar Dyulgerov, Harper College (CCI) – Supervisors: Jason St. John and Brian Schupbach, Fermilab

Motivation

Fermilab strives to operate its accelerator complex without interruption, and unplanned outages of the Linac (Linear Accelerator) impose down time on the entire complex which it feeds. The Linac accelerates H- ions to more than 70% the speed of light, at 400 MeV kinetic energy. This is achieved with a carefully tuned series of RF (Radio Frequency) accelerating cavities. For the RF stations and support systems, the Linac presents about 2800 control system devices, monitored and operated by the Main Control Room. Through Machine Learning (ML), data streams from these devices can be used to foresee and automatically mitigate unplanned Linac outages, maximizing up time and conserving lab resources.



(Figure 1) Linac (Top right, blue star) within the Fermilab accelerator complex

Methods

Recorded Linac-fault data from 24 selected Linac devices in 10-second outage intervals underwent dimensional reduction by UMAP (Uniform Manifold Approximation and Projection), obtaining a 2D plane with a data point from each outage interval. This plane from UMAP enables cluster analysis by fault type, and defining the boundaries of clusters allows prediction of fault type for future faults. With extensive use of Pandas and Scikit-learn, K-Means and the Gaussian Mixture Model were both used in order to come up with their respective predicted labels and clusters of the outages. In order to measure the performance of the clustering, homogeneity and completeness metric scores were assed on a 0 (worst case) to 1 (best case) scale.

Timeline

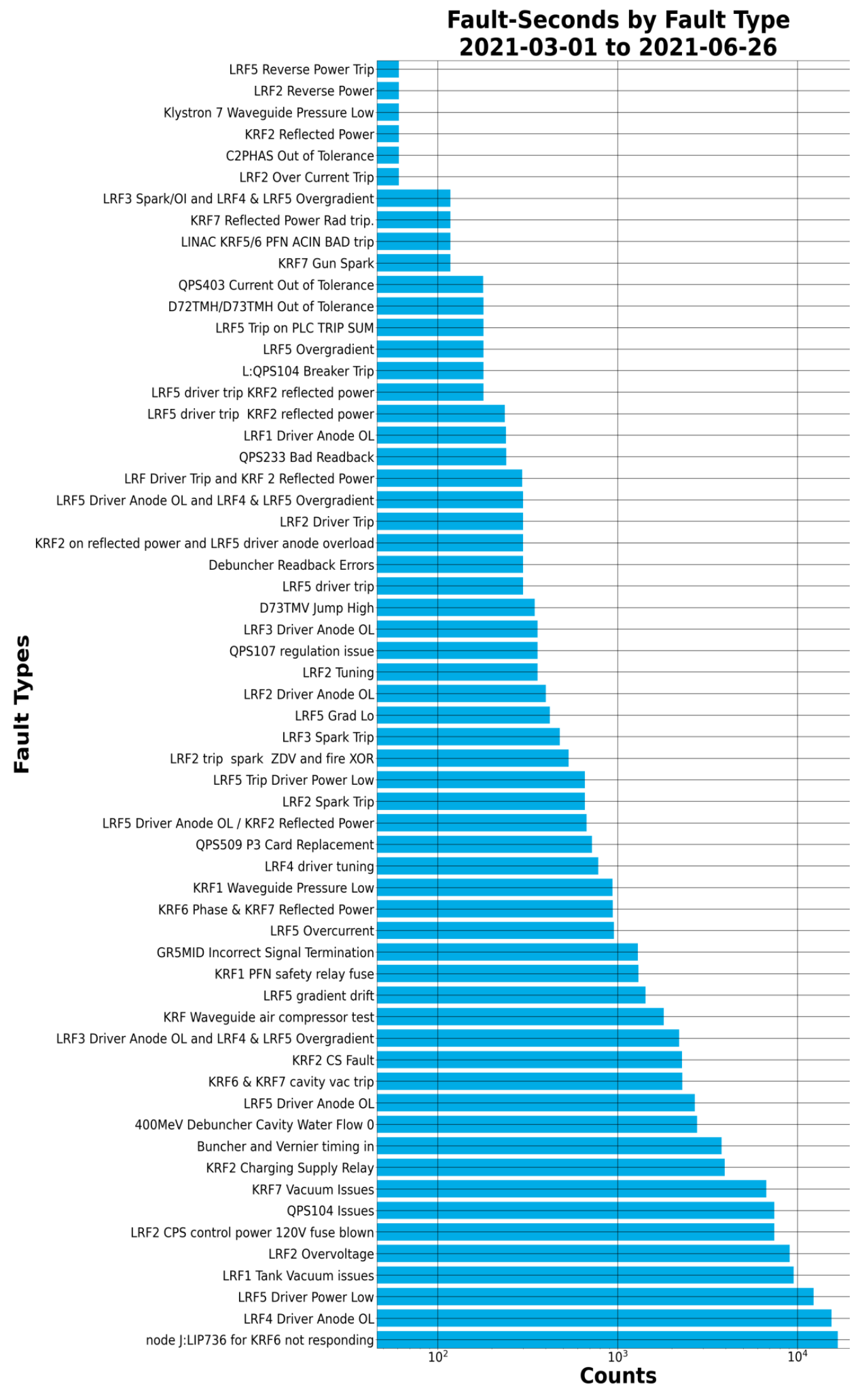
Initially – Explored Linac-fault data from March-July, 2021 to better understand the recorded frequencies of each fault type, plotted clustering results, and performed clustering studies in the UMAP plane.

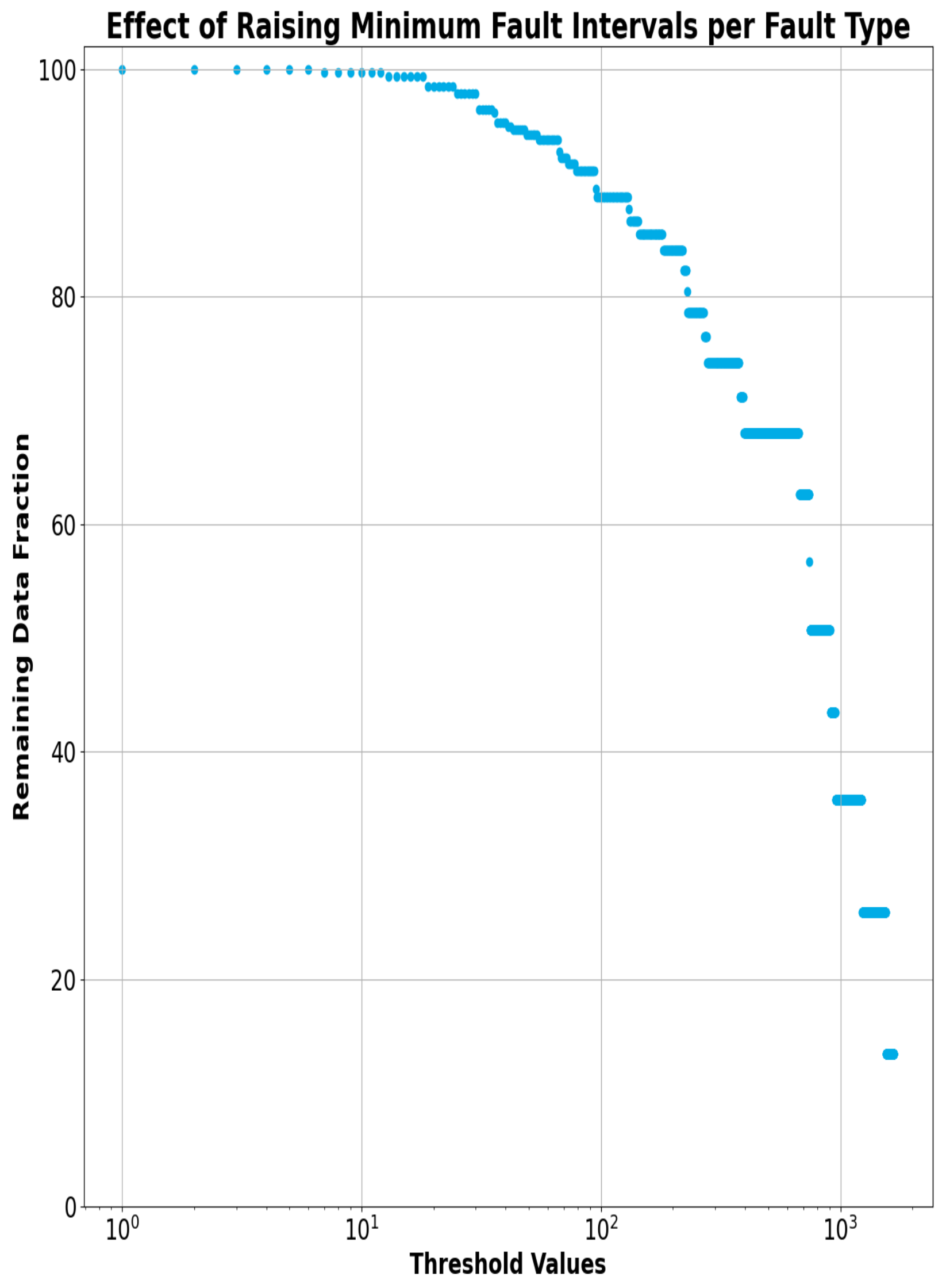
Afterwards – Duplicate fault types and planned downtime faults were removed from the Linac-fault data.

At the same time – Understood Linac-fault data to be fault-seconds of each fault type rather than instances of each fault.

Finally – Reproduced clustering results and obtained performance metric scores: homogeneity and completeness. [See the last section]

Linac-fault data





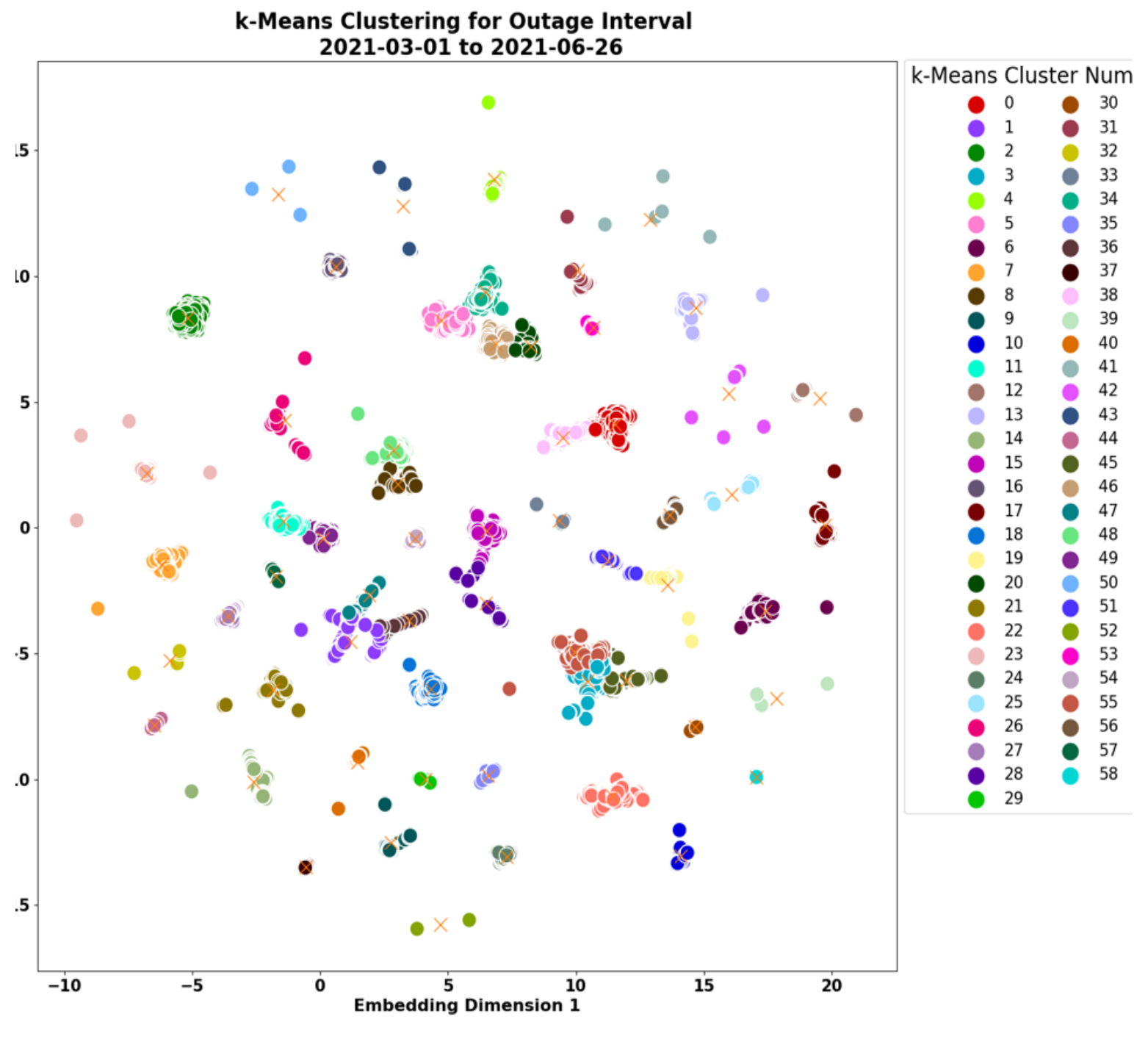
(Figure 2) Fault-10 Second Interval by Fault Type and Count

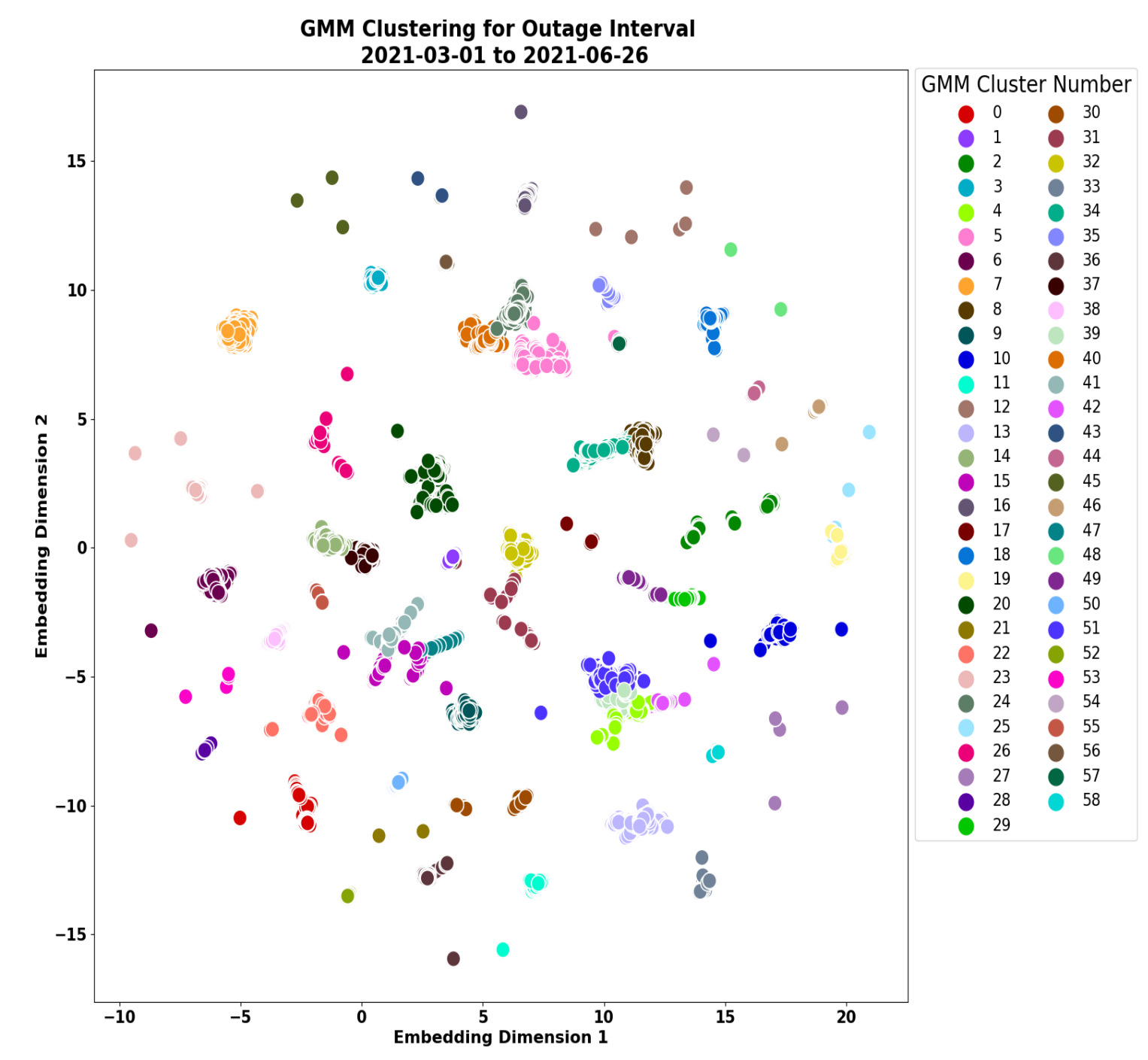
(Figure 3) Remaining data fraction vs minimum fault-intervals (10-second intervals) per fault type

Improved data cleaning, eliminated planned beam outages, and merged some fault types for unplanned beam outages, removing 55% of data volume. Investigated cutting away the rarest fault types to improve statistical robustness. Fig. 2 (right) shows that requiring at least 300 10-second intervals per fault still leaves about 80% of total fault-seconds.

Clustering

Unsupervised K-Means and GMM algorithms each learned to label points on the 2D UMAP plane by their location alone. Label count was matched to ground-truth label count, and random seeds set for reproducibility. K-Means partitions the plane into non-overlapping areas by nearest cluster center, placed iteratively. Gaussian Mixture Method instead uses overlapping 2D Gaussians.

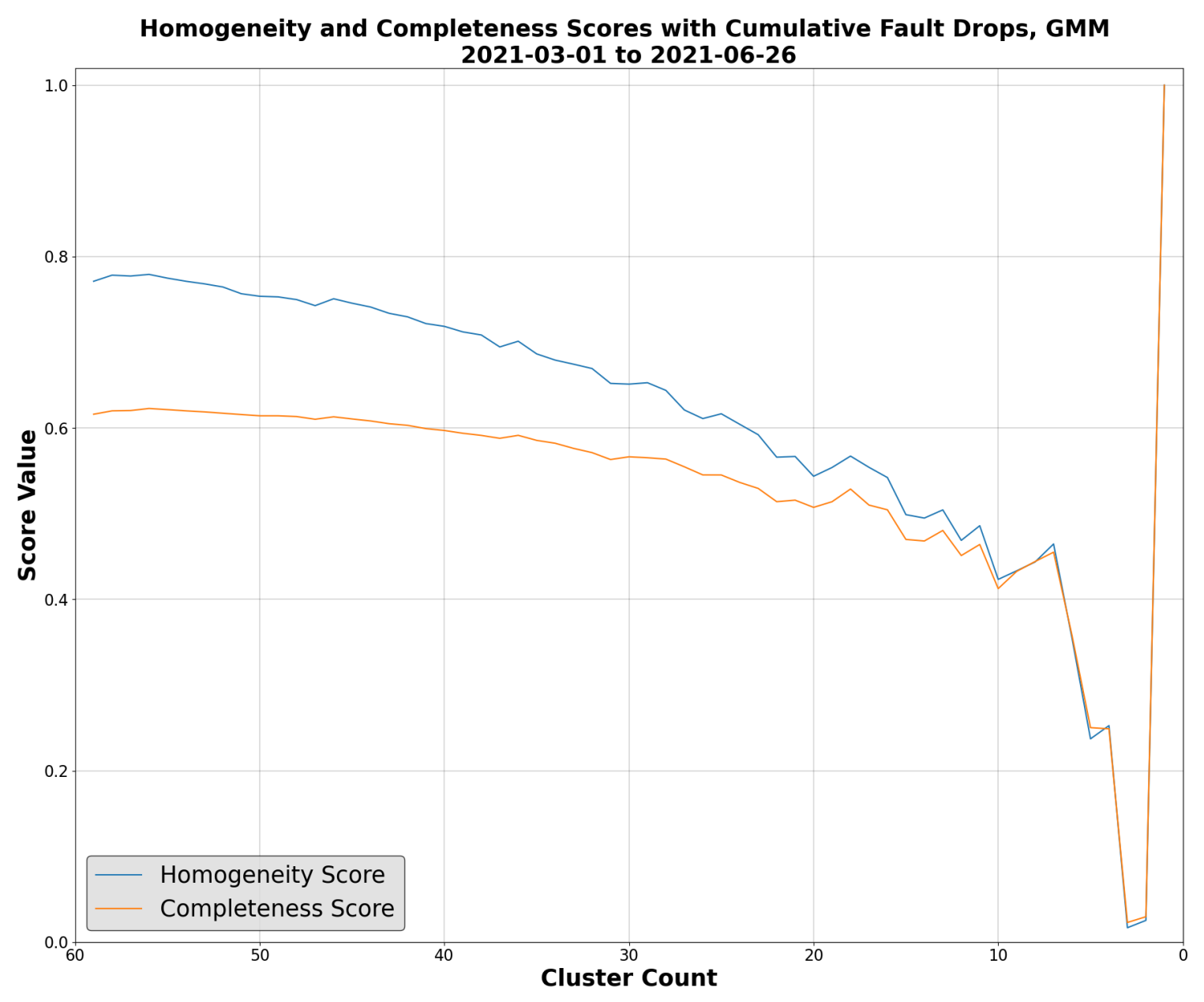


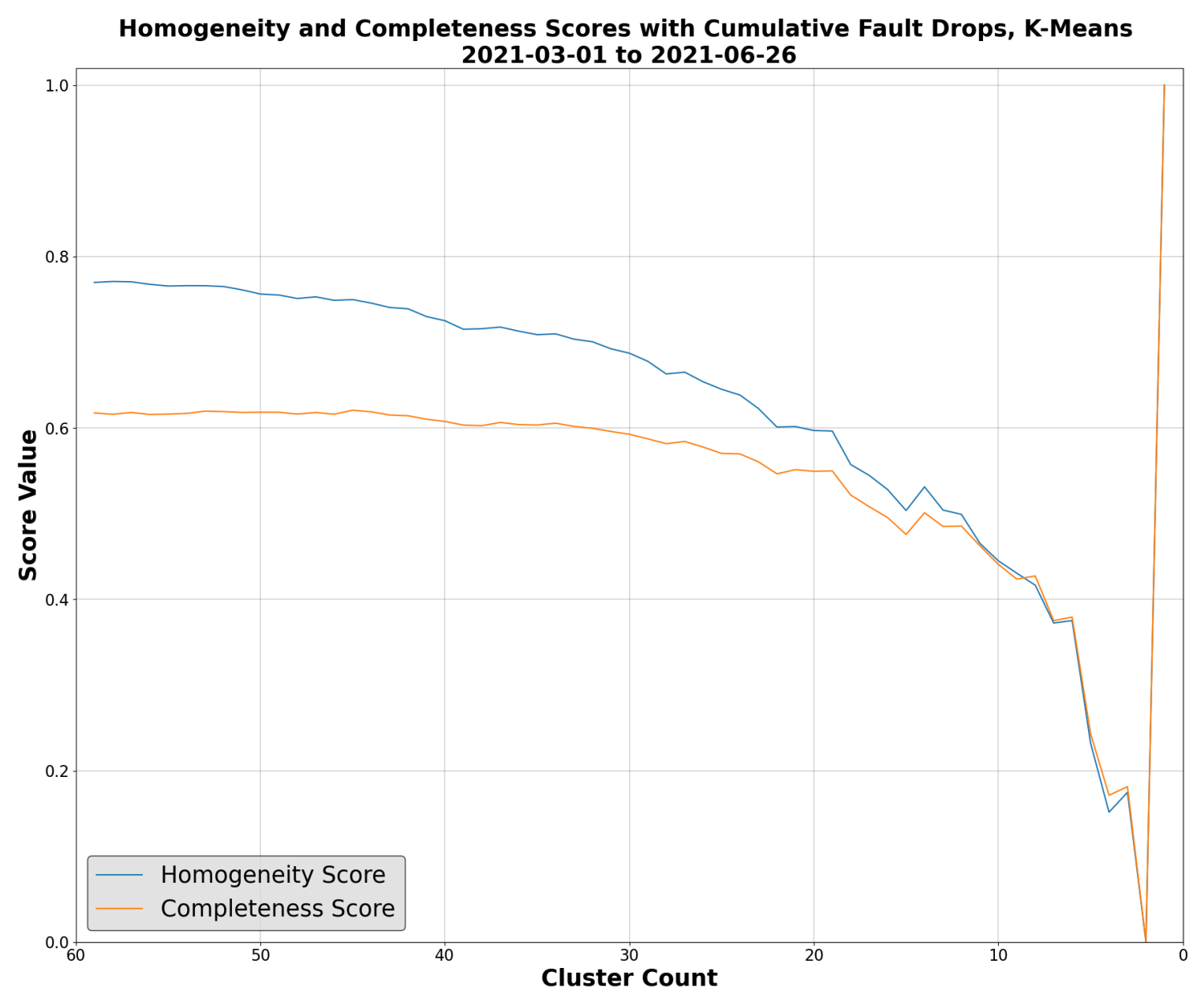


(Figure 4) K-Means Clustering for Outage Interval

(Figure 5) GMM Clustering for Outage Interval

Metrics





(Figure 6) GMM homogeneity and completeness scores vs Cluster count

(Figure 7) K-Means homogeneity and completeness scores vs Cluster count

Homogeneity shows what amount of predicted clusters contain only members of a single class, and completeness measures if all members of a class are assigned to the same cluster. Both metrics use a scale of 0 to 1 where a higher score is better. Figs. 6 and 7 were produced by iteratively dropping the rarest fault types and re-clustering. Results showed K-Means and GMM initially behaved the same, but the 23 to 1 cluster count range shows superior performance in K-Mean's homogeneity and completeness metrics.

Fermi National Accelerator Laboratory

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.