Preliminary Analysis

Flight data anomaly detection.

Lucas Coelho e Silva

17/09/2020

Contents

List of Figures

1 Introduction

This report describes the inital findings of the exploratory data analysis conducted for building the foundation of the anomaly detection algorithm to be developed. Is is divided in the following sections:

- **Flight data**: brief description of the flight data collection used;
- **Preliminary data visualization**: introductory charts for flight data familiarization;
- **Benchmark**: presentation of the results obtained with PyMKAD;
- **Landing drill down**: discussion focused on the landing fase;
- **Anomalies**: introductory discussion of the potential anomalies found in the data (quantitatively and qualitatively);
- **Clustering**: initial findings of the application of clustering algorithms.

This report currently focuses on the Landing phase, for which the MKAD algorithm detected anomalies.

The underlying code for analyzing the flight data and plotting charts can be found in *eda.py* and *eda_clustering.py*.

2 Flight data

For this prelimnary analysis, the flight data used was the one publicly available with the Multiple Kernel Anomaly Detection (MKAD) algorithm (https://ti.arc.nasa.gov/opensource/projects/mkad/). It is composed by 112 flights, available as *csv* files. Each flight contains observations for the following parameters:

- Time;
- Altitude;
- AirSpeed;
- Landing_Gear;
- Thrust_Rev;

- Flaps;
- Param1_1;
- Param1_2;
- Param1_3;
- Param1_4;
- Param2;
- Param3_1;
- Param3_2;
- Param3_3;
- Param3_4;
- Param4.

2.1 Flights heterogeneity

The flights available clearly encompasses operations for different routes, given the diverse flight times and distances from touchdown presented on the charts in the *Preliminary data visualization* chapter. This raises a few questions:

- Should anomaly detection be restricted to a set of flights that are supposed to be similar? In a sense that the algorithm must beware that those flights refer to the same operation (landing at a given airport, e.g.)
- What are the disadvantages of restricting the sets of flights? One could lose track of a rare but potentially hazardous anomaly that happened in two different operations but with the same root cause. In order to avoid these scenarios, the methodology could contemplate the anomalies also being clustered amongst themselves.

3 Preliminary data visualization

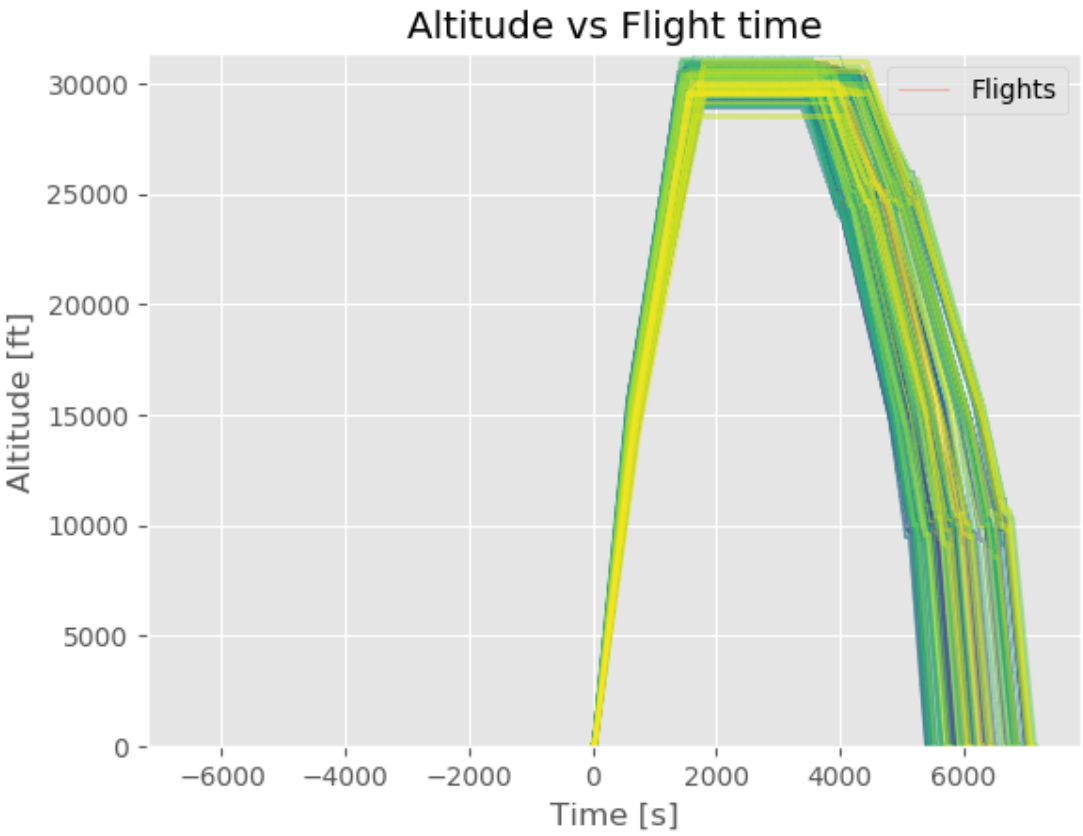

Figure 1: Altitude vs flight time

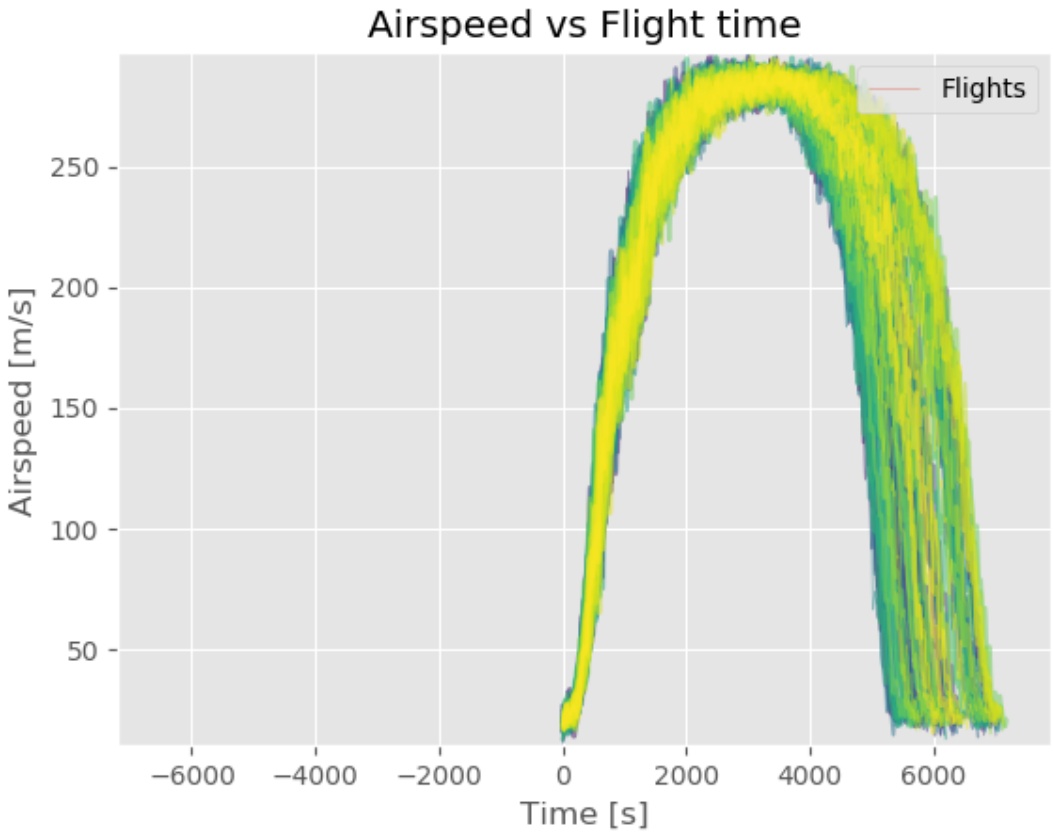

Figure 2: Airspeed vs flight time

Since the landing phase is the one currently being considered, it is more adequate to visualize the data in terms of the distance from touchdown.

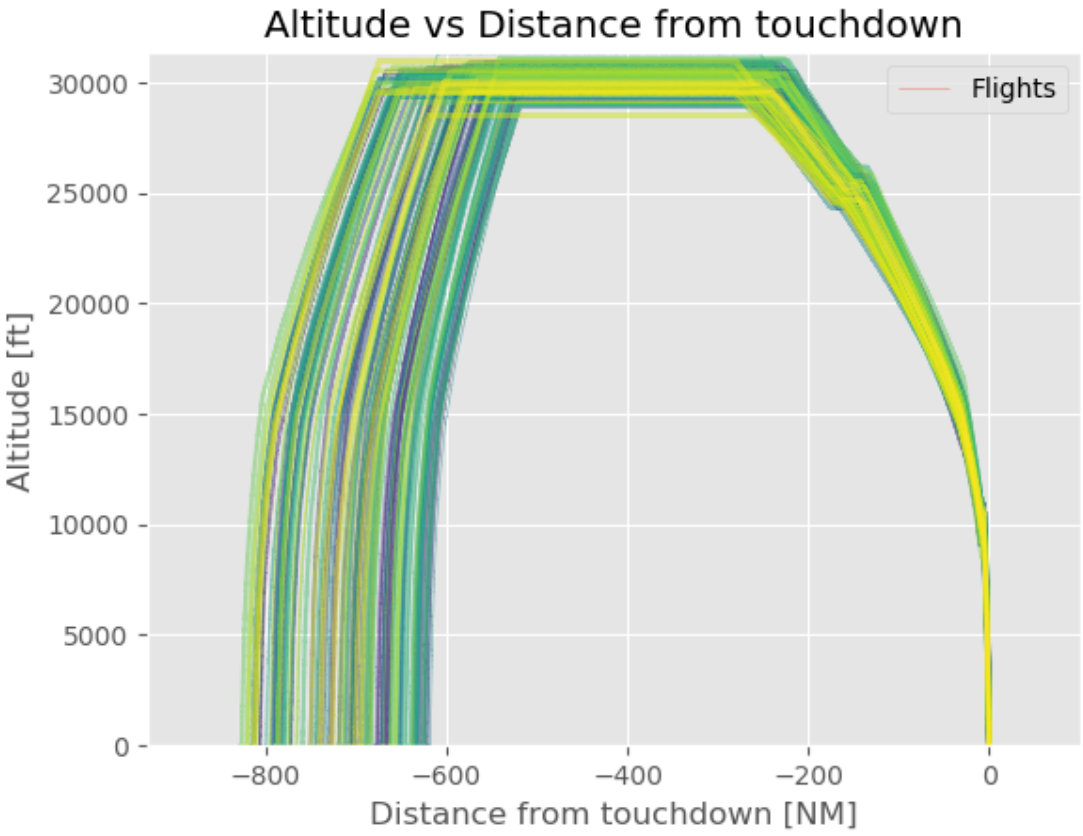

Figure 3: Altitude vs distance from touchdown

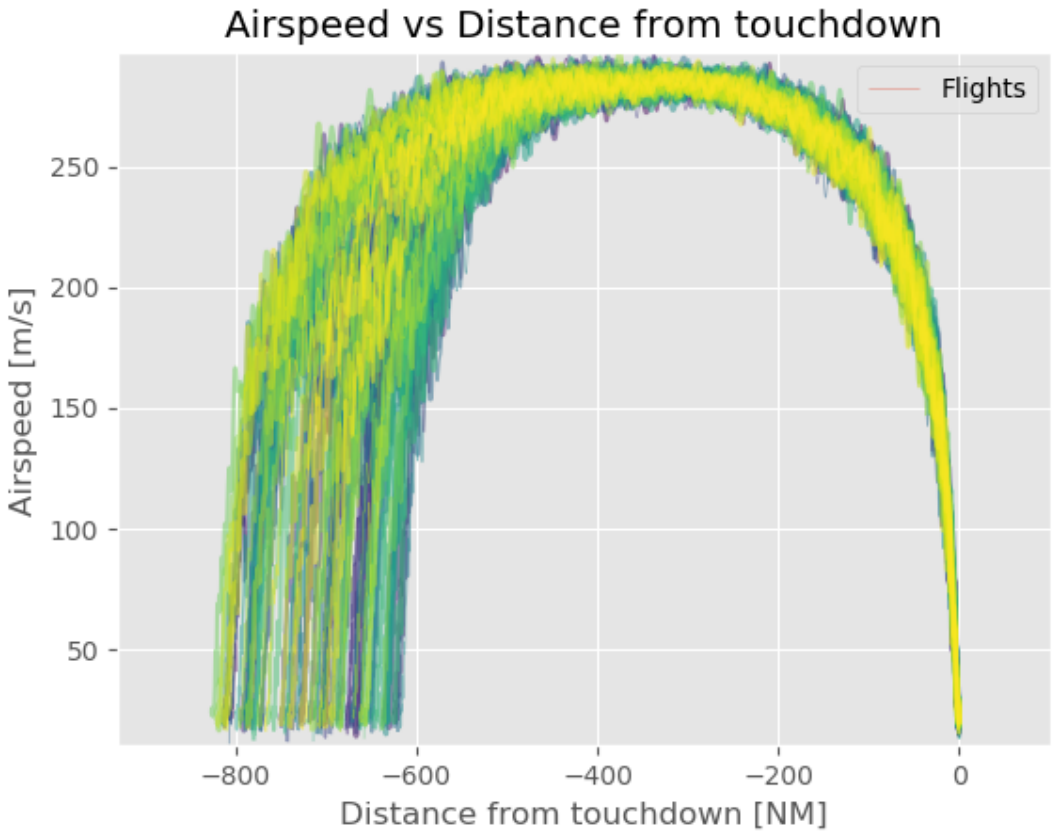

Figure 4: Airspeed vs distance from touchdown

4 Landing drill down

As mentioned, this report will currently focus on the Landing phase, for which the MKAD algorithm detected anomalies. The charts are then redrawn to emphasize the landing window.

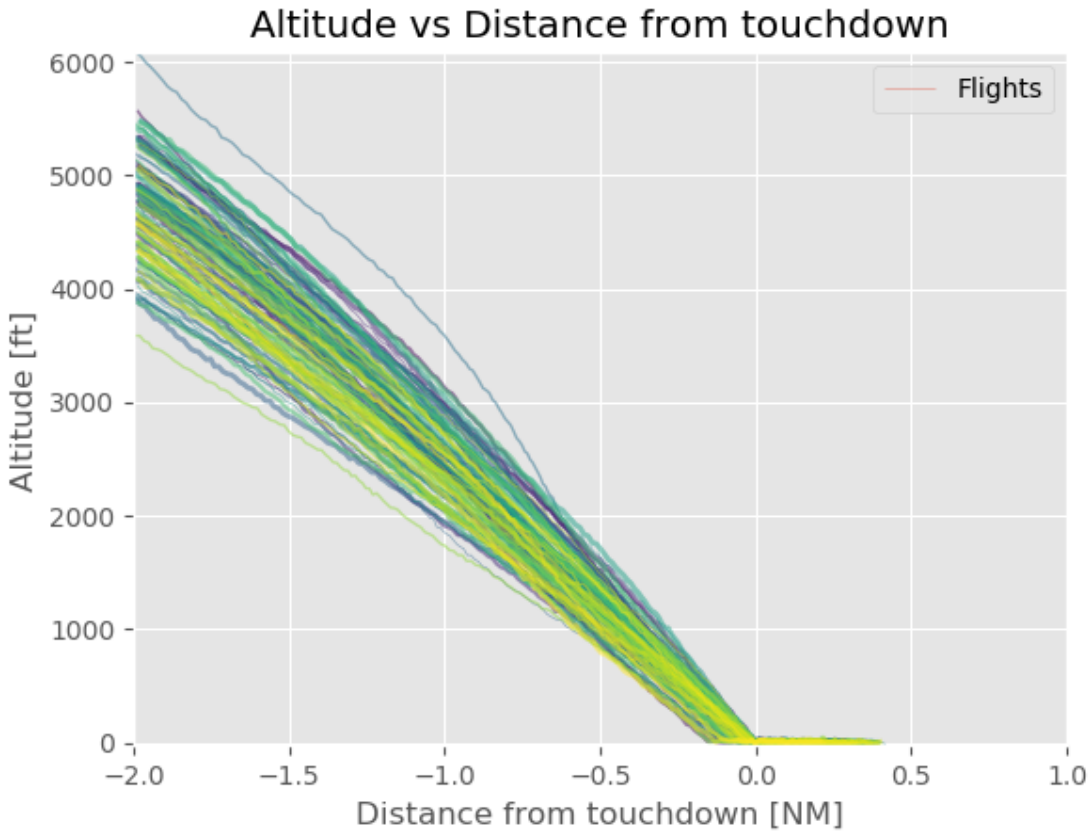

Figure 5: Altitude vs distance from touchdown (landing)

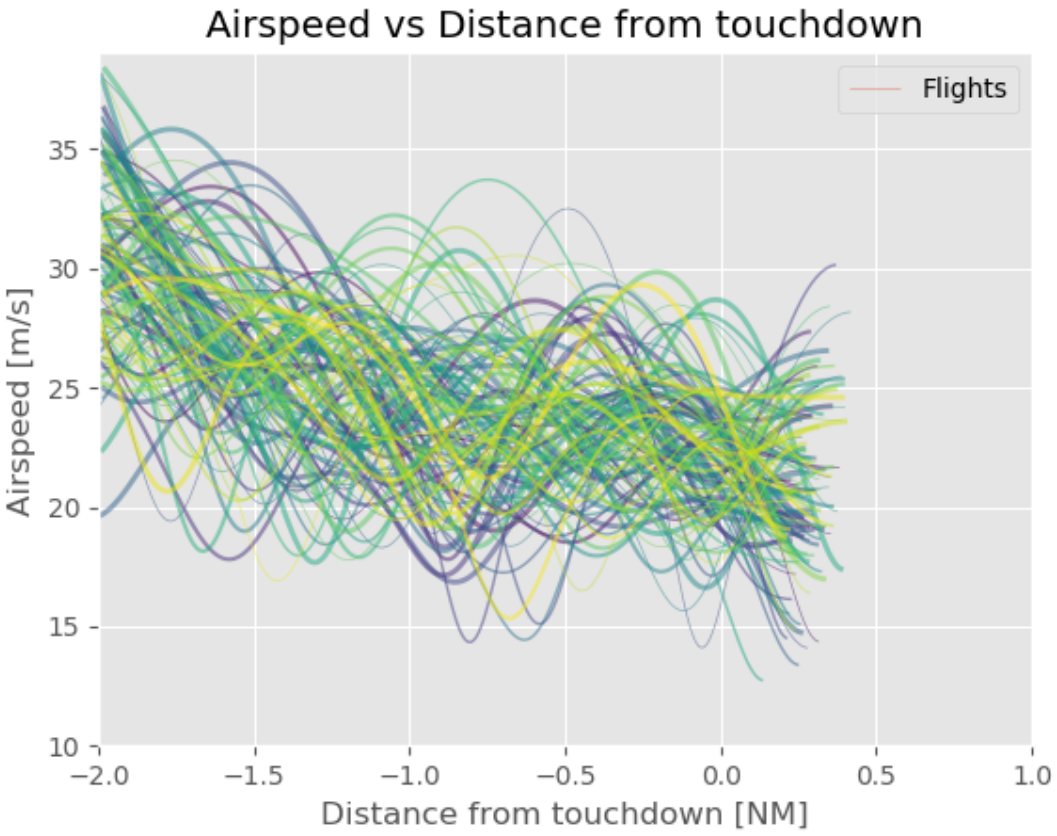

Figure 6: Airspeed vs distance from touchdown (landing)

5 Anomalies

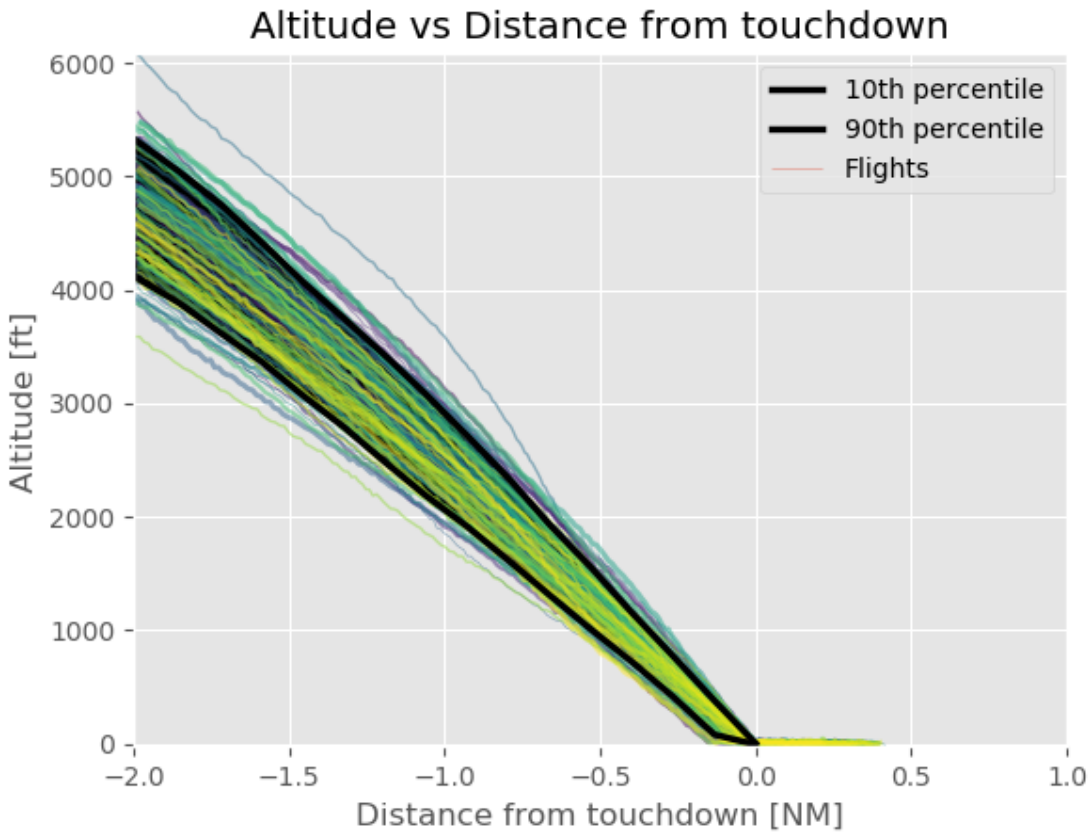

Figure 7: Altitude vs distance from touchdown (landing), bounded

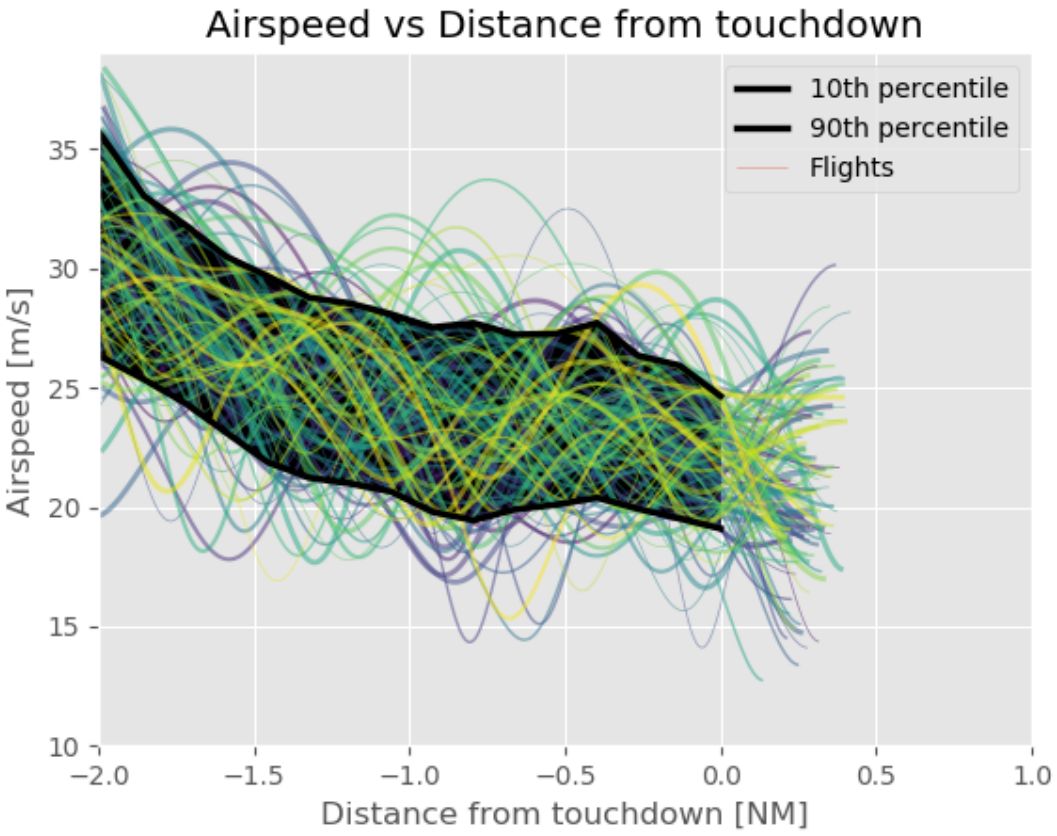

Figure 8: Airspeed vs distance from touchdown (landing), bounded

Include discussions.

6 Benchmark

Include anomalies detected by MKAD.

7 Clustering

Initially, DBSCAN (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html) is being used as the clustering algorithm. The Python code used for clustering can be found in *eda_clustering.py*. It aims to be conceptually similar to ClusterAD-Flight, defined by Li and Hansman [1].

The flights marked as “anomalies”" (isolated clusters) are listed on the table below.

flight	cluster
12fdaef5792eb79ae2685ffffc15efe1.csv	1
37efcdaac9667889617cd5189e6ba5b8.csv	1
146dd4496e5ef2eba89aa55d9d4ace4c.csv	2
15eb86afde34f1c6a1ef3ef33bb81982.csv	3
1631c1efd7963737ecd73659ff6a166f.csv	4
19df64ae9a8cce2682bef85b77546214.csv	5
1cd65e48f8d35f96d741def4198774c6.csv	6
21aecaefd9e5bd94e9d57332571b727a.csv	7
21ef7536b9eca257b33a6c17fe1dbfb1.csv	8
24bfd17a76b33838b31bbb8bcef1281b.csv	9
262136efefc3926fe35c1b7df274af2c.csv	10
2b8baa6ae15f9788cef115b9abf9f6c5.csv	11
2f997c335e6f28b3734b427e7efe87ab.csv	12
38b276c9c7aaa31ef65894fc48c34b81.csv	13
3c4effde731b8997dda917444c53edaa.csv	14
3e9e7f26e3d28ef52f8493f57c3682d4.csv	15
41961efd5d83b78b46236b326cc2429a.csv	16
44ec864377d5a988f15a7962158a69ef.csv	17
4f1412735ee11b25938638ef34e84383.csv	17
4553e45c2691b923e12bfe2a51cfefea.csv	18

In the figures below, the red lines mark the “anomalous” flights. For now, only the Altitude and Airspeed charts are presented, even though the other parameters might be the ones indicating an anomalous behavior.

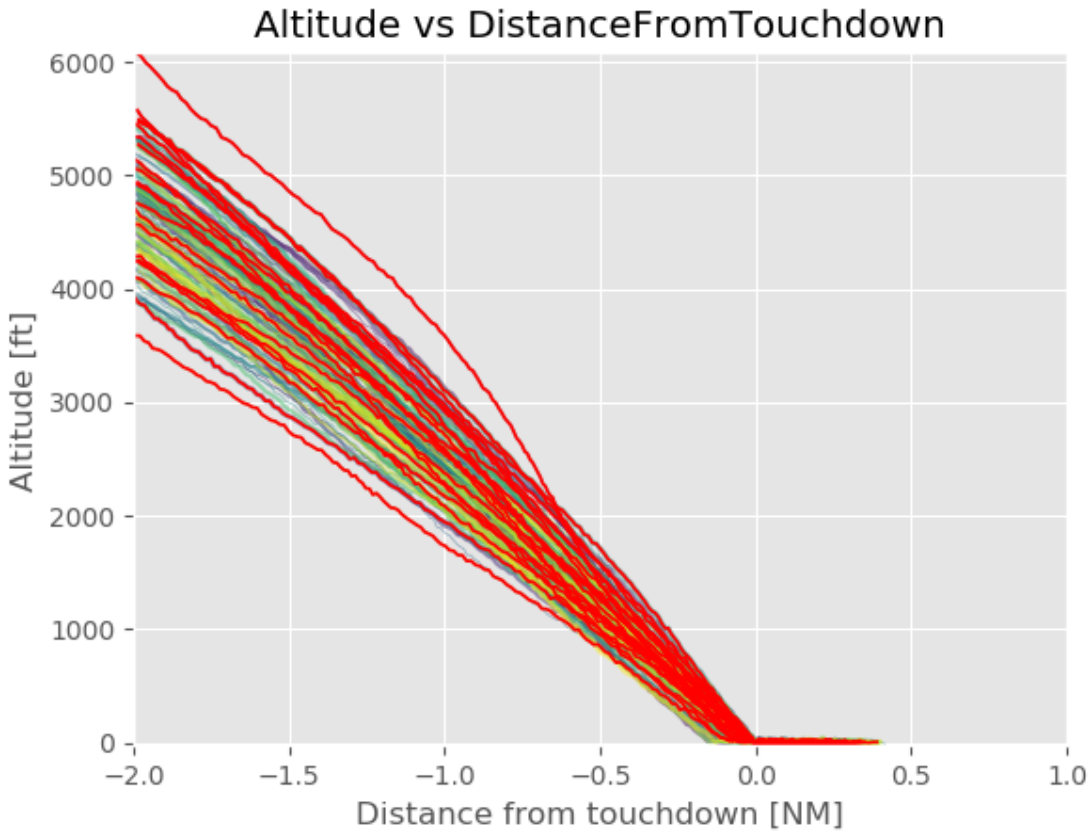

Figure 9: Altitude vs distance from touchdown (landing)

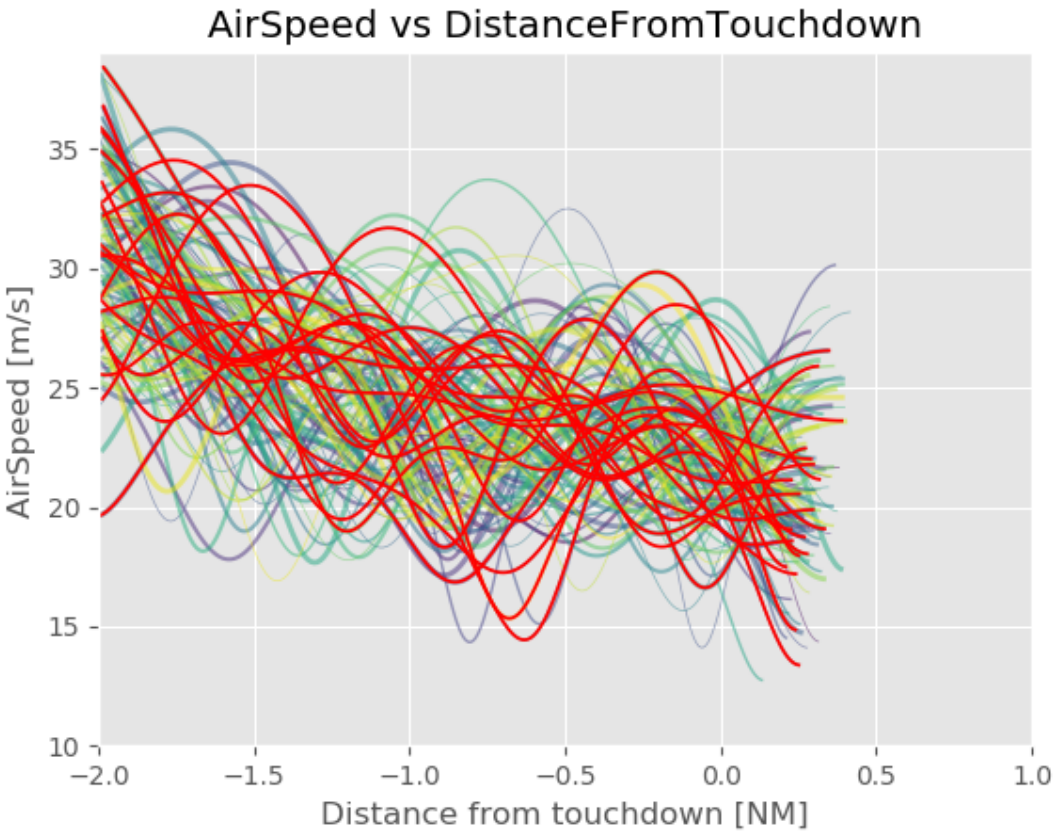

Figure 10: Airspeed vs distance from touchdown (landing)

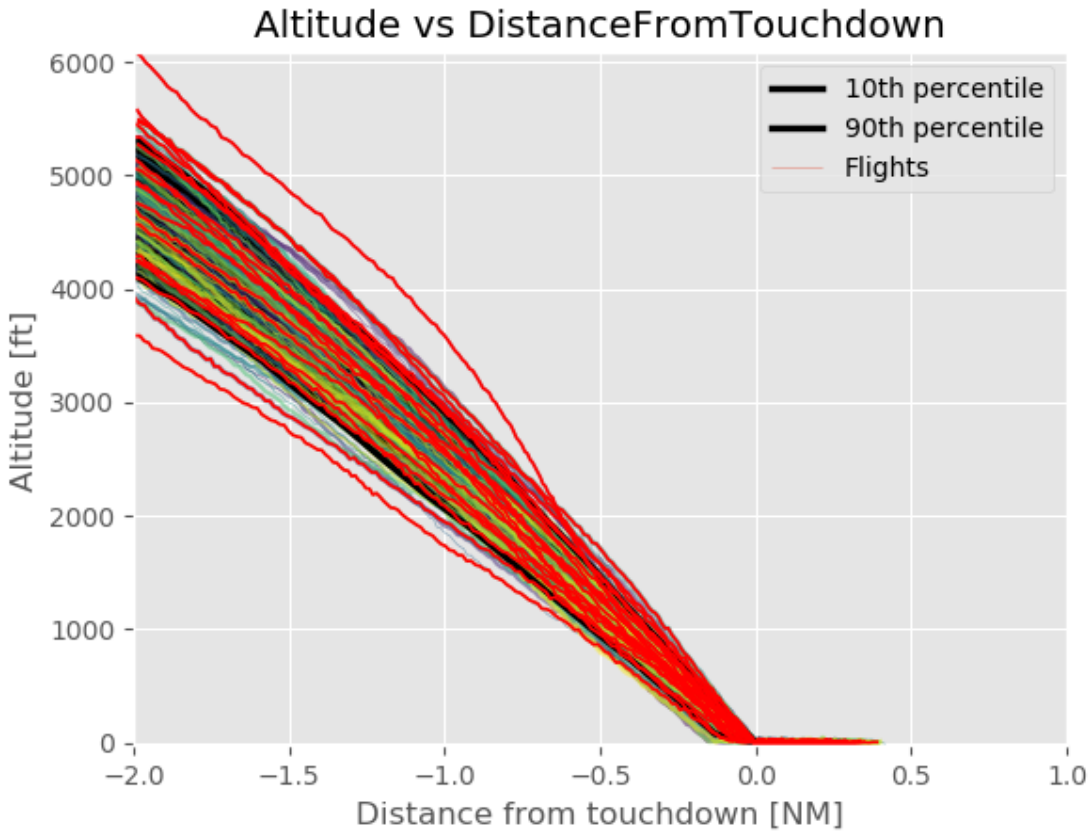

Figure 11: Altitude vs distance from touchdown (landing), bounded

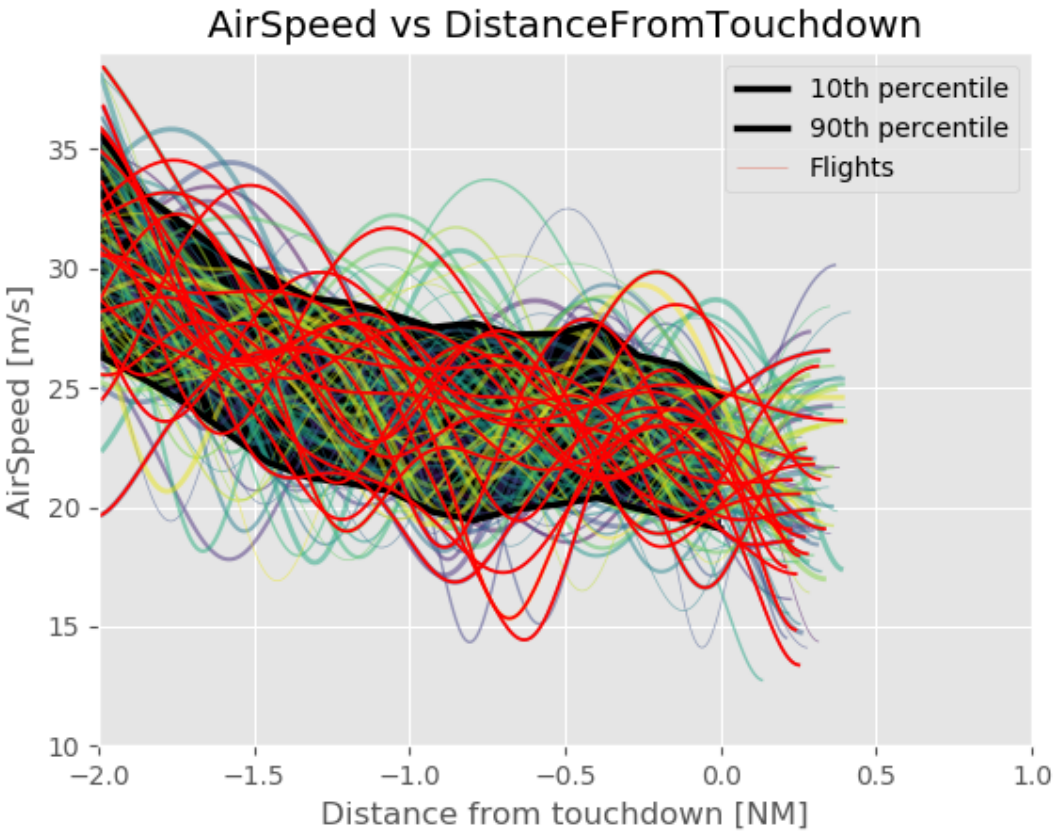

Figure 12: Airspeed vs distance from touchdown (landing), bounded

Discuss clustering methodolgy and results. Improve visualization without color repetition.

References

[1] Lishuai Li and R John Hansman. “Anomaly Detection in Airline Routine Operations Using Flight Data Recorder Data”. en. PhD thesis. Cambridge, MA: Massachusetts Institute of Technology, June 2013.