# Using Big Data to improve Aircraft metadata (and NWP)

Bruce Ingleby[1], Micky Yun Chan[2] and Mohamed Dahoui[1]

[1] ECMWF, Reading, UK [2] Latvian University of life sciences and technologies

*bruce.ingleby@ecmwf.int*

## Background

Wind and temperature from commercial aircraft in AMDAR format form an important and growing input to Numerical Weather Prediction (NWP), Petersen (2016, BAMS). Unfortunately there is very little metadata in AMDAR reports and the aircraft identifier has been anonymized. Ideally we would like to know the aircraft type, the airline, the software (avionics) and processing used. Aircraft temperatures are typically biased high by up to about 1 K and NWP centres have to bias correct them based on the AMDAR identifier. The bias may depend on aircraft type (Drüe, 2008, QJ) although there are probably other factors as well.

Relatively recently ECMWF realized that a subset of wind directions from Boeing 787 aircraft have large errors (figure 1). We now have a partial list (over 500 B787 aircraft reporting AMDARs) but this has shown up our frustrating lack of metadata – resulting in somewhat unsatisfactory 'solutions', rejecting many good winds as well as bad ones, and not all the bad ones.

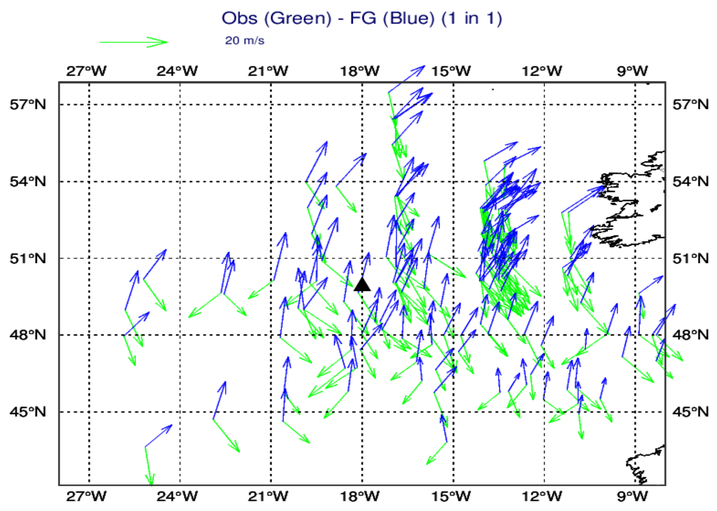


Figure 1. Observed (green) and forecast (blue) winds for suspect reports on 1 Jan 2019.
The B787 bug reverses the sign of the v component.

The 2019 ECMWF Summer of Weather Code (ESoWC) offered the opportunity to team up with an external computer scientist (MYC) and explore ways of using Big Data from flightaware and flightradar24 websites to match AMDAR identifiers to aircraft type, airline and tail-number (identifiers used by the aviation industry).

## Guess the airport

Only a minority of AMDAR reports include departure/arrival airports (as in Figure 2). In most cases the airports have to be estimated from the AMDAR positions. If these extend close to the ground then the airport can usually be determined with high confidence, if the reports stop at cruise level then the nearest airport (within 250 km) is usually chosen. Different lists of airport positions have been tried, with up to about 8000 airports globally (we plan to cut this down to those used by commercial airports, excluding eg heliports and military bases).

Figure 3 shows an example with estimated airports. We are fairly confident of the match (but less confident than with Figure 2). The black line shows long haul flights to/from Charles de Gaulle airport, Paris (CDG). The corresponding airports estimated from the AMDARs (red) show various airports within about 300 km of Paris – because ascents/descents at this end are not reported.

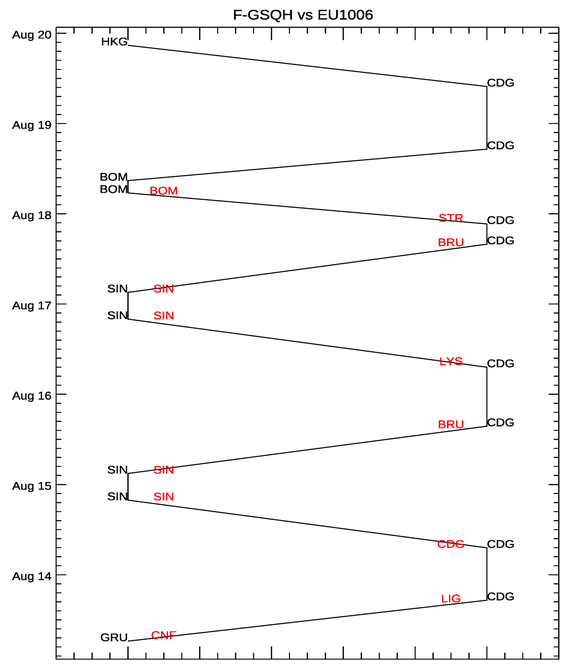


Figure 3. As figure 2 but with estimated airports in red.
CDG is Paris, SIN is Singapore and BOM is Mumbai.
F-GSQH is a Boeing 777.

## Online data vs AMDAR data

The information in black in the schematic Figure 2 summarises the data available for tail-number G-EZAS from flightradar24 for a week in August 2019: it has times of take-off and landing and a three letter code for each airport (EDI is Edinburgh). The information in red is the time of the first and last AMDAR in each flight and (in this case) the departure and arrival airport reported in the AMDARs. There are some flights without AMDARs, but for those with AMDARs the airports all match giving confidence that this AMDAR system (EU0001) is on this aircraft (tail number G-EZAS, type Airbus 319).
Splitting the AMDAR reports into "flights" requires some assumptions. If there is no report for an hour this may indicate that the aircraft is on the ground – but some long-haul flights have in-air gaps of more than an hour and some short-haul flights have airport stops of less than an hour. One of the main complications for us is that AMDARs can start or stop in mid-air without providing profiles at one or both airports.

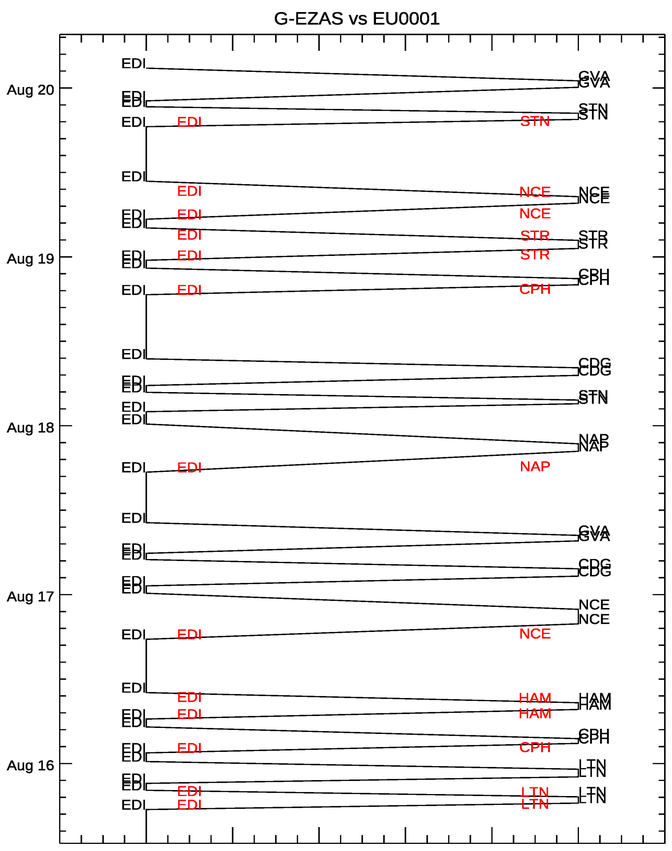


Figure 2. Schematic of aircraft movements for about a week, with date/time on y-axis.
**Black: online data with IATA airport codes.**
**Red: AMDAR data with reported airport codes.**

## Results

Early attempts at matching single flights were very time consuming (because of the number of combinations to try) and generated many spurious matches and few good ones. Trying to match over a period of 5-7 days proved more fruitful.

First one needs to know (or guess) an airline that provides AMDAR reports, then for all the tail numbers in that fleet obtain summary listings of the recent flights from the internet. These summary listings (as in the black data of figures 2 and 3) are then compared with summaries for all 'likely' AMDAR identifiers. Preliminary results produced in the second half of August 2019 include:

- USA program 1062/7105 identifiers matched: 15%
- Europe: 505/1186 matched
- Japan: 142/252 matched … 6 other national/regional programs

Excluding the USA about 50% of the AMDAR identifiers were matched with moderate confidence, this ranged from 19/19 for South Korea (they use tail numbers as AMDAR identifiers!) to 0/12 for Canada. Given more data/work the proportion will increase, but how to combine/cross-check results for different periods will need to be addressed. There are various tunable parameters in the matching and different values work better for different airlines.

*Overall matching AMDAR identifiers with online data has proved more difficult than originally expected, but useful progress has been made.*

## Future

As well as some improvements to the matching procedure as above we hope to :

- Use the metadata to try to improve ECMWF monitoring, quality control and bias correction of aircraft data
- Convince the data producers to provide more metadata
- Make the case that the anonymisation of AMDAR identifiers is outdated given the existence of such online resources (particularly important given the emergence of new aircraft data, such as MODE-S, which may have completely different identifiers for the same aircraft).