Predictive Risk Modelling in Aviation Incidents

Prakash Gawas [*] Hyuntae Jung [†] Denis Larocque [‡]
Michael R Lindstrom [§] Guillaume Poirier [¶]
Ahmed Sid-Ali [‖]

August 2020

Abstract

In the 2020 Montréal Industrial Problem Solving Workshop, the International Air Transport Association posed a challenge to participants: to identify anomalies in time series data for flights, across different air craft types and airport origins/destinations. Within this anomaly detection were two questions: how to identify a time series as anomalous and how to identify when a new record is anomalous relative to previous data in the time series. We present our analysis showing a novel method of time series anomaly detection using an extension of kernel density estimation.

keywwords: anomaly detection, time series, air travel

1 Introduction

Currently, the global aviation safety risk identification is mainly reactive with the approach, "we don't know what can be the problem until we face the problem." The International Air Transport Association (IATA) is interested in proactively identifying potential risk areas before they evolve into an accident. Thus, 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.

In this report, we focus upon the two problems outlined below. Some members of this team have also published a paper on the novel method developed for Problem 1 and we refer a reader there for a greater, more technical exposition [1].

[*]Polytechnique Montréal
[†]International Air Transport Association
[‡]HEC Montréal
[§]University of California, Los Angeles
[¶]IVADO
[‖]Carleton University

1.1 IPSW Challenge Target 1: Anomaly Detection

Develop a model to give hints to safety analysts where to look instead of needing to query every criteria one-by-one. The model should examine the set of incident reports by, for example, drilling down into specific aircraft type finding:

- Aircraft Type A does not show significant difference to the global rate
- Aircraft Type B shows anomalous behaviour relative to the global rate, which may indicate prominent safety risk.

Once the model automatically identifies such "anomalies" with statistical evidence, a flag will be raised, so that human safety analysts can perform deeper investigation.

1.2 IPSW Challenge Target 2: Predictive Analysis

Develop a model to predict event rate based on historical records, and flag if the actual rate is exceptional. For example: suppose we are given monthly rates of Event A (with the seasonal pattern). After training with, say, 2 years of historical incident data, the model should make a prediction for the next month with a given interval of confidence. However, the actual data for the next month may be out of the boundary. In this case, that should be flagged as anomalous.

1.3 Data - Incident Reports & Sector

We were provided with Incident Reports. Approximately $621,000$ reports specifying many details. For example, one report could include

- Report ID: 7723515
- Year: 2018
- Month: May
- Fleet Family: ACType5
- Location: Airport162
- Location Country: Country256
- Phase: Approach
- Event: Weather - Windshear

We also were provided with Sector Data, to normalize the flights by the number of flights flying between a given source and destination over a give time window. The data were provided on a quarterly basis. We might for example have:

- Quarter: 2018 Q2
- Fleet Family: ACType5

- Departure: Airport162
- Departure Country: Country256
- Arrival: Airport359
- Arrival Country: Country26
- Sectors: 3,631.

This allows us to compute the flight statistics on a per 1000 flight basis, for example.

2 Problem Solving

We present a series of ideas that could be used in studying anomalies.

- Vectorized representation for data and Logistic Regression
- Neural Network
- Naive Bayes Classifiers
- Functional KDE
- Functional Isolation Forest
- Time-series forecasting (e.g. forecast and prophet R package)

2.1 Data Preparation

As preliminary work, we wrote scripts to process the raw data into a form that could be analyzed for anomalies. The scripts allowed a user to specify certain descriptors of the events they seek and then to obtain time series for those events by fleet or location. For example, a user could obtain the time series for all air craft types for records that listed both "Windshear" and "Turbulence".

2.2 Anomaly Detection

We used two methods for anomaly detection: in the first method, we extended kernel density estimation in a novel fashion to score the time series for their level of anomalousness; in the second methods we used Heirarchical Curve Clustering with the dtwclust `R` package.

2.2.1 Functional KDE Anomaly Detection

Our thought process in developing an extension to KDE for time series can be summarized by:

- Think of an anomaly is 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.
- Since we don't know the distribution we use Kernel Density Estimation.

Kernel Density Estimation (KDE) Review 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 probability density function (pdf),

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

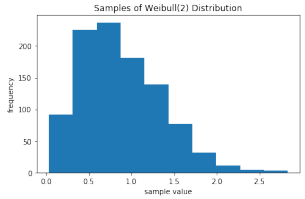

Figure 1: Sample of Weibull distribution

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 — see Figure 1. We also remark that values whose probability density are very low tend to be anomalous as depicted in Figure 2. Thus, if we could ascribe a "probability density" to time series, which are points in a Hilbert space, then we could likewise identify anomalies time series as depicted in Figure 3.

Simple Functional KDE In our first approach, we can think of our time series as samples of signals $x : [0, T] \to \mathbb{R}$ or as being in the Hilbert space, $\mathcal{H}$, say $L^2(0, T)$ or $\mathcal{H}^1(0, T)$. Hilbert spaces have induced norms, $|| \cdot ||$, which can be thought of as generalized distances. The idea is to place a Gaussian kernel over $\mathcal{H}$ at each time series $x_i(t)$ and construct a probability density functional. We can formally, i.e., without rigor, define an empirical pdf over $\mathcal{H}$ with:

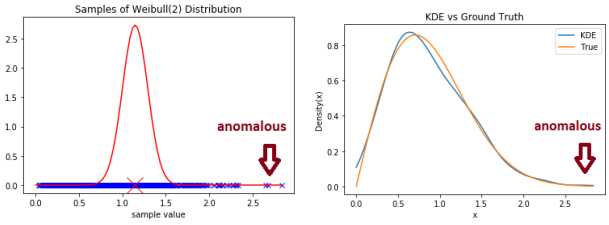

Figure 2: Intuition of anomalies being points far away from the peak density values.

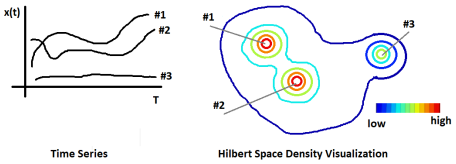

Figure 3: Two curves are very close and there is one anomaly. That curve then, when abstractly mapped to a probability density, has a lower probability density in its vicinity.

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

$$\rho(a) = \sum_{x \in S} e^{-\frac{1}{2\sigma^2}(x-a)^2} \tag{2}$$

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

For "High-Energy/Unstable Approach," scores ≤ 10 seem anomalous (by inspection) — see Figure 4.

Discrete Fourier Transform Functional KDE

In our second approach, we note that $L^2([0,T])$ and $H^1([0,T])$ have countable bases $\{e2\pi in/T\}_{n \in Z}$. Fix M and suppose $x_j(t) \approx \sum_{n=-M}^{M} \hat{x}_n^j e^{2\pi in/T}$. Suppose that each $\hat{x}_n \sim \epsilon_n$ for some pdf ϵ_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

$$f(x_j) = \prod_{n=-M}^{M} \zeta_n(\hat{x}_n^j) \tag{3}$$

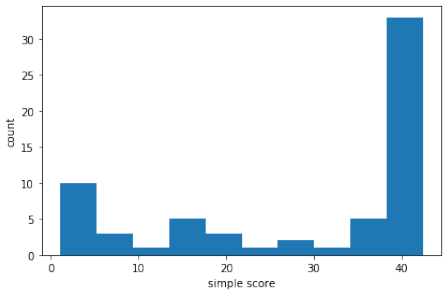

Figure 4: Histogram of scores using simple approach.

In practice: use Discrete Fourier Transform (DFT) since our signal is discrete and finite. A method is summarized below:

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

$$\rho[a] = \prod_{n=0}^{M} \zeta_n(\bar{x}_n) \tag{4}$$

- Assign to each x_j a score $s_j = \rho[x_j]$.
- Identify anomalies by a histogram of $s_j, j = 1, \ldots, N$.

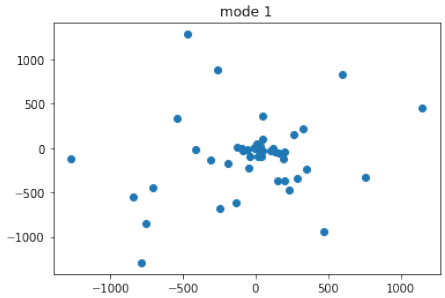

Figure 5: Distribution of Discrete Fourier coefficients at $m = 1$ mode number.

In Figure 5, we display an example distribution of $\hat{x}_1$ values. KDE is done upon this in each Fourier mode. For "High-Energy/Unstable Approach," scores ≤ -510 are anomalous by inspection — see Figure 6.

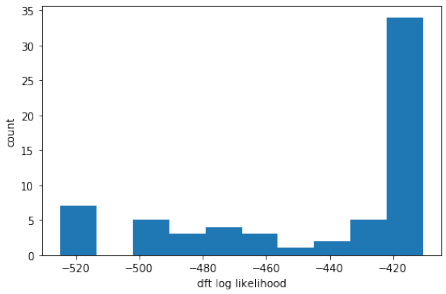

Figure 6: Histogram of time series scores using Fourier approach.

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

Table 1: Anomalies air crafts for different methods for different event types.

Simple vs DFT Comparisons For select events, we plot the anomalies air craft numbers for the two methods in Table 1. Anomalous flight IDs have significant overlap between the two methods. Everything the DFT method finds is also found in the Simple method.

We plot the time series of "High Energy/Unstable Approach" in Figure 7; the ordinary curves are in blue, based on the DFT classification. The anomalous curves are red. Interpretation is an open question: identifying why a curve is anomalous.

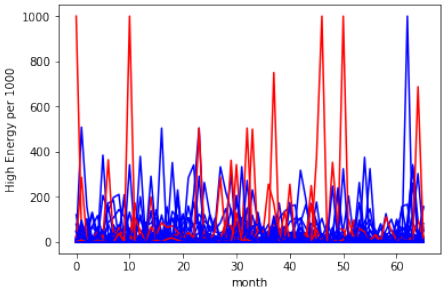

Figure 7: Time series of "High Energy/Unstable Approach" events with anomalous series in red.

Anomaly Detection – Hierarchical Curve Clustering

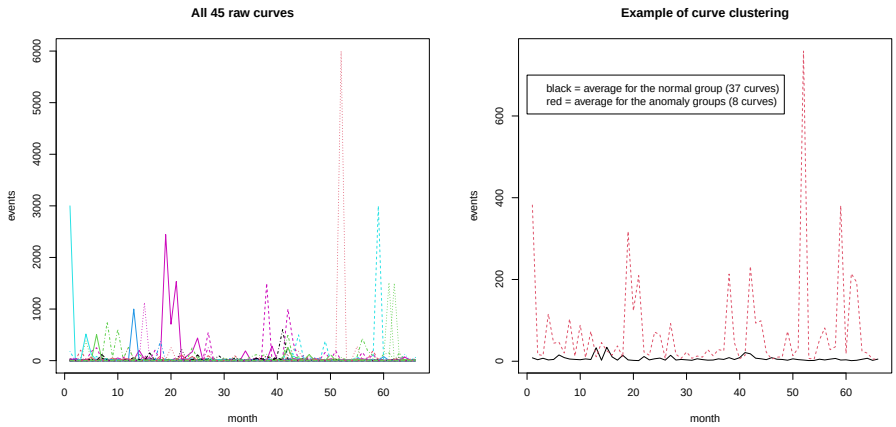

Figure 8: Descriptor: Windshear. Curve by Fleet Family (2013 2018 Aggregated)

2.2.2 Predictive Analysis

Here we have 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. See Figure 9.

3 Next Steps

Following the workshop, there are now a number of next steps that could be taken to further develop our methods and make the results more useful. We list a few below:

- Automate the data creation and management, including the verification of the data quality.
- Try and compare several anomaly detection methods to find which perform the 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.
- One way to proceed would be to have a MSc student from HEC Montréal do a supervised project (internship) at IATA.

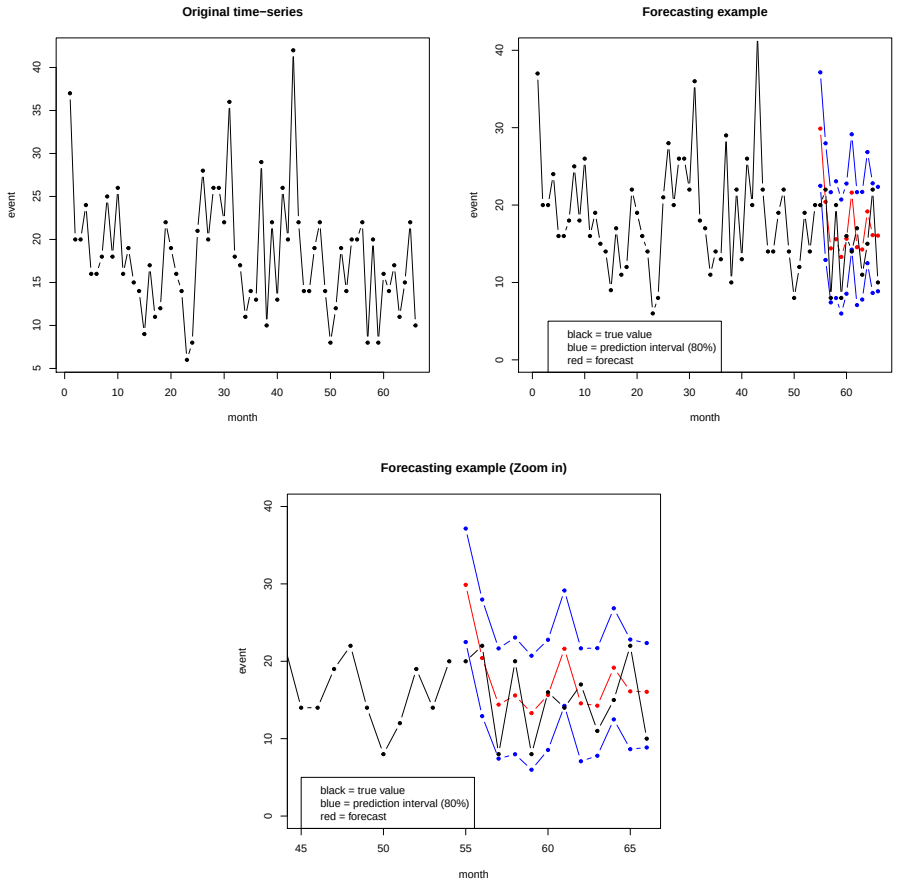

Figure 9: **Event Type:** Landing Gear. **Fleet Family:** Aircraft Type 1. Monthly # of events

- A supervised project consists in 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.

References

[1] M. R. Lindstrom, H. Jung, and D. Larocque, *Functional Kernel Density Estimation: Point and Fourier Approaches to Time Series Anomaly Detection*, Entropy, 22 (2020), p. 1363.

Acknowledgments

We would like to thank the Centre de recherches mathématiques (CRM) and the Institute for Data Valorization (IVADO) for hosting the workshop online. We also thank Odile Marcotte for organizing the workshop.

Author Contributions

PG helped in data processing. HJ helped by presenting the problem, providing data, and analyzing/interpreting the results. DL helped in developing and running methods and writing this report. MRL helped in data processing, developing and running methods, and writing this report. GP helped in coordinating the work. AS helped in writing this report.

A Additional Anomaly Detection Methods

- 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.