Clustering Aviation Accidents

Sandy Preiss

# Background

The National Transportation Safety Board (NTSB) provides a wealth of data on aviation accidents in the US and international waters. However, with dozens of variables as well as narrative descriptions, it is difficult to get an overall sense of the different types of aviation accidents that occur. For this brief, we used NTSB data to group accidents into clusters and develop initial lists of high-level aviation accident types.

# Analysis

We used two rounds of k-means clustering to identify natural groups in the data. First, we used all the structured data in the NTSB database to cluster accidents. Next, we used the narrative descriptions of each accident and its probable cause to develop a second list of clusters. The tables below list the clusters and some general interpretation of their key features.

**Table 1: Clusters Generated from Structured Variables**

|  |  |  |
| --- | --- | --- |
|  | **Overall Description** | **Key Variables** |
| 1 | Common Personal Aviation (post-2000) | FAR description, Purpose, Lat/Lon (recorded post-2000) |
| 2 | Common Personal Aviation (pre-2000) | FAR description, Purpose, Lat/Lon (not recorded pre-2000) |
| 3 | Air carrier accidents | Damage (minor), Engine (turbo fan), Total involved (many) |
| 4 | Very severe, over-ocean commercial | Country (not US), Damage (destroyed), Max injury (fatal) |

**Table 2: Clusters Generated from Narrative Text**

|  |  |  |
| --- | --- | --- |
|  | **Overall Description** | **Top Terms** |
| 1 | Common personal aviation | engine power airplane pilot loss carburetor landing force oil fuel |
| 2 | Very severe | pilot airplane flight foot airport condition time impact terrain passenger |
| 3 | Fuel-related, severe | fuel tank engine pilot power airplane gallon landing flight selector |
| 4 | Landing-related, not severe | airplane runway pilot landing gear wind left right control nose |
| 5 | Abbreviations used | acft plt rwy ft eng flt may landing provide gear |
| 6 | Helicopter accidents | helicopter rotor pilot tail blade autorotation flight main hover ground |

The structured data and narrative text both separate more- and less-severe accidents. Beyond that, they offer different perspectives on the natural groupings of accidents.

The structured data separates most personal aviation accidents from air carrier and commercial accidents. It further separates the small group of very severe commercial accidents that happened in international waters. The separation of personal aviation accidents into pre- and post-2000 is largely because the latitude and longitude of accidents was not recorded prior to 2000, and thus is not very useful.

The text clusters provide more nuance on the cause of the accident. In addition to separating severe from common, the narrative data produces a cluster related to running out of fuel, which unsurprisingly tends to result in severe accidents. It also produces a cluster of accidents related to landing, which tend to be less severe. Another cluster is mostly driven by the use of abbreviations in the narrative, which is not very useful.

# Conclusions

The most important factors distinguishing high-level types of aviation accidents are purpose (personal vs. commercial), severity, and cause. Narrative and probable cause descriptions are essential for determining the cause of accidents.