

# Predictive Risk Modelling in Aviation Incident

## The 10<sup>th</sup> Montreal Industry Problem Solving Workshop

Virtual workshop organized by the CRM and IVADO, August 13-27, 2020



# Thanks to the team!

**Participants:** Joanna Chen, Behrouz Ehsani, Prakash Gawas, Mike Lindstrom, Mandana Mazaheri, Najmeh Nekooghadi, Ahmed Sid-Ali, Aida Vahdani.

**IATA:** Hyuntae Jung, Andrea Mulone, Yuval Yakubov.

**IVADO Advisor:** Guillaume Poirier.

**Coordinator:** Denis Larocque (HEC Montréal).

# Brief Recap of the IATA's challenge

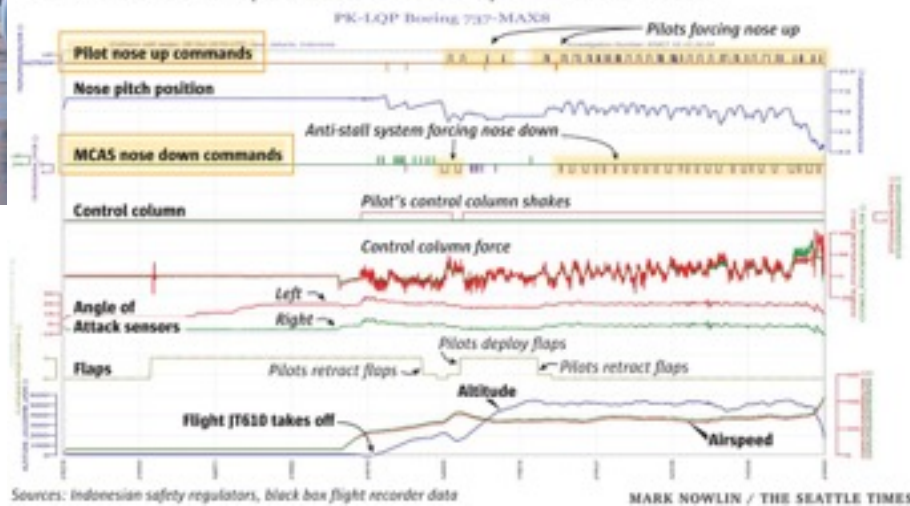
Predictive Risk Modelling in Aviation Incidents

# Background



## The jet's nose is repeatedly pushed down

The new anti-stall system on the Boeing 737 MAX forced the nose of Lion Air JT610 down 26 times in 10 minutes before the pilots lost control and the plane dived into the sea.



Currently, the global aviation safety risk identification is mainly reactive.

**“We don’t know what can be the problem until we face the problem.”**

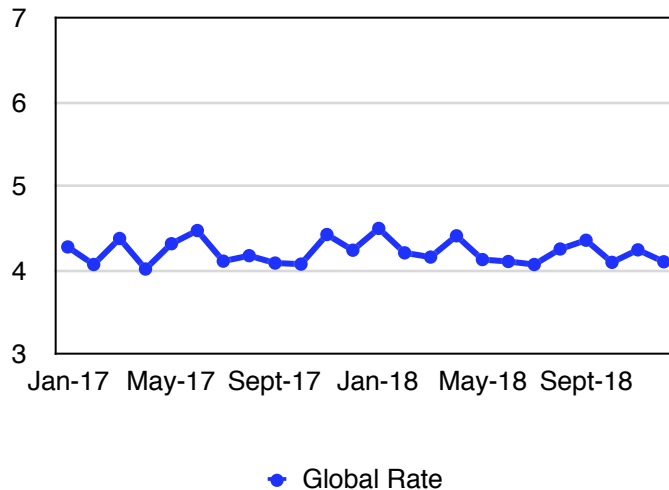
To proactively identify the potential risk areas before it evolves to an accident, we need to look at the data that may have “hints” about where to focus.

In a global scale, manually collecting, processing and analyzing these datasets are unsustainable. We need automation support on continuously monitoring the risk area.

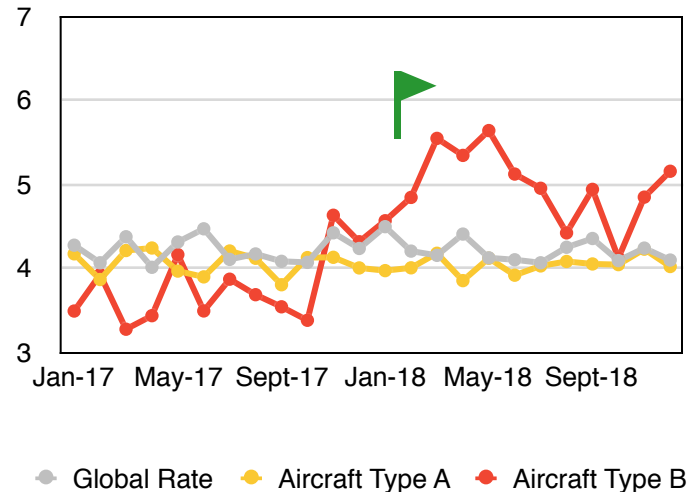
# IPSW Challenge Target – (1) Anomaly Detection

**Model to give hints to safety analysts where to look before querying every criteria one-by-one.**

Example: Global Aggregated Hard Landing Rate



Example: Hard Landing Rate by Aircraft Type



The model examines the set of incidents by drilling down into specific aircraft type, finding:

- Aircraft Type A does not show significant difference to the global rate
- Aircraft Type B shows a spike over the global rate (green flag), which may indicate prominent safety risk.

Once the model automatically identify such “anomaly” with statistical evidence, the flag will be raised, so that human safety analysts can perform deeper investigation.

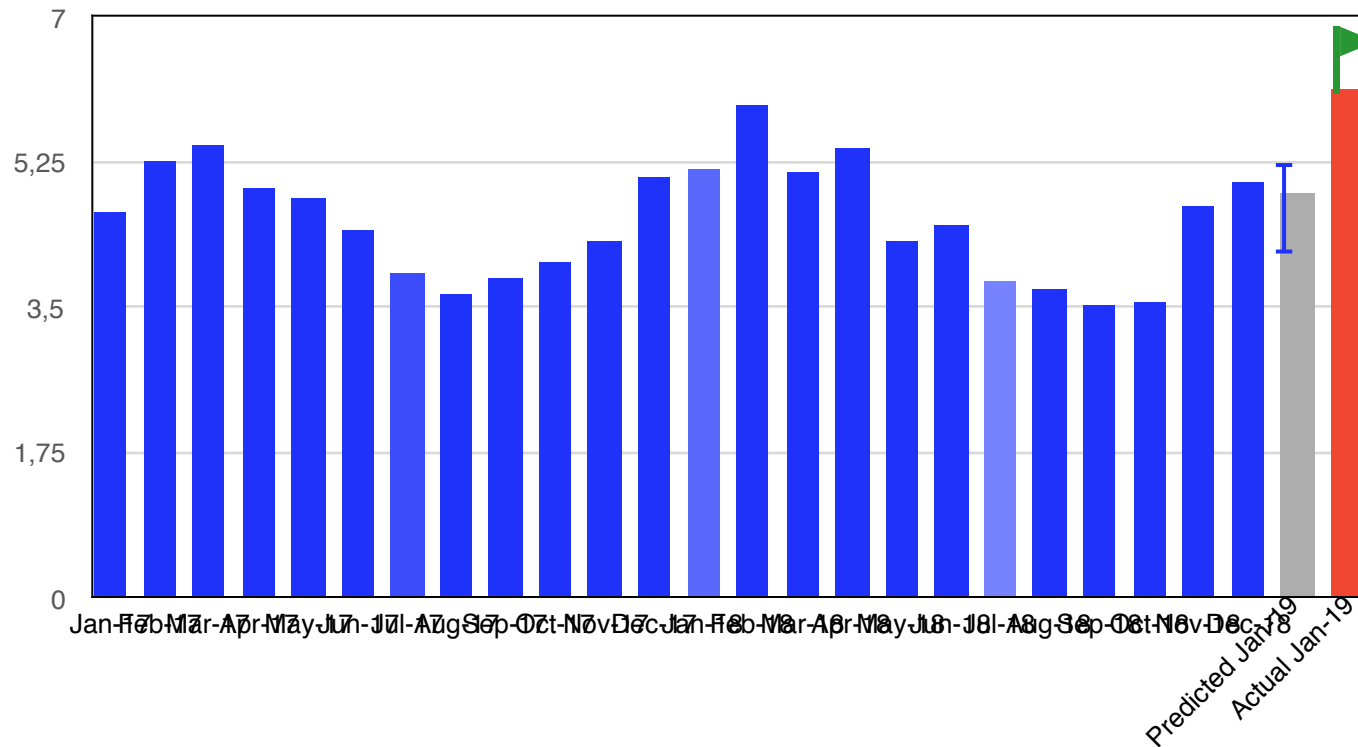
## Slicing the data with combinations of multiple attributes

- Events  
(e.g. Hard Landing, Engine Overheat, etc.)
- Date of Occurrence  
(e.g. Seasonal factors)
- Geographic  
(Region, Country, Airport Level)
- Aircraft Type
- Flight Phase

# IPSW Challenge Target – (2) Predictive Analysis

**Model to predict event rate based on historical records, and flag if the actual rate is exceptional.**

Example: Monthly rate of Event A (with the seasonal pattern)



In this example, the historical data for Event A shows seasonal pattern – higher rate in the winter season and lower rate in the summer season.

After trained by the 2 years of historical incident data, the model makes a prediction for January 2019, with given interval of confidence. However, the actual data from January 2019 was out of the boundary.

If the actual rate of the certain incident is out of the boundary, this may mean there might be significant change in the risk profile. The model will flag this finding, so that human safety analysts can perform deeper analysis.

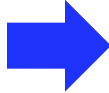
# Data – Incident Reports & Sector

## Incident Report (621K)

- **Report ID:** 7723515
- **Year:** 2018
- **Month:** May
- **Fleet Family:** AType5
- **Location:** Airport162
- **Location Country:** Country256
- **Phase:** Approach
- **Event:** Weather - Windshear

“What event happened”

normalize



## Sector Data

- **Quarter:** 2018 Q2
- **Fleet Family:** AType5
- **Departure:** Airport162
- **Departure Country:** Country256
- **Arrival:** Airport359
- **Arrival Country:** Country26
- **Sectors:** 3,631

“How much flights”



**Event Rate**

# Problem Solving

## Predictive Risk Modelling in Aviation Incidents



# Work accomplished during the workshop

1. Understand the available data and problems.
2. Identify potential ways to solve them.
3. Do a literature review to find state-of-the-art methods that could be used in practice.
4. Proofs of concepts, using IATA's data, for potentially useful and practical data analysis methods.

# Brainstorming and Literature Reviews

- During the brainstorming, each participant suggested a possible approach. Then the team reviewed various methodologies such as:
  - Vectorized representation for data and Logistic Regression
  - Neural Networks
  - Naive Bayes Classifiers
  - Functional KDE
  - Functional Isolation Forest
  - Time-series forecasting (e.g. *forecast* and *prophet* R package)
- The team decided to try multiple approaches in parallel according to their preferences and then present the results during daily meetings.

# Data Preparation

- Data aggregation was made to filter out by certain type of events (descriptors) with the given set of attributes.
  - First input to the script includes the most interesting event descriptors for the user.
  - Second input is the aggregation level desired by the user. The input given is a subset of all the attributes for each event.
  - Example : “([Windshear, Turbulence], [Fleet Location, Phase])”
  - The script then aggregates the count of the events descriptors occurring together for each of the aggregate level attributes.
  - This procedure can be directly called for use in any model that the user wants to implement, thus saving time in data preparation. It also allows for comparison between different values of the attribute (for example comparison between different fleet families).

# (1) Anomaly Detection

- Basically, the problem consist of findings anomalous patterns in time-series curves. Clustering methods for curves can be used.
- The team used two models to detect anomalies in the curves:
  - (1) **Simple and Discrete Fourier Transform KDE Anomaly Detection**
  - (2) **Hierarchical Curve Clustering** with the dtwclust R package.

# Functional KDE Anomaly Detection

Think of an anomaly as being distant from the rest of the data.

If the data come from some distribution, anomalies should have correspondingly small probability densities.

Using our data, a collection of time series, we want to ascribe a score to represent these densities so that comparatively low scores represent anomalies.

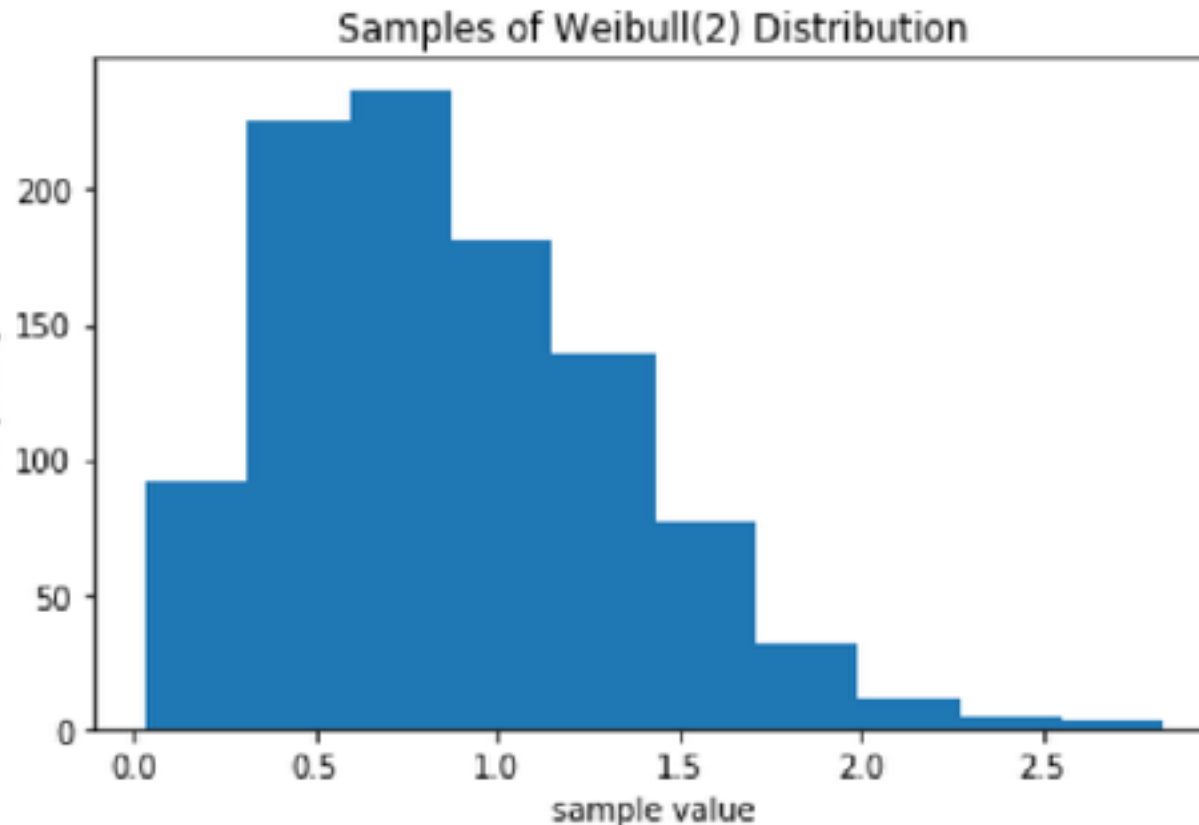
But there is a problem: we don't know the distribution...

# Kernel Density Estimation (KDE)

**Kernel Density Estimation (KDE)** uses sums of Gaussian kernels to infer empirical, continuous probability distributions for data.

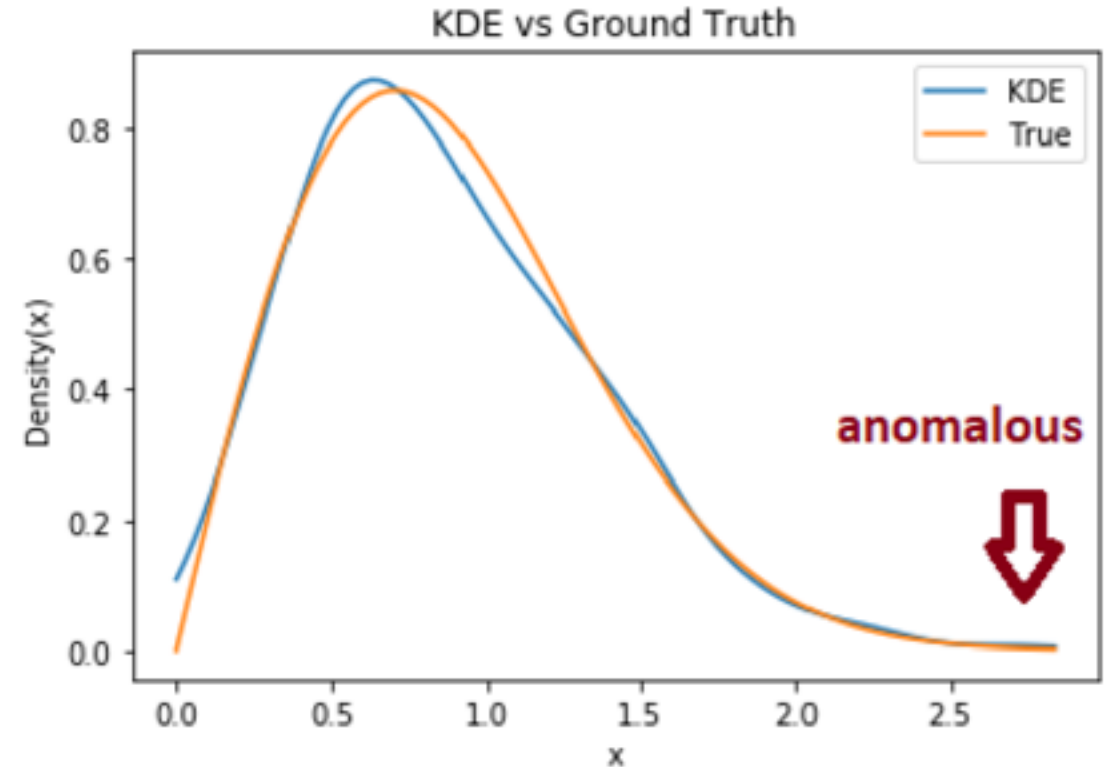
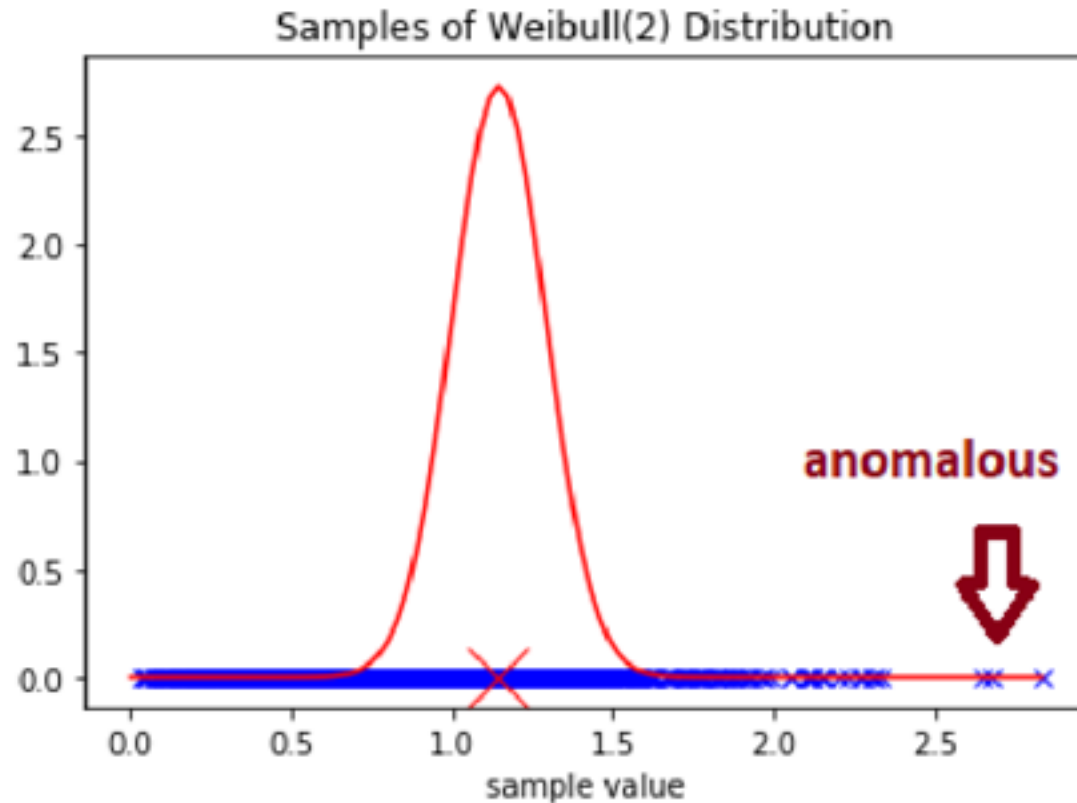
Consider discrete samples of a Weibull distribution with pdf

$$f(x) = kx^{k-1}e^{-x^k} \text{ for } k = 2.$$



# Kernel Density Estimation (KDE)

If at each point, we place a Gaussian kernel...



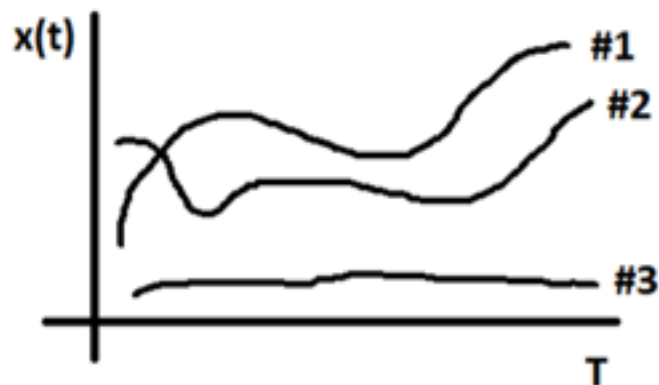
Then the sum of all such kernels gives an estimate for the true pdf with lower pdf values indicating anomalies.

# Simple Functional KDE

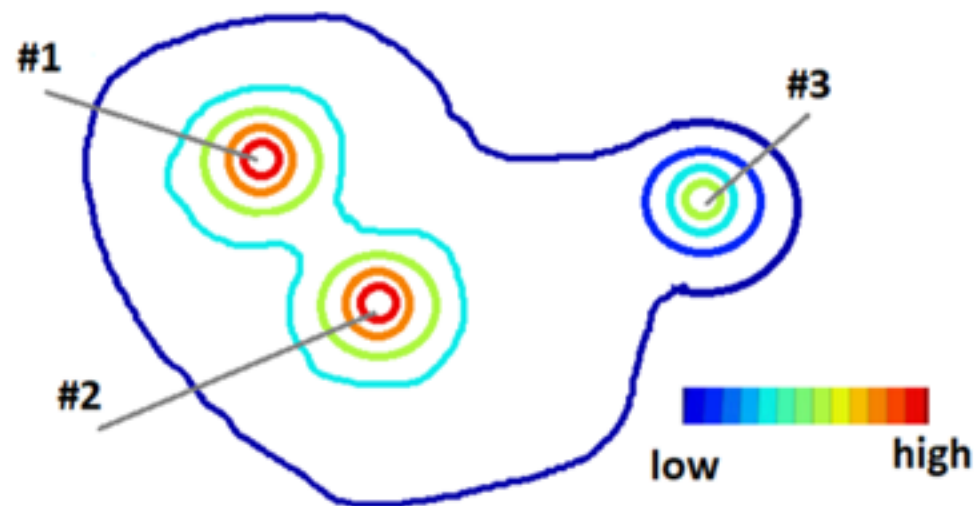
We can think of our time series as samples of signals  $x: [0, T] \rightarrow \mathbb{R}$  or as being in the Hilbert space,  $\mathcal{H}$ , say  $L^2([0, T])$  or  $H^1([0, T])$ .

Hilbert spaces have induced norms,  $\|\cdot\|$ , which can be thought of as generalized distances.

**Idea:** place a Gaussian kernel over  $\mathcal{H}$  at each time series  $x_i(t)$  and construct a probability density functional.



Time Series



Hilbert Space Density Visualization



# Simple Functional KDE

We can formally define an empirical pdf over  $\mathcal{H}$ .

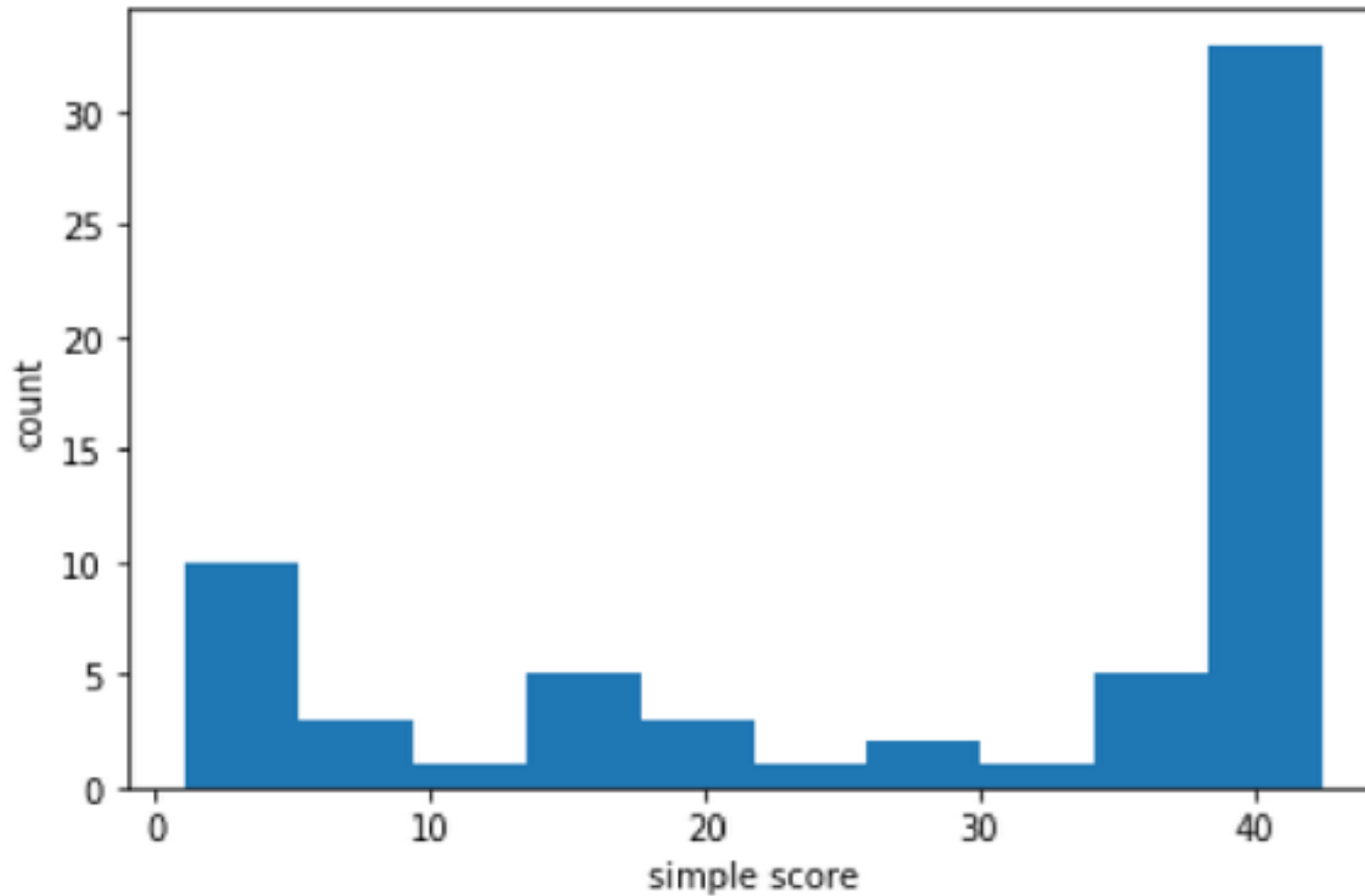
## Simple Functional KDE Method:

- Begin with a multiset of curves  $S = \{x_j(t)\}_{j=1}^N$  where  $x_j \in \mathcal{H}$  for  $j = 1, 2, \dots, N$ .
- Choose  $\sigma > 0$  a hyper-parameter.
- Define the probability density functional

$$\rho[a] = \sum_{x \in S} e^{-\frac{1}{2\sigma^2}\|x-a\|^2}$$

- Assign to each  $x_j$  a score  $s_j = \rho[X_j]$ .
- Identify anomalies by a histogram of  $\{s_j\}_{j=1}^N$ .

# Simple Functional KDE



For **High-Energy/Unstable Approach**, scores  $\lesssim 10$  are anomalous.

# Discrete Fourier Transform Functional KDE

The spaces  $L^2([0, T])$ ,  $H^1([0, T])$  have countable bases  $\{e^{2\pi i n/T}\}_{n \in \mathbb{Z}}$

Fix  $N$  and suppose  $x_j(t) \approx \sum_{n=-N}^N \hat{x}_n^{(j)} e^{2\pi i n/T}$ .

Suppose that each  $\hat{x}_n \sim \varepsilon_n$  for some pdf  $\varepsilon_n$  with corresponding density over  $\mathbb{C}$  of  $\zeta_n(z)$ .

Then to each curve  $x_j$ , we can ascribe a pdf value in  $\mathbb{R}^{2N+1}$  with

$$f(x_j) = \prod_{n=-N}^N \zeta_n(\hat{x}_n^{(j)}).$$

**In practice:** use Discrete Fourier Transform (DFT) since our signal is discrete and finite.

# Discrete Fourier Transform Functional KDE

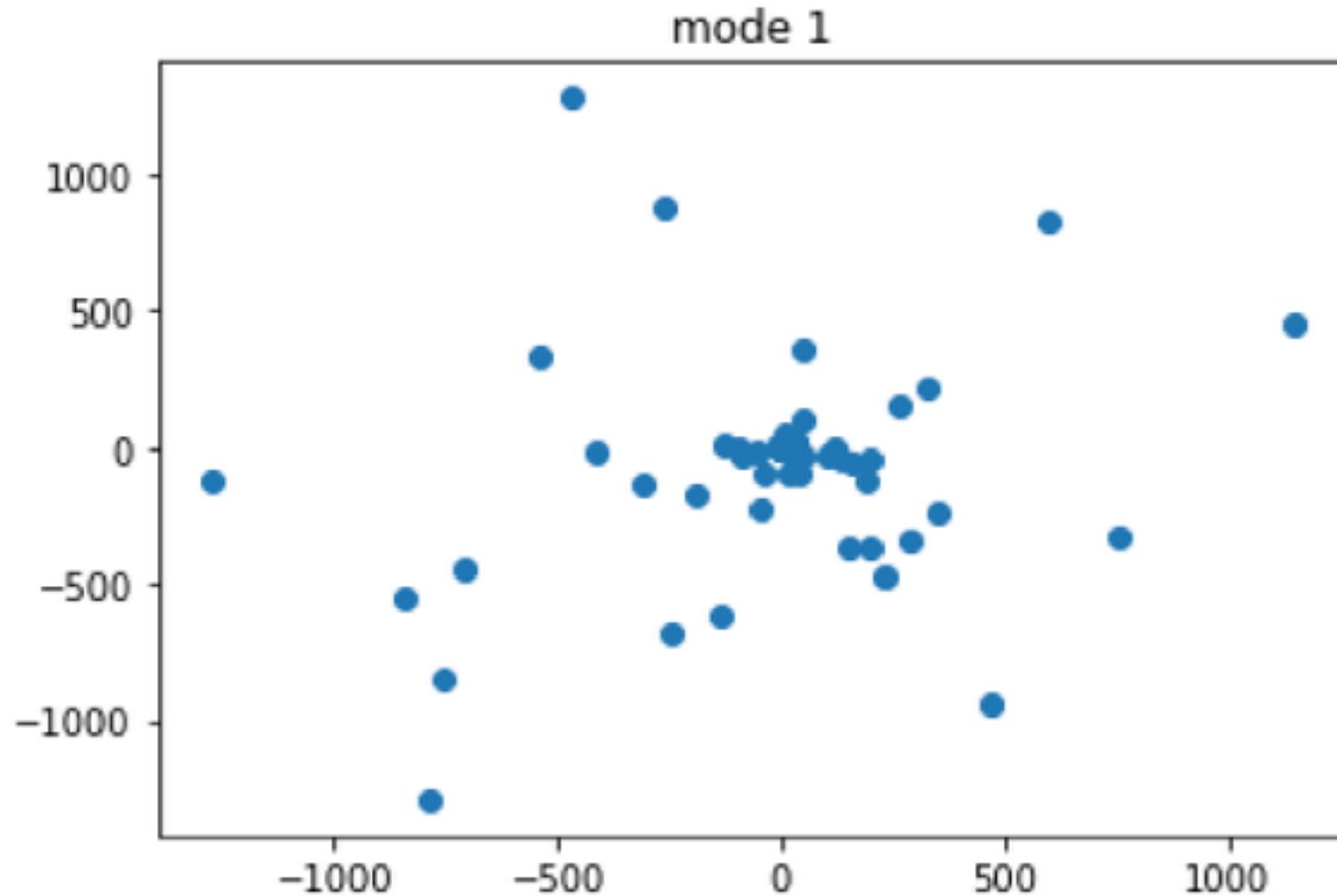
## DFT Functional KDE:

- Begin with a set of curves  $S = \{x_j(t)\}_{j=1}^N$  where  $x_j \in \mathcal{H}$  for  $j = 1, 2, \dots, N$ .
- Use a Discrete Fourier Transform to compute  $\{\hat{x}_n^{(j)} \mid j = 1, 2, \dots, N; n = 0, 1, \dots, N - 1\}$ .
- Use KDE to estimate pdf of  $\hat{x}_n$ , call it  $\zeta_n$  for  $n = 0, \dots, N - 1$ .
- Define the probability density at  $a \in \mathcal{H}$  as

$$\rho[a] = \prod_{n=-N}^N \zeta_n(\hat{x}_n^{(j)}).$$

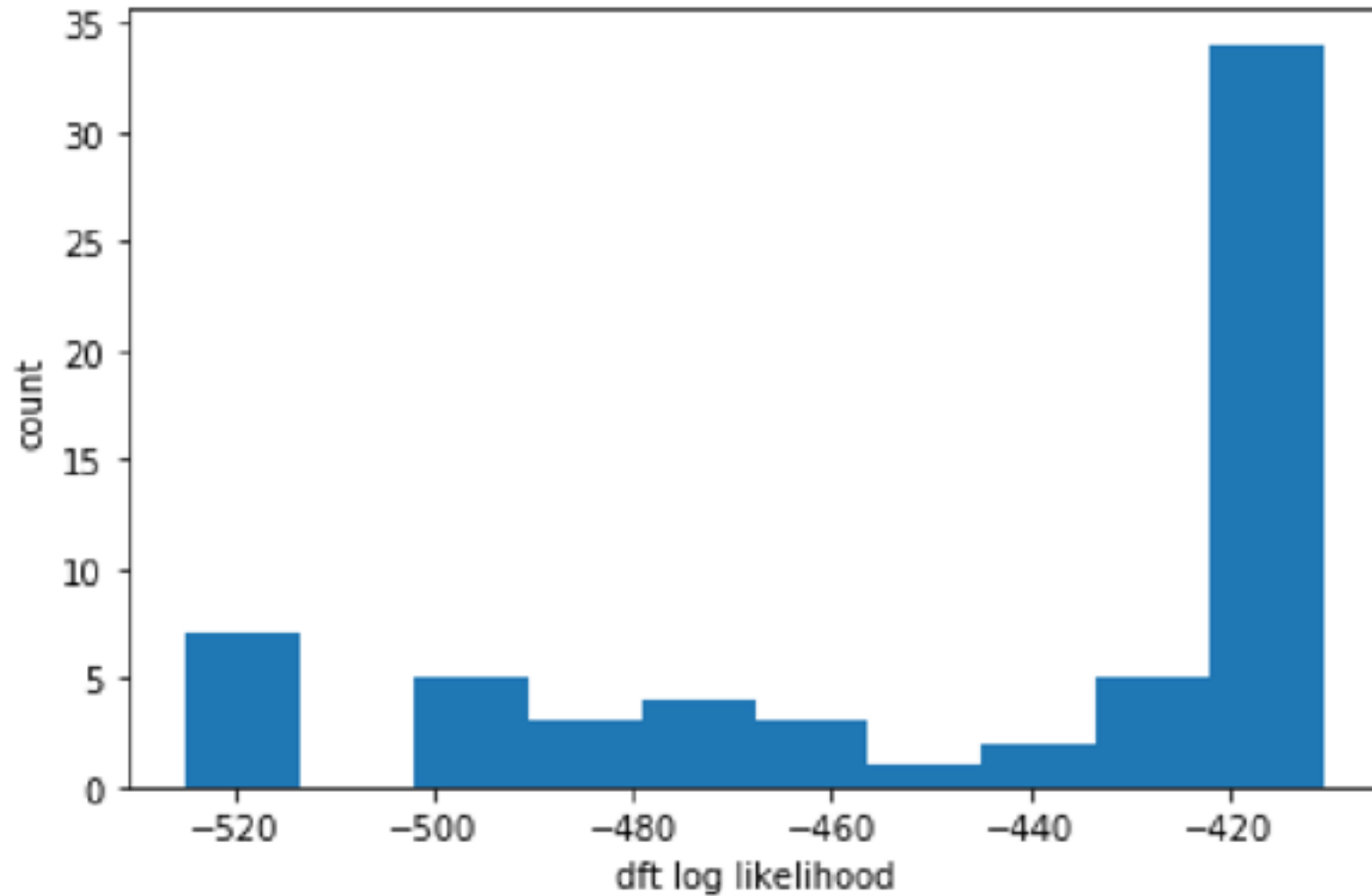
- Assign to each  $x_j$  a score  $s_j = \rho[x_j]$ .
- Identify anomalies by a histogram of  $\{s_j\}_{j=1}^N$ .

# Discrete Fourier Transform Functional KDE



Example of distribution of  $\hat{x}_1$  values. KDE is done upon this in each Fourier mode.

# Discrete Fourier Transform Functional KDE



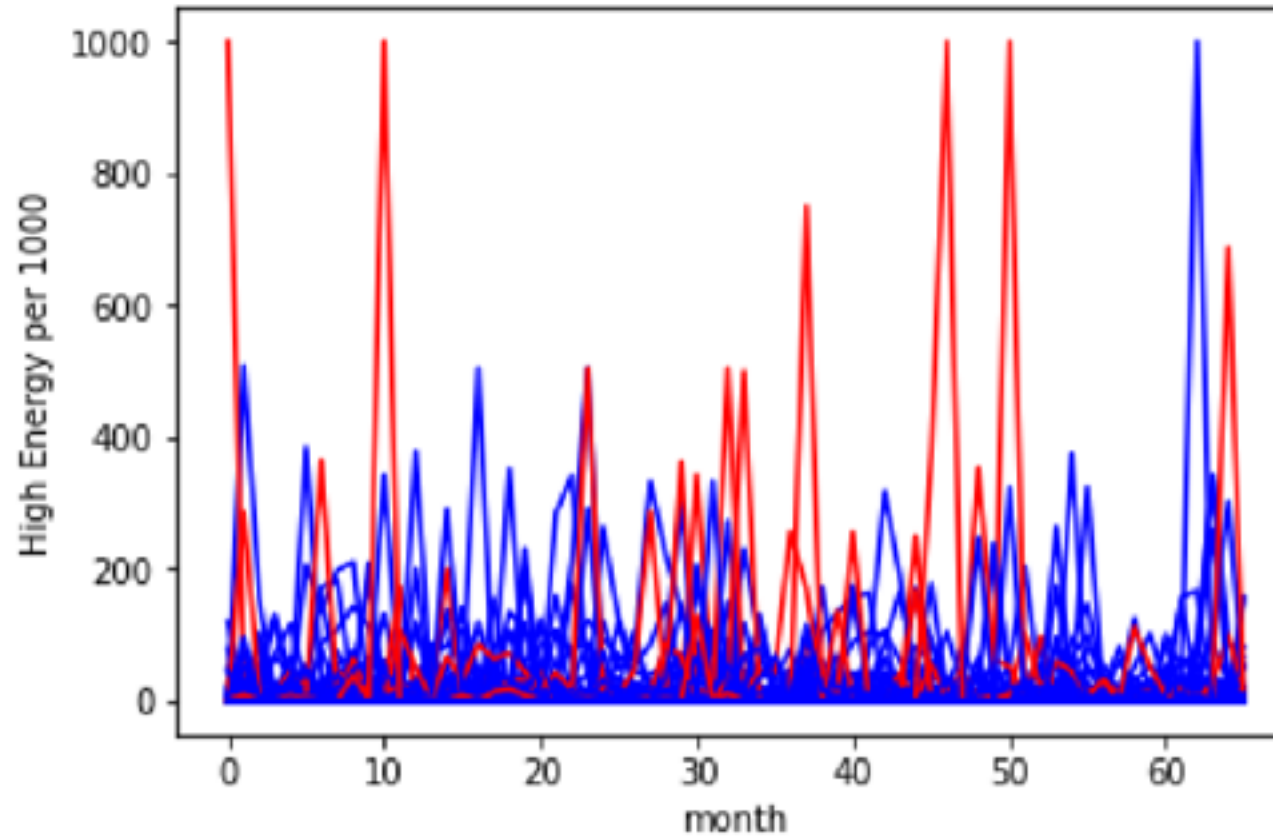
For **High-Energy/Unstable Approach**, scores  $\lesssim -510$  are anomalous.  
**Separation is much clearer.**

# Simple vs DFT Comparisons

| Events                        | Simple  | DFT  |
|-------------------------------|---|--|
| Landing Gear System           | 6, <b>11</b> , 12, 13, 14, <b>23</b> , <b>25</b> , 29, <b>33</b> , <b>48</b> , <b>52</b>              | <b>11</b> , <b>23</b> , <b>25</b> , <b>33</b> , <b>48</b> , <b>52</b>            |
| High Energy/Unstable Approach | 11, 12, <b>13</b> , 14, 16, 18, <b>19</b> , 20, 22, 23, <b>30</b> , <b>36</b> , <b>52</b> , <b>57</b> | <b>13</b> , <b>19</b> , <b>30</b> , <b>36</b> , <b>52</b> , <b>57</b>            |
| Windshear                     | <b>8</b> , 9, <b>12</b> , 13, <b>14</b> , 16, <b>20</b> , 21, 22, <b>26</b> , <b>30</b> , <b>51</b>   | <b>8</b> , <b>12</b> , <b>14</b> , <b>20</b> , <b>26</b> , <b>30</b> , <b>51</b> |

Anomalous flight IDs have significant overlap between the two methods. Everything the DFT method finds is also found in the Simple method.

# Anomalous Fleets

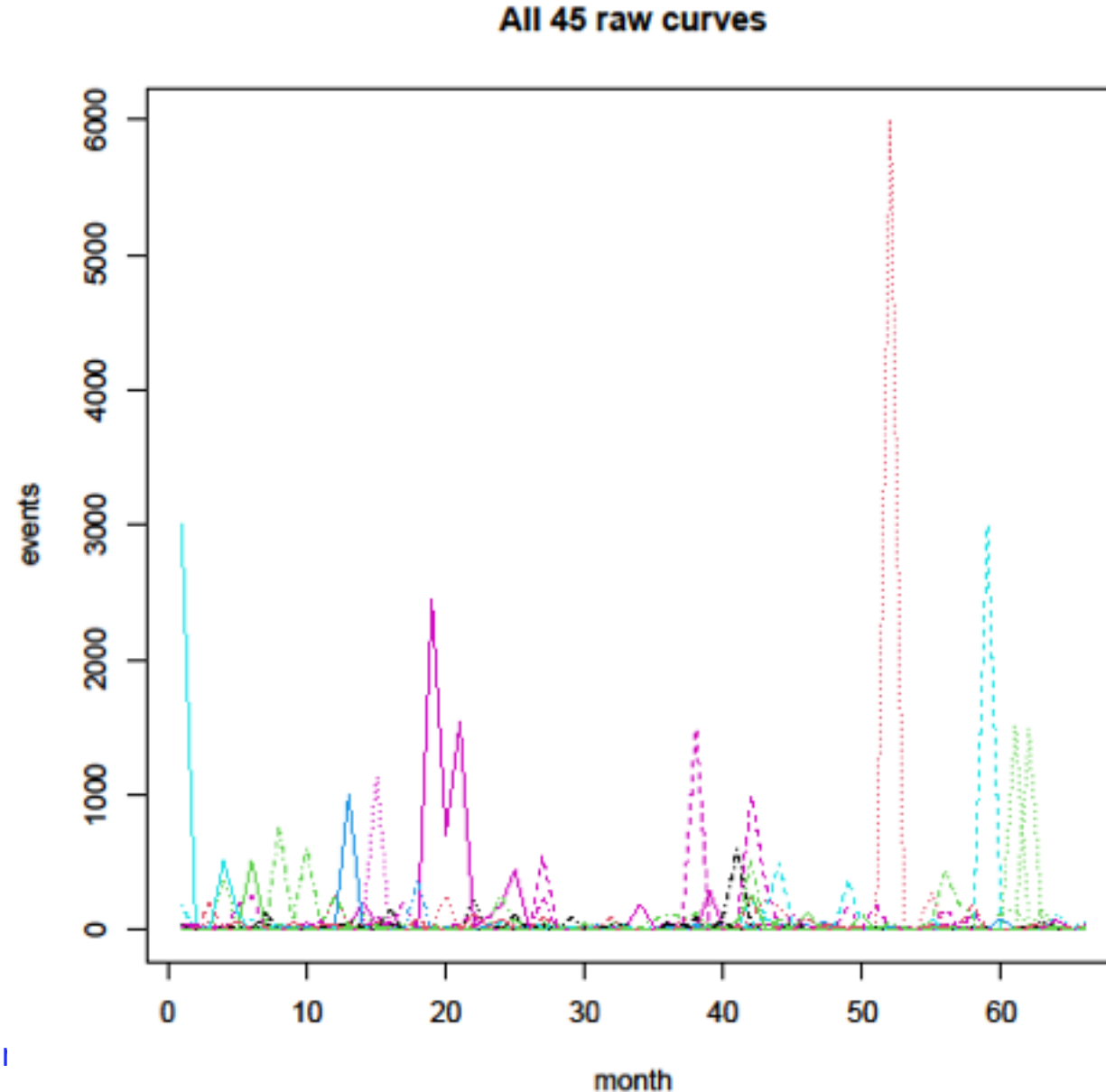


We plot the anomalous “**High Energy/Unstable Approach**” curves in red; the ordinary curves are in blue, based on the DFT classification.

Interpretation is an open question: identifying why a curve is anomalous.



# Anomaly Detection – Hierarchical Curve Clustering



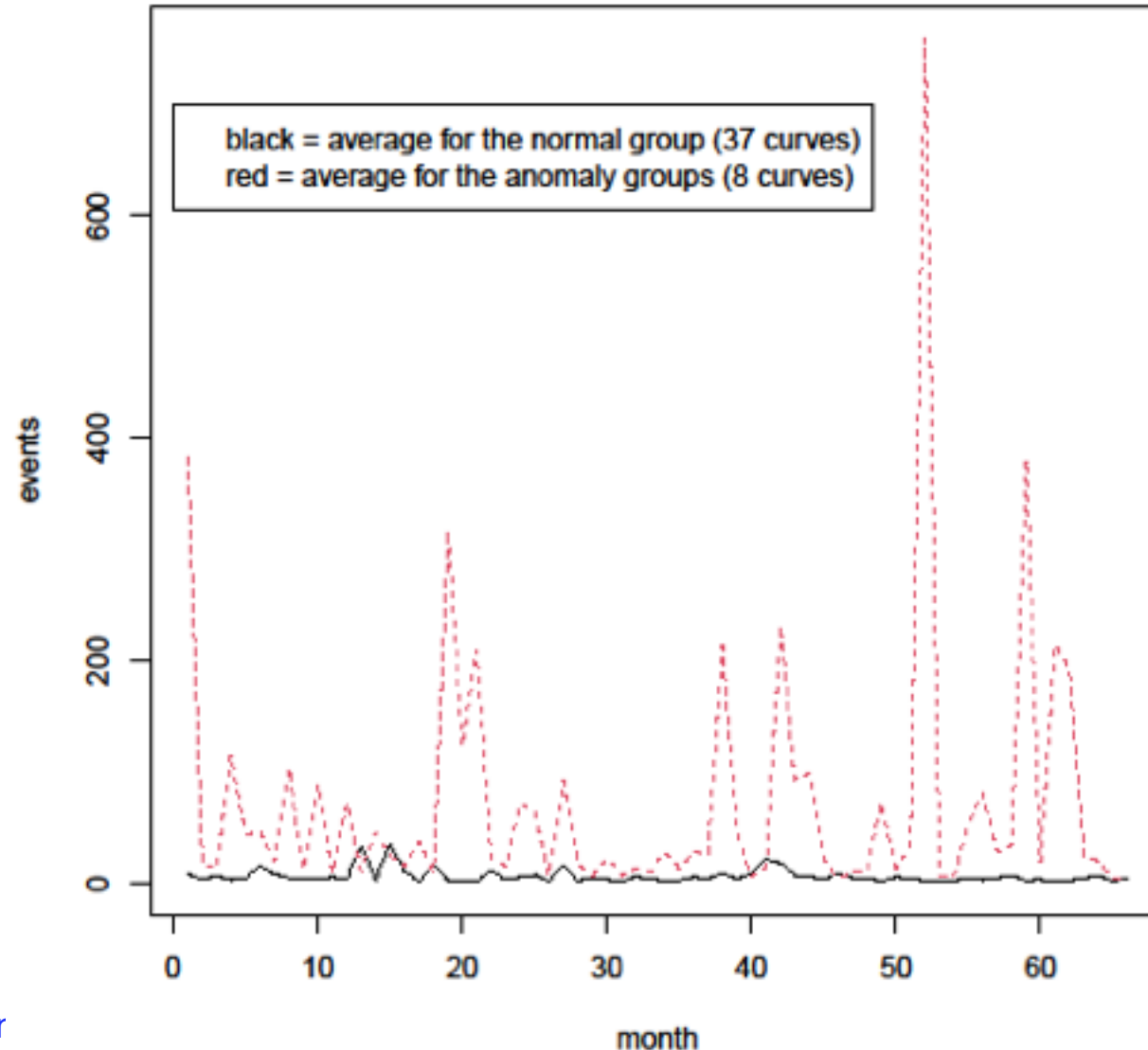
**Descriptor:** Windshear

Curve by **Fleet Family**

(2013 ~ 2018 Aggregated)

# Anomaly Detection – Hierarchical Curve Clustering

Example of curve clustering



**Descriptor:** Windshear

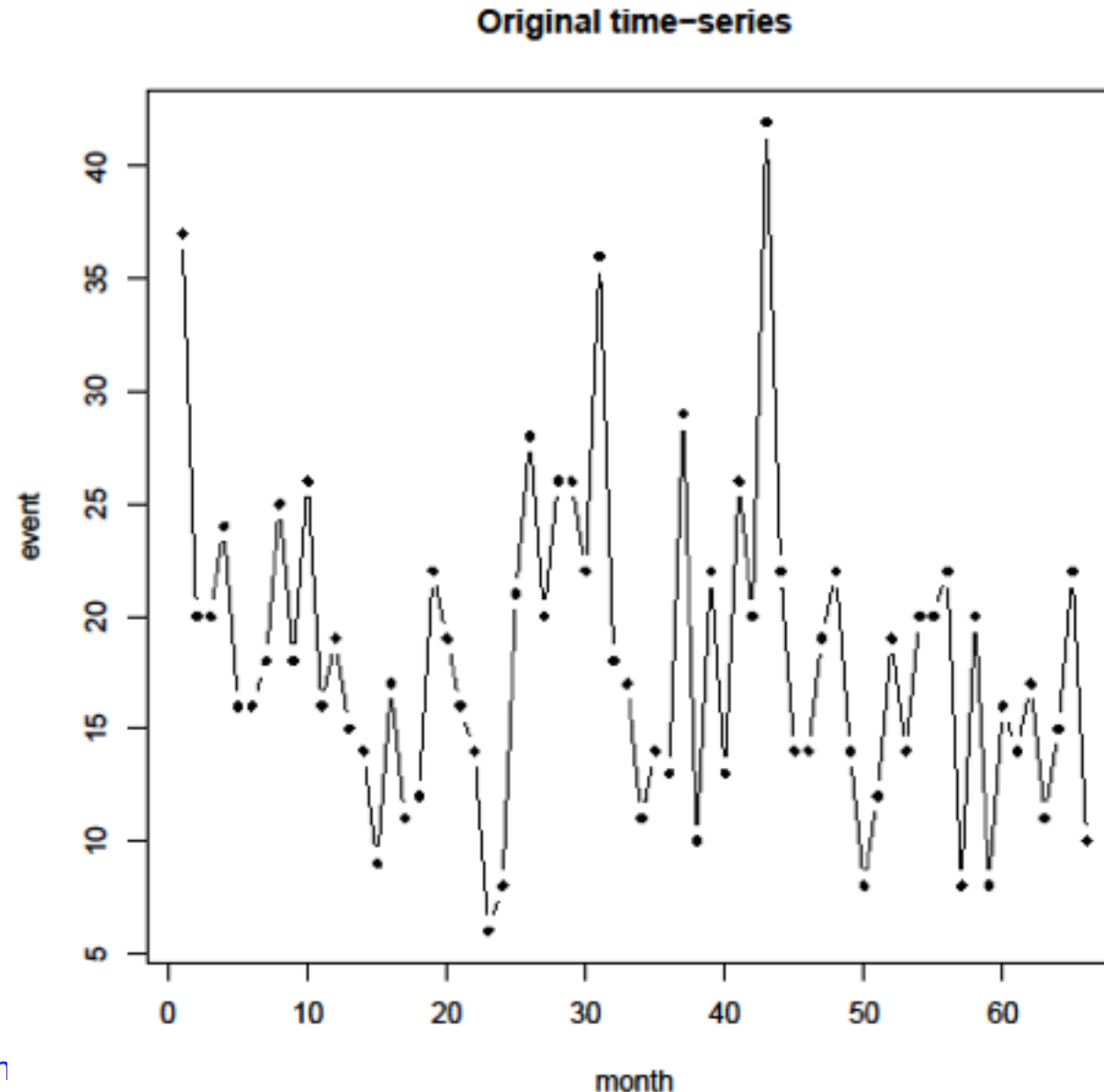
Curve by **Fleet Family**

(2013 ~ 2018 Aggregated)

# Predictive Analysis

- This is a classical time-series forecasting problem.
- Several methods are available and implemented in readily available software/languages like R.
- The following example uses a **moving window scheme** to forecast each of the last 12 months, using the previous months to fit the model with the prophet R package.

# Predictive Analysis – Moving Window Scheme

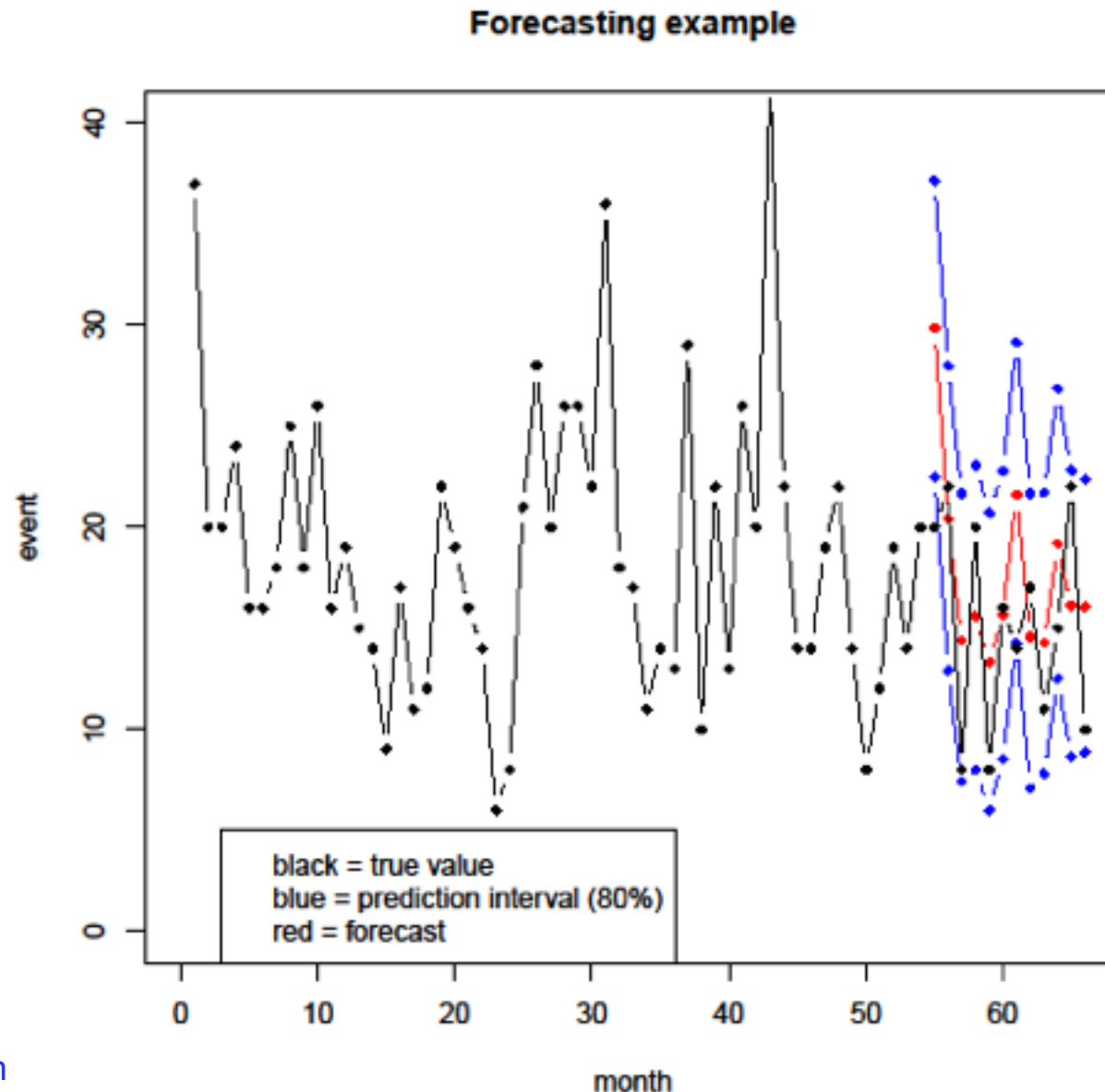


**Event Type:** Landing Gear

**Fleet Family:** Aircraft Type 1

Monthly # of events

# Predictive Analysis – Moving Window Scheme



**Event Type:** Landing Gear

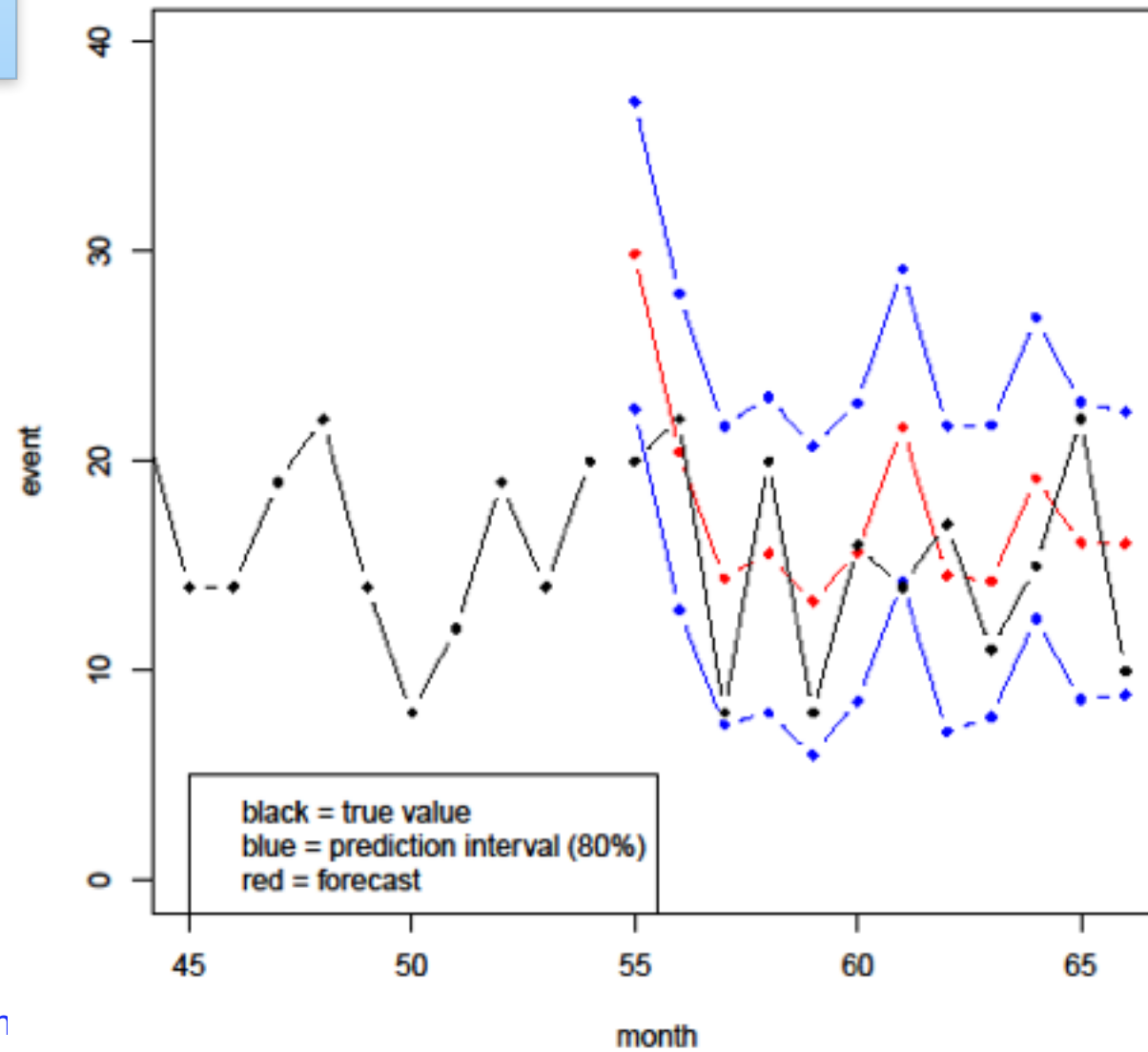
**Fleet Family:** Aircraft Type 1

Monthly # of events

I can explain the graph tomorrow. Basically, you see only one point outside the prediction interval (the one for the first forecast where the black dot (true value) is outside the interval)

# Analysis – Moving Window Scheme

Forecasting example (Zoom in)



**Event Type:** Landing Gear  
**Fleet Family:** Aircraft Type 1  
Monthly # of events

# Next Steps

- Automate the data creation and management, including the verification of the data quality.
- Try and compare several methods for both problems, to find which performs best and suits IATA's needs. See the appendix for a list of methods.
- Automate the data analysis, including the data extraction.
- Prepare visualization and reporting tools, dashboards etc.

# Next Steps

- One way to proceed would be to have a MSc student from HEC Montréal do a supervised project (internship) at IATA.
- A supervised project consists of 400 hours of work within one semester (4 months).
- Students in the specializations *Business Intelligence* or *Data Science and Business Analytics* are perfectly equipped with the technical and managerial skills required for this project.



# Appendix

A few possible methods for problem 1 (anomaly detection):

- Time-series clustering (R package *dtwclust*).
- Functional isolation forest (Python code: <https://github.com/Gstaerman/FIF>). <https://arxiv.org/abs/1904.04573> .
- Robust archetypoids (R package *adamethods*). <https://link.springer.com/article/10.1007/s11634-020-00412-9> .
- Control chart for functional data (R package *qcr*). <https://www.mdpi.com/1099-4300/20/1/33> .

Possible methods for problem 2 (time-series forecasting).

- Numerous R packages available: <https://cran.r-project.org/web/views/TimeSeries.html> .
- e.g.: *fable*, *forecast*, *prophet*.