EPiC Series in Computing

Volume 67, 2019, Pages 51–62

Proceedings of the 7th OpenSky Workshop 2019

Anomalies in the Sky: Experiments with traffic densities and airport runway use

Axel Tanner[1] and Martin Strohmeier[2]

[1] IBM Research-Zurich, Rüschlikon, CH
axs@zurich.ibm.com
[2] armasuisse, Thun, CH
martin.strohmeier@armasuisse.ch

Abstract

Anomalies in the airspace can provide an indicator of critical events and changes which go beyond aviation. Devising techniques, which can detect abnormal patterns can provide intelligence and information ranging from weather to political events. This work presents our latest findings in detecting such anomalies in air traffic patterns using ADS-B data provided by the OpenSky network [8]. After discussion of specific problems in anomaly detection in air traffic data, we show an experiment in a regional setting, evaluating air traffic densities with the Gini index, and a second experiment investigating the runway use at Zurich airport. In the latter case, strong available ground truth data allows to better understand and confirm findings of different learning approaches.

1 Introduction

Today's air traffic is a highly complex system by volume as well as the number of different actors, structured partially top-down through regulations, but also with emerging behaviour bottom-up. Patterns exist on very different scales from global flight routes to patterns due to local conditions. These patterns follow normal changes through time of day and season, but also deviate in the macro and micro level for many different reasons. Anomaly detection tries to find interesting normal patterns as well as deviations from that, but ultimately the point of interest of an investigation decides what is to be considered an anomaly and what the 'normal' behaviour.

This report presents recent experiments looking for anomalies in air traffic data. As background, we provide a short definition and characterization of anomalies and specific challenges in the aviation context in Section 2. A fundamental problem is the availability of ground truth data usable for the learning process or the evaluation of anomalies flagged by the respective detection approach. After a short discussion of potential sources of ground truth in Section 3 we describe an experiment in a regional setting using the Gini index to evaluate the changing air traffic densities in Section 4.1 and different approaches in a local setting looking at the runway use at the Zurich airport in Section 4.2.

2 What are anomalies?

	single observations	**trajectory level**	**groups of trajectories**
point	squawk alerts	runway excursion go-around atypical/non-compliant approaches special trajectories (e.g., surveillance)	changing flight route 737 MAX grounding
contextual	wrong sensor readings	schedule anomalies/delays late-night arrival	weather related unusually high/low traffic volume at given times
collective	gusty wind behaviour	deviation due to conflict resolution	blocked regions (like WEF) political unrest/crisis zones volcanic activity

Table 1: Examples for anomalies in the context of air traffic control

Anomalies have been defined as "patterns in the data that do not conform to a well-defined notion of normal behaviour" [3]. This fits the intuitive view, but also makes it explicit that anomalies can only be evaluated versus a well-defined understanding of some behaviour deemed normal. As such, the search for anomalies is very dependent on the question one is interested in. Chandola et al. [3] also review different general aspects of anomaly detection, including the *nature of the data* and the *type of anomaly*:

Nature of the data: depending on the problem statement, a data point can, e.g., represent a single ADS-B message, a whole trajectory for a certain flight or aircraft, or specific groups of trajectories like clusters of multiple trajectories describing a flight route, landing corridor or grouping by aircraft types etc.

Type of anomaly: *point* anomalies represent data points that are by themselves not normal. In contrast, with *contextual* anomalies the data points are not anomalous by themselves, but given a specific, defined context, e.g., in time or space. Finally, there are *collective* anomalies, where individual data points may again not be anomalous, but the collection of points is (not only in the respective context) .

Table 1 gives a selection of examples for potential anomalies in context of air traffic control, categorized by these criteria.

3 Challenges of anomaly detection

There are several challenges that are specific to the detection of anomalies. Firstly, the specifics of the approach strongly depends on the particular question of interest - normal behaviour in

local	regional	global	comments
weather (general sources, METAR)			usually not freely available, METARs only at airports
airport setup and operation guidelines			explicit, but highly specific
NOTAM			well-defined including time & space
specific squawk codes			highly specific, only for single flights
News / Social Media			broad, but fuzzy
political, conflict zone (Ukraine, Qatar)			rare, large scale events
volcano			

Table 2: Examples for sources of ground truth in the context of aviation

one context can be an anomaly in a different setting, therefore there is no generic and general approach. Many different techniques have been developed, each with their range of applicability and their specific strengths and weaknesses. For a good review of methods specifically used in the context of aviation, see the upcoming review by Basora et al. [2].

In addition, anomalies are often rare, creating a strong imbalance in the available data. On the one hand, this makes it hard to learn specific models for these anomalies, i.e., to explicitly find a model for the anomalies. On the other hand, in the general case, one needs to ensure that the anomalous behaviour should not be part of the training data, but if anomalies are very rare, modeling the normal behaviour might not incorporate the anomalous behaviour into the model due to its scarcity, so that even without excluding them from the training data these rare anomalies might be flagged as outliers.

The current work shows the challenge around *ground truth*: depending on the learning method, ground truth is either required for labeling the data (supervised and semi-supervised approaches) or to evaluate findings of data flagged as anomaly (unsupervised approach).

Some examples for sources of ground truth in the context of aviation are shown in Table 3. In general, sources range from exact and specific (e.g., METAR information for weather at the airport) to very general, but inexact or fuzzy (e.g., Twitter data). In the end, detailed expert knowledge is often the best and only source to finally evaluate data flagged as anomaly by a chosen approach, but is of course a rather scarce and highly valuable resource hard to access for experimentation.

4 Experiments

We present two different examples of anomaly detection scenarios, the first one in a regional setting, the second in a local setting.

All ADS-B data is obtained from the OpenSky Network [8], a crowd-sourced receiver network that collects air traffic control data on a large scale and makes it available to researchers.

4.1 Regional setting: observing air traffic densities

Motivated by visibly changing geographic distributions in air traffic densities, as exemplified in Fig. 1, we ran experiments using the Gini index as a high-level measure of the distribution

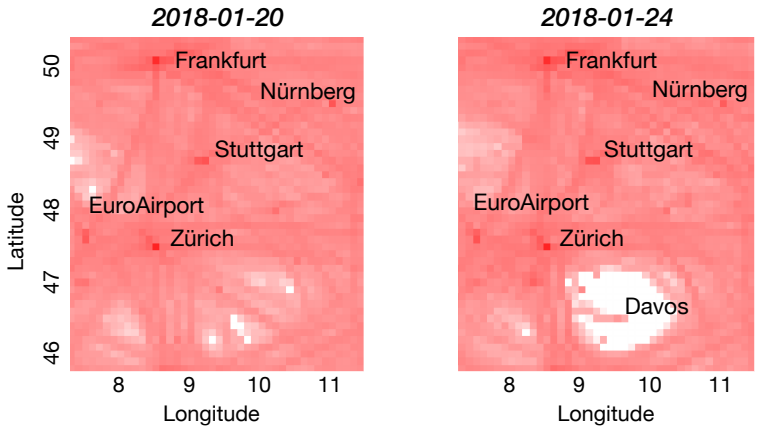

Figure 1: Distribution of ADS-B messages for two different days January 2018

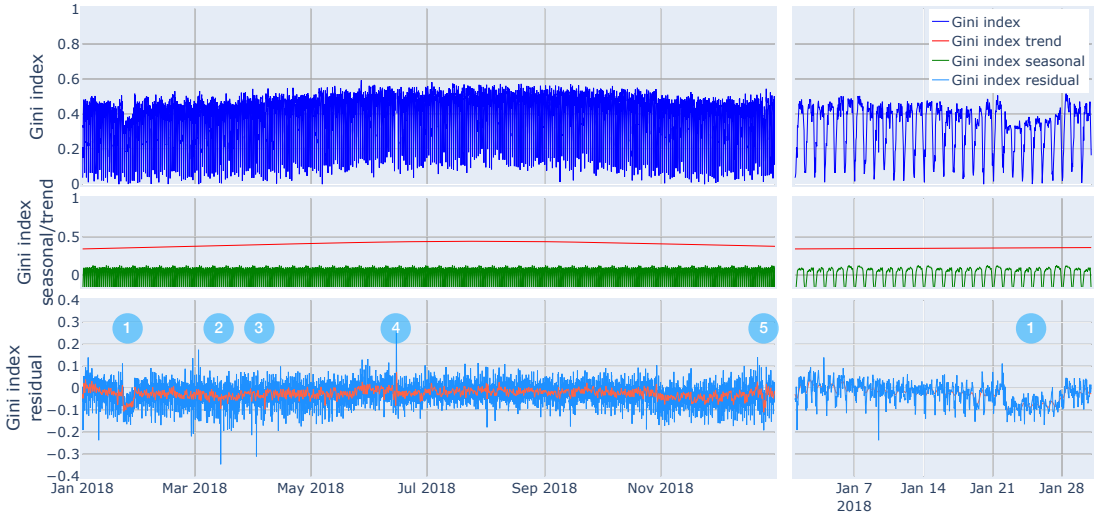

Figure 2: Gini index for hourly densities of air traffic: upper part shows the direct Gini index over time in blue,; the middle part shows seasonal and trending parts of the decomposed Gini index; the lower part shows the residual signal (light blue) with a smoothed line for better visibility (in orange). On the left the data for the full year is shown, on the right the data for January 2018. For markers in the lower part, see text.

of densities. For this, we used geographic cells (0.1° by 0.1° degrees longitude/latitude) and collected the number of ADS-B messages per hour per cell through the year 2018. For each hour, the Gini index of these cell densities is calculated. The Gini index measures the equality of the values in the distribution, ranging from 0 (all values are identical) to 1 (all values maximally unequal). A resulting curve for the Gini index for a large part of Switzerland as function of time is shown in the upper part of Fig. 2. As of course the densities themselves vary very strongly through the hours of each day, from very low traffic in the night (meaning a rather equal distribution) to high traffic during the day (unequal distribution), there is a strong regular

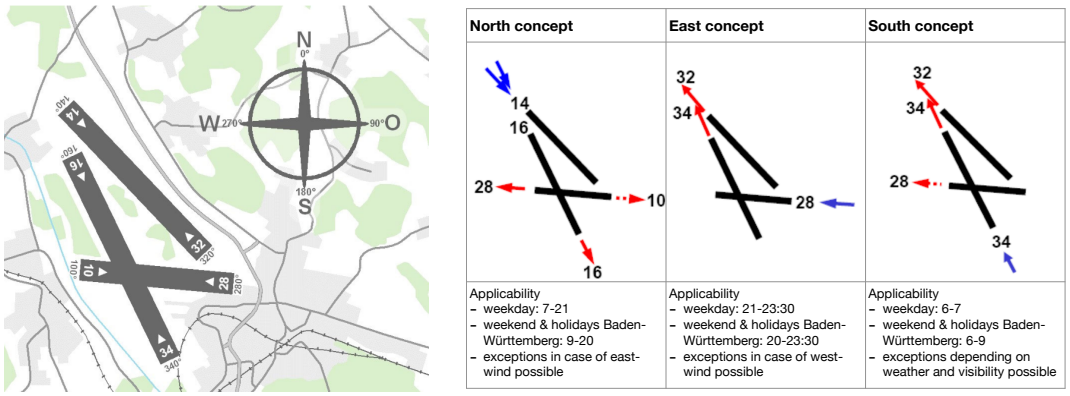

Figure 3: Zurich airport ZRH: runways (left) and operating concepts (right - translated from [1]). Baden-Württemberg is the German province bordering to the north of the airport.

daily pattern in this data. Trying to extract a more detailed signal, we decomposed the data into seasonal, trend and residual parts, using the `STLDecompose` library [6] (using weeks as the duration of the 'season'). As can be seen in the middle row of Fig. 2, the trend signal extracts the slow and weak yearly dependence, the extracted weekly repeating pattern is visible in green, showing the regular daily pattern with slight weekend deviations.

The lower part of Fig. 2 shows the residual data that we used as our main signal: it is obvious that the signal is still very noisy. The strongest anomaly can be seen close to the end of January 2018: the block-out of air traffic due to the World Economic Forum around Davos, taking place January 23-26 in 2018, marked with *1* in the figure. This event generates a large change in the traffic pattern, visible in the right-hand picture of Fig. 1, that correspondingly changes the Gini index during these days significantly. A second strong signal, *5* in the figure, is seen in December, when traffic is sufficiently different after Christmas to leave a trace. Other observable changes are either a result of problems in the data capture (*4* in the figure), or mark events (*2* and *3*) that we could not track to a specific event.

Therefore overall, though large scale events can be seen with this method, we found it too noisy and unclear to be helpful in finding and understanding smaller scale events.

4.2 Local setting: runway use at the Zurich airport

In our second, more extensive example, we present the usage of runways at the Zurich airport in 2018 in the context of time of day, with respect to the official operating concept and weather related deviations. We describe runway and operating concepts of the airport, the extraction of the runway information from OpenSky data, the wind situation and then delve into different experiments to learn from the data in unsupervised, supervised and semi-supervised approaches.

4.2.1 Airport runway situation and operating concepts

The Zurich airport, IATA code ZRH, has three runways, which are utilized following different *operating concepts* that are defined partially due to several statutory and political requirements — the airport is close to the German border. All three runways and the three different official operating concepts are shown in Fig. 3. Part of the political requirements are reflected in the dependency on holidays in Baden-Württemberg , the neighboring *German* province.

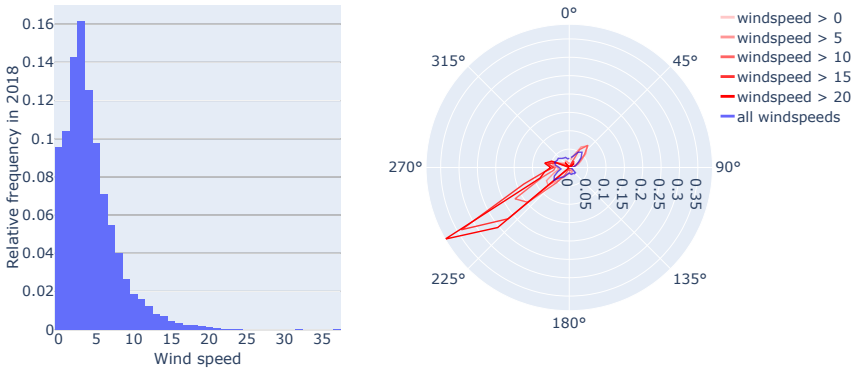

Figure 4: Wind speed and direction distribution as observed at the Zurich airport from METAR data throughout 2018

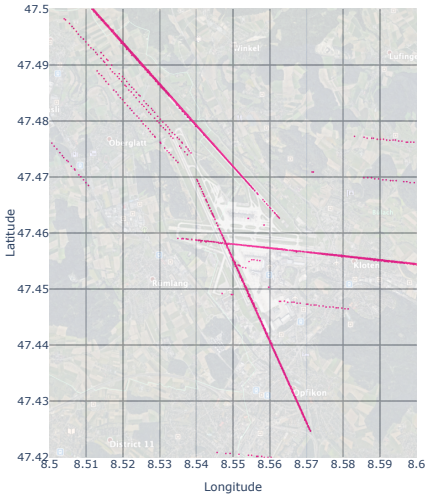

Figure 5: Position data from OpenSky around the Zurich airport for 2018-01-01 with visible 'ghost runway' tracks

4.2.2 Wind Situation

We are using the published METAR data for the Zurich airport[1] as the source for weather-related information. This information is published twice per hour and includes data for wind speed and wind direction. Fig. 4 shows summary information about the distribution of wind speeds and wind directions in 2018. The right-hand side shows wind directions for all speeds (in blue) and selected ranges of wind speeds (in red). The latter highlights that nearly all stronger wind comes from west-southern directions.

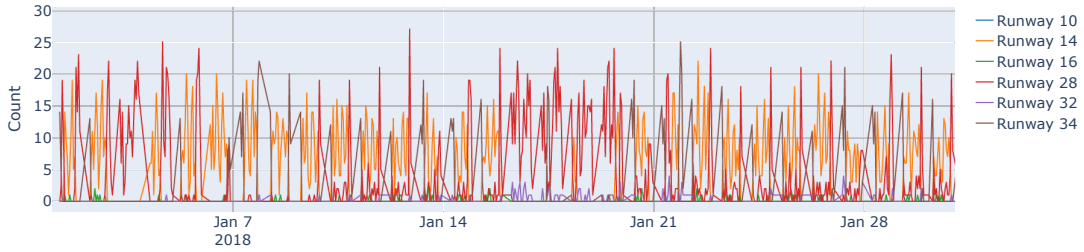

Figure 6: Hourly take-offs and landings per runway at the Zurich airport for January 2018

4.2.3 From OpenSky data to runway use

Although there is good line-of-sight coverage down to the ground by OpenSky sensors near the Zurich airport, the detection of runway use is not completely straightforward due to non-precise position data and spurious data, e.g., after landing respectively before take-off. As an example, Fig. 5 shows pure position data for a day in the area of the airport below 600m — clearly visible are 'ghost' runways, parallel ADS-B recordings that in general belong consistently each to a specific aircraft transmitting false position data.

The current evaluation uses a combination of selecting data to a close neighbourhood to the airport and below an altitude of 900m (altitude of the airport is 432m), then evaluating per callsign the final stretch of heading values consistent with a straight path. Mean `vertrate` of this track is used to detect whether the plane is starting or landing. The resulting data is pruned to exclude borderline or unclear events (e.g., for go-arounds), currently leading to an artificial imbalance between observed take-offs and landings. The resulting data is then in the form of hourly values for take-offs and landings per runway (i.e., including direction), exemplified in Fig. 6.

For our anomaly detection experiments, we regroup this data into vectors describing the use of the runways for a full day. Therefore, a data point captures the count of take-offs and landings on the different runways per hour-of-day for a given day, i.e., represented as one vector with $24(\text{hours}) * 6(\text{runway-directions}) * 2(\text{take-off/land}) = 288$ dimensions.

4.2.4 Unsupervised learning through clustering

In unsupervised learning, we use the available data without a specific classification/labeling of the data points. Fig. 7 shows the different clusters obtained with HDBSCAN [5] as clustering method. Depending on the detailed parameters (*min_cluster_size*, *min_samples*), slightly different sets of clusters are obtained, but these differ mostly in a more or less finer differentiation of smaller clusters while the larger clusters are very stable.

As described above, each data point corresponds to the usage of the different runways during one day and is represented in the figure after folding the large 288-dimensional vector back into the more meaningful space, where the count of take-offs and landings is shown by the hour of the day for the different runways (represented by different colors). Each row corresponds to a cluster found by HDBSCAN, so represents a subset of the days of 2018. The first cluster is the *noise* cluster generated by HDBSCAN, other clusters are sorted by size.

Looking at the dominating colors throughout the day, it is quite clear that the clusters indeed collect days of similar runway use and, overall, runways 34, 14 and 28 are used predominantly.

[1] available, e.g., at https://mesonet.agron.iastate.edu/request/download.phtml?network=CH__ASOS

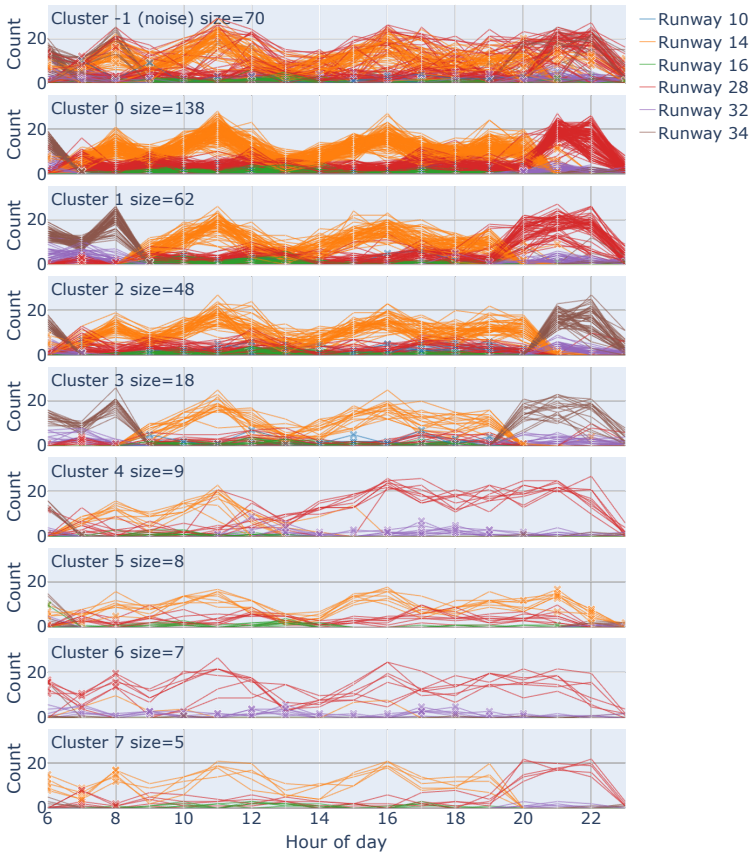

Figure 7: Visualizing the clusters of runway use found with HDBSCAN

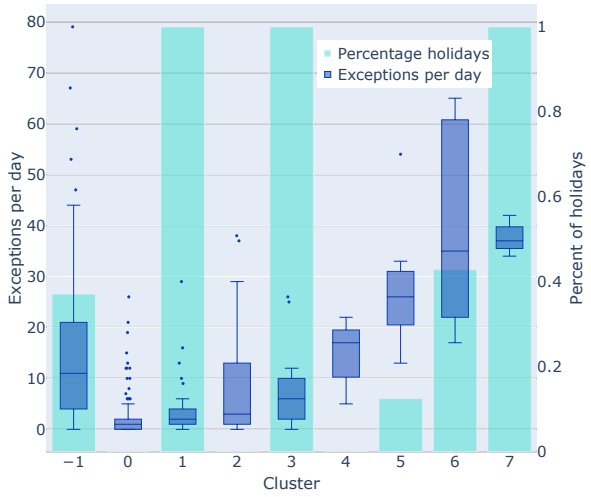

Figure 8: Showing the distribution of exceptions per day (blue box plot) and percentage of holidays (green bars) for the clusters found be HDBSCAN and shown in Fig. 7

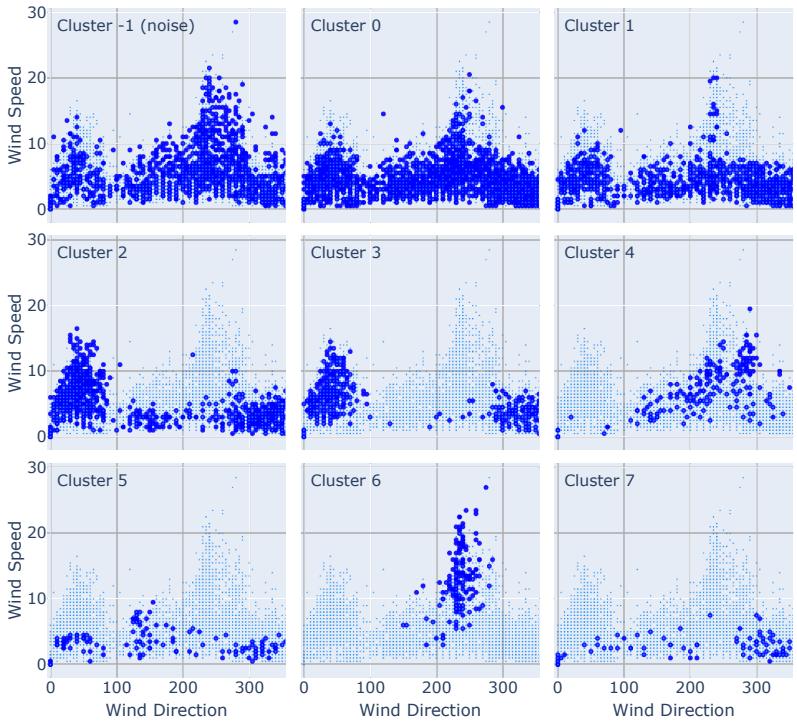

Figure 9: Wind speed and direction observations for days and hours part of the respective clusters

One can also see that cluster 0 relates to cluster 2, and similarly cluster 1 to 3, in the sense that the order of runway use is very similar, with the difference that morning and evening parts are longer in the respective second cluster of the pair. The prominent difference between the two pairs, on the other hand, is the runway use in the evening (runway 28 vs 34). The remaining clusters are smaller, showing different runway use patterns through the day.

The different operating concepts mentioned in Section 4.2.1 also mean that for each hour of the day certain runways are allowed respectively preferred under normal conditions. In Fig. 6 runway use that is not following the operating concepts is marked with an x-symbol. Looking at the clusters we find runway use outside of the operating concepts to a larger degree in smaller clusters 4-7, namely during the day in clusters 4 and 6, in the morning and evening hours in cluster 5 and in the morning hours only in cluster 7.

In the following figures, the available ground truth data is used for evaluation of these findings in the resulting clusters. Fig. 8 shows the percentage of holiday days in each cluster (green bars), indicating as expected that the clusters 0 and 2 correspond purely to weekdays, whereas clusters 1 and 3 correspond purely to weekends and holidays. This figure also shows as box plot the exceptions from the operating concepts per day for each cluster, where we see a growing part of exceptions for the smaller clusters.

Finally, Fig. 9 shows the wind situations observed for the hours and days in the different clusters. Whereas for the noise cluster and the 'normal' clusters 0 and 1 the wind direction and wind speed is quite mixed, we can see that for clusters 2 and 3 (as discussed before, mainly differing from 0 and 1 in their evening runway use) have more specific wind situations with

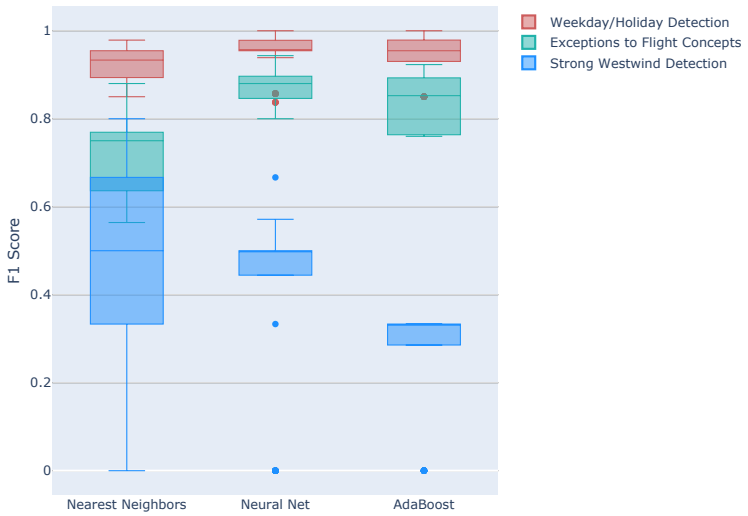

Figure 10: F1 Score calculated with different classifiers from `sklearn` (5-fold stratified with 20 runs each) for different detection topics

mostly northern winds. Most obvious in its specificity is the wind situation for cluster 6 — this clearly corresponds to the stronger west-southern wind leading to the exceptional use of runway 32 during the day.

In summary, clustering of the data results in clearly separated clusters of runway use that we can understand through the ground truth of operating conditions, the dependence on weekday versus holiday, and wind conditions.

4.2.5 Supervised learning

In this setting, we use the available ground truth as labels for supervised learning. As we have seen in the previous section in the unsupervised clustering, there are clear differences between weekdays and holidays, as well as well-observable differences when days have exceptions from the normal operational use concept. A weaker signal has been seen especially in cluster 6 in Fig. 9 corresponding to strong western wind situations. We use these three different labels in turn as binary classifiers, namely the binary *Weekday/Holiday* signal, the number of observed exceptions from the normal flight concept per day after thresholding (using 5 as threshold) (*Exceptions to Flight Concepts*) and the count of strong western winds per day (defined as wind situations with wind speed $\geq$ 15kt and wind direction in the range between 220° and 280°, with a threshold of 3 per day).

High-level results of the experiment are presented in Fig. 10 showing the *F1* score for the three classification tasks for selected classifiers of the Python `sklearn` [7] module (without further tuning of the classifiers). Different classifiers show different capabilities to correctly classify, but it becomes clear that the *Weekday/Holiday* difference can be learned quite well, the *Exception to Flight Concepts* is slightly harder to detect, but still has a high selectively, whereas the *Strong Westwind* situations are much harder to detect. This latter fact is not too surprising, as Fig. 9 hints that there is not such a clear connection between runway use and strong western winds, as these wind situations also can be found in clusters 0 and 1 and to

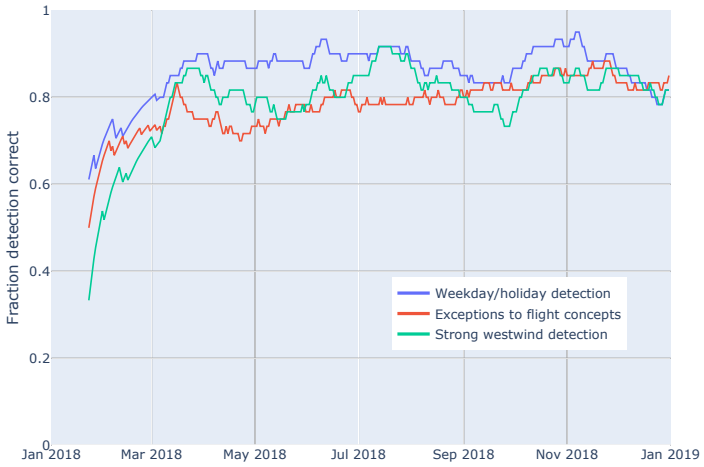

Figure 11: Detection success over time in 'historical' semi-supervised learning approach

some extent in cluster 4.

4.2.6 Semi-supervised learning experiments

After confirming with supervised learning that the considered signals are detectable quite well, we made initial tests for the idea of a "historical" semi-supervised learning approach for the different signals. We consider the setting that we train a neural network autoencoder model (for the method compare, e.g. [4]) by selected past observations, then on each day evaluate the reconstruction error for the new observation of that day, using a threshold to decide whether this new observation is anomalous compared to the selected signal or not. The threshold is here defined as the mean of the *past* reconstruction data of the training data multiplied by a factor 1.5. On the next day, it is decided whether the previous data point should be added to the training data or not, thereby extending the selected training data set over time with the expectation that a growing training data set should increase the detection rate over time.

We again used the three separate classification types for testing, namely detection of weekday/holiday in the data (which of course is trivial in reality, just serving for testing here), detection of a high level of exceptions compared to the normal operating concepts and detection of strong western wind situations.

Our initial findings are represented in Fig. 11 where for each day we find the success rate of determining the classification correctly for the new observation based on the previous training data set, in a moving 60-day window. It can be seen that in all cases the success rate starts at the level of random guessing, but then increases over time as more training data is modeled in the autoencoder, getting above 80% success rate, somewhat surprisingly, for all signals, though quite heavily fluctuating over time. We include these findings as our initial tests of work in progress, obviously more rigor and deeper analysis is required to further develop this idea.

5 Discussion

In this report we have discussed anomaly detection with their specific challenges in the context of aviation. We explored the topic of ground truth that is available for experiments and found that due to the very high complexity of the subject, it is often hard to evaluate results of a detection approach as the specific conditions during that region in time and space are often unknown except possibly to subject matter experts, who are unfortunately a scarce resource purely for experimenting.

We observed this in our work investigating regional air traffic patterns with the Gini index, as we were able to find and understand strong signals like the World Economic Forum in Davos and the Christmas days, but beyond this, other findings were very hard to attribute to specific conditions.

Therefore, we turned in our experiments to the more local setting of runway use at an airport, where more of the direct conditions are known, be it in the form of well-defined operating procedures or well-recorded local weather conditions. In this setting, it is possible to find stronger signals and patterns that are understandable, like finding the weekday/holiday patterns, exceptional use of runways, partially attributable to specific weather conditions, serving as test cases to tackle also other, more real-world questions in the future.

There are many opportunities for future work in this context: we would like to repeat our work on the runway use at the Zurich airport with other airports that are more complex, have a lower data visibility, and potentially a less well-defined usage pattern. We also would like to follow up on the idea of semi-supervised learning with neural network models, which we just started to extend by using LSTM-neural networks to capture the time component more explicitly in the modeling.

References

[1] Flughafen Zürich AG. Pistenbenutzungskonzepte. https://www.flughafen-zuerich.ch/~/media/flughafenzh/dokumente/das_unternehmen/laerm_politik_und_umwelt/pistenbenutzungskonzepte_2018.pdf, 2018. Accessed: 2019-11-08.

[2] Luis Basora, Xavier Olive, and Thomas Dubot. Recent Advances in Anomaly Detection Methods applied to Aviation. preprint, MATHEMATICS & COMPUTER SCIENCE, September 2019.

[3] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, July 2009.

[4] Thomas Dubot. Predicting sector configuration transitions with autoencoder-based anomaly detection. In *Proceedings of the International Conference for Research in Air Transportation*, June 2018.

[5] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), mar 2017.

[6] Josh Montague. STL Decompose. https://github.com/jrmontag/STLDecompose, 2017. Accessed: 2019-11-08.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Matthias Schäfer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias Wilhelm. Bringing up OpenSky: A large-scale ADS-B sensor network for research. In *Proceedings of the 13th international symposium on Information processing in sensor networks*, pages 83–94. IEEE Press, 2014.