Contrast-Set Mining of Aircraft Accidents and Incidents

Zohreh Nazeri, Daniel Barbara, Kenneth De Jong,
George Donohue, and Lance Sherry

George Mason University, 4400 University Drive,
Fairfax, Virginia, 22030, USA
{znazeri,dbarbara,kdejong,gdonohue,lsherry}@gmu.edu

Abstract. Identifying patterns of factors associated with aircraft accidents is of high interest to the aviation safety community. However, accident data is not large enough to allow a significant discovery of repeating patterns of the factors. We applied the STUCCO[1] algorithm to analyze aircraft *accident data* in contrast to the aircraft *incident data* in major aviation safety databases and identified factors that are significantly associated with the accidents. The data pertains to accidents and incidents involving commercial flights within the United States. The NTSB accident database was analyzed against four incident databases and the results were compared. We ranked the findings by the *Factor Support Ratio*, a measure introduced in this work.

Keywords: contrast-set mining, aviation safety, data mining, aircraft accident analysis, aircraft incident analysis, knowledge discovery.

1 Introduction

An aircraft *accident* is an occurrence associated with the operation of an aircraft in which people suffer death or injury, and/or in which aircraft receives substantial damage; an *incident* is an occurrence which is not an accident but is a safety hazard and with addition of one or more factors could have resulted in injury or fatality, and/or substantial damage to the aircraft [1]. Previous research on aircraft accidents has focused on studying accident data to determine factors leading to accidents. In his Why-Because Analysis (WBA) [2] to understand involving causal factors to accidents, Ladkin aims to reveal the causal reasoning behind the events and circumstances leading to an accident. He applied his method to individual aircraft accidents to show how it can improve understanding of the factors involved in those accidents [3]. Dimukes [4] studied 19 airline accidents focusing on pilot errors; his study showed characteristics and limitations of human cognition in responding to different situations and suggested accidents are caused by confluence of multiple factors. Van Es [5] studied Air Traffic Management (ATM) related accidents worldwide and showed *flight crew* is a more important factor in ATM-related accidents than *air traffic control* is. He also

[1] STUCCO algorithm is developed by S. D. Bay and M. J. Pazzani, University of California, Irvine.

P. Perner (Ed.): ICDM 2008, LNAI 5077, pp. 313–322, 2008.

reported no systematic trends were found in the accident dataset when performing a trend analysis. While these studies help understanding individual accidents and their causal factors, the low rate of accidents however, makes it difficult to discover repeating patterns of these factors.

Other research has analyzed larger sets of data available on incidents to determine the causal factors of incidents. Majumdar [6] applied log-linear modeling technique to analyze seven factors involved in loss-of-separation incidents. Hansen and Zhang [7] tested the hypothesis that adverse operating conditions lead to higher incident rates in air traffic control. NASA [8] studies voluntarily submitted incident reports, mostly by pilots, and publishes the results monthly. While studying incident data is helpful to understand incident causal factors, it does not identify the relationship between the incident factors and accidents. Since the ultimate goal of studying aviation safety data is to reduce accidents, in this research, we analyzed both accident and incident data to show the relationships between the two classes of events and to identify factors that are significantly associated with accidents.

2 Data

The data used in the study consists of accidents and incidents pertaining to commercial flights (part-121) from 1995 through 2004. The accidents were obtained from:

- National Transportation Safety Board (NTSB) database, containing reports of all accidents

The incidents were obtained from four major national databases:

- Federal Aviation Administration Accident and Incident Database System (FAA/AIDS), containing reports of incidents investigated and/or documented by the FAA
- National Aeronautics and Space Administration Aviation Safety Reporting System (NASA/ASRS), containing self-reported errors voluntarily submitted mostly by pilots
- FAA Operational Errors and Deviations (OED), containing mandatory reports of Air Traffic Control errors
- FAA System Difficulty Reports (SDRS), containing reports of mechanical problems with the aircraft system or components

Each report in these databases consists of structured fields plus an unstructured narrative explaining the event. In this study we used the structured fields only. The structured fields contain causal and contributory factors which are identified either by the person reporting the event or by a domain expert who has reviewed the report afterwards. Our analysis used these factors.

2.1 Data Constraints

Some constraints imposed by the data need to be considered. All accidents in the United States involving civil aircraft are investigated by the National Transportation

Safety Board (NTSB), an independent organization, and are reported in the NTSB database. Accident data, therefore, can be assumed complete and free of bias. Incident data however, are under-reported and subject to self-reporting bias. To address these constraints, our study analyzes the underlying factors of accidents and incidents *qualitatively* (and not a quantitative analysis such as regression). The historical data on incidents is large enough to represent these factors qualitatively. Also, we consider *all* factors that have been present in an event, regardless of their primary or contributory role in leading to the event. This minimizes the impact of the bias in reporting the factors.

2.2 Data Selection

Since the purpose of the analysis is to identify operational factors under normal conditions, accidents and incidents due to the following causes were filtered out from the data:

- passenger and cabin-crew related problems, such as passengers being injured due to hot coffee spilling on them
- medical and alcohol related events, such as pilot being sick
- terrorism and security events, such as bomb threats
- bird/animal strike, such as aircraft encountering a deer on the runway
- events during the phases of operation when the aircraft is not operating (parked, standing, preflight)

Also, reports pertaining to the Alaska region were excluded since flight environment and procedures in this region are different from the rest of the regions in the United States and require a separate study.

After applying the filters, there were 184 accidents, and the following sets of incidents left in the data for analysis: 2,188 reports in the AIDS dataset, 29,922 reports in the ASRS dataset, 10,493 reports in the OED dataset, and 85,687 reports in the SDRS dataset.

2.3 Data Preparation

We first normalized the data across the databases and then developed an ontology by developing a hierarchy of factors and sub-factors common across the databases.

Normalization of the values was needed so that all databases use the same term to refer to the same factor or condition. For example, the action where pilot executes a maneuver to avoid an object on the runway is referred to by one database as 'ground encounter' and by another as 'object avoidance'.

Eight high-level categories of factors were identified in the data, each containing corresponding sub-factors. These factors and examples of their sub-factors are shown in Table 1. The 'Other' category contains all sub-factors which didn't fit under the other seven categories and were not big enough to have their own separate category.

We transformed the reports into vectors consisting of fields that indicate presence or absence of each of the common factors and sub-factors in the event (accident or incident). We then analyzed these vectors.

Table 1. Common ontology across multiple databases

Factor	*Sub-Factor examples*
Aircraft	Engine, Flight control system, Landing gear
Airport	Snow not removed from runway, Poor Lighting, Confusing marking
Air Traffic Control	Communication with pilot, Complying with procedures
Company	Procedures, Management, Training
Maintenance	Compliance, Inspection
Pilot	Visual lookout, Altitude deviation, Decision/Judgment
Weather	Wind, Thunderstorm, Ice
Other	Factors not in the other categories; FAA oversight, Visibility

3 Analysis

We applied the STUCCO algorithm [9] to perform four sets of analyses. In each analysis, the accident vectors were paired with incident vectors from one of the four

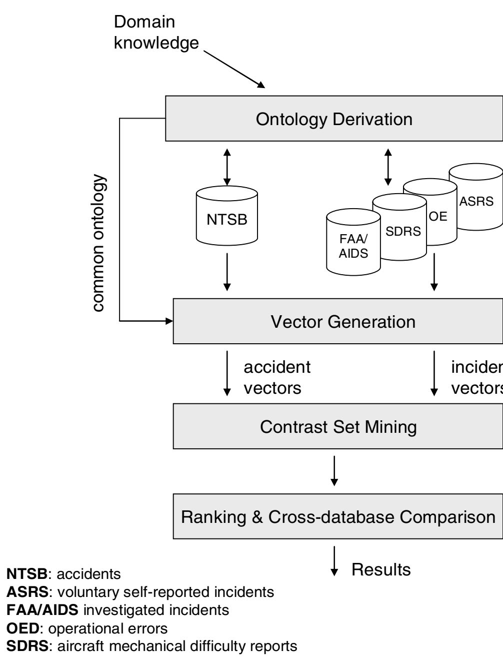

Fig. 1. Analysis of accident data in contrast to incident data

incident databases. Each analysis identified patterns of factors which are significantly associated with accidents (or with incidents). We ranked the findings of each analysis, using the *Factor Support Ratio* measure described below. Final results of the four analyses were compared at the end. Figure 1, depicts the analysis process.

3.1 Algorithm

The STUCCO algorithm finds conjunctions of attribute-value pairs that are significantly different across multiple classes. In the case of our data, there are two classes: accidents and incidents. Attribute-values are binary values (1 or 0) for the factors in each event vector, implying presence or absence of the factors in that event. Figure 2 shows the algorithm used in our study.

In an A-Priori-like process [10], the factors and their children are examined for their support in each class. For each factors-set, *deviation* is calculated as absolute value of the difference between accident support and incident support for the

```
Input accident and incident vectors
C = set of factors in the input vectors
D = set of deviations, initially empty

1. While C is not empty
2.   Scan input data and count support ∀c ∈ C
         supp_acc = (accidents containing the
                          factor/total accidents)*100
         supp_inc = (incidents containing the
                          factor/total incidents)*100
3.   For each factor-set c ∈ C :
4.    If (count_acc > min cell frequency AND
          count_inc > min cell frequency)
5.      If ( |supp_acc - supp_inc| > dev_min )
          then factor-set is large
6.        If (Chi Square test passed)
            then factor-set is significant
                 Add factor-set to candidates D
7.        Generate children (factor-set, C)
8.        For each child
            If (supp_acc> dev_min OR supp_inc> dev_min)
              Then add child to C'
9.   C = C'

10. Rank candidates in D by Factor Support Ratio
```

Legend:
$supp_{acc}$= accident support
$supp_{inc}$= incident support
dev = deviation

Fig. 2. STUCCO algorithm used in this research

factor-set. Factor-sets with a *deviation* of more than a minimum threshold are tested for the statistical significance of their distribution over the two classes. Chi Square test is used to perform the test. Factor-sets whose test results in a p-value of more than 0.05 are rejected, the rest are added to the list of candidates whose children will be generated and go through a similar test. The contingency table shown in Table 2 is used for the Chi Square test.

The step shown in Figure 2 are slightly different than the original algorithm discussed in [9]. The difference is in step 7 in Figure 2. In this step, the original algorithm generates children for a factor-set if the factor-set is both *large* and *significant.* Here the significance criterion is relaxed for child generation. (Note that this criterion is relaxed only for child generation, passing the significance test is still required for a factor-set to be added to the candidates.) The reason for this modification is to allow for discovery of factors that might not be individually associated with accidents, but if combined together they could be significant accident factors. Discovery of such cases is one of the objectives of the analysis. This modified step generated two additional significant two-factor-sets whose individual factors did not pass the significance test individually.

Table 2. Contingency table used for Chi Square significance test

	accidents	**incidents**
factor-set true	accidents containing the factor-set	incidents containing the factor-set
factor-set false	accidents not containing the factor-set	incidents not containing the factor-set

3.2 Ranking

Once significant factor-sets are identified by the algorithm, we rank them by their *Factor Support Ratio* measure. We calculate the Factor Support Ratio for each factor-set as the probability of the factor-set given an accident, divided by the probability of that factor-set given an incident, or the ratio of the factor-set's support in accident dataset over its support in the incident dataset (1) where F = factor-set, acc= accident, inc=incident, P(F|acc) = probability of factor-set given an accident, #Facc = number of accidents containing factor F, #acc = total number of accidents.

$$\begin{aligned} \text{Support Ratio} &= \frac{P(F \mid acc)}{P(F \mid inc)} \\ &= \frac{P(acc \mid F)P(F)/P(acc)}{P(inc \mid F)P(F)/P(inc)} \\ &= \frac{\#Facc/\#acc}{\#Finc/\#inc} = \frac{\%Facc}{\%Finc} \end{aligned} \tag{1}$$

$$\text{Support Ratio} = \frac{Support_{accident}}{Support_{incident}}.$$

The information conveyed by the Support Ratio measure about the factor-set is different than that of the *deviation* that is used in the algorithm. Deviation is the difference between the factor-set's accident and incident supports. Support Ratio is the probability of a factor-set being involved in an accident divided by its probability of being involved in an incident. To see the significance of this distinction, consider factor-sets A and B and their corresponding measures in Table 3.

Table 3. Support Ratio vs. deviation

factor-set	***accident supp***	***incident support***	***Dev***	***Support Ratio***
A	60%	50%	10%	1.2
B	11%	1%	10%	11

Both factor-sets A and B have a deviation of 10% between their accident support and incident support. However, in the case of factor-set B, the support in accidents is 11 times more than in incidents. This can be interpreted as: occurrence of factor-set B in an accident is 11 times more likely than its occurrence in an incident. This is a more distinctive distribution than that of factor-set A which has a Support Ratio of 1.2. We can use this measure to compare factor-sets A and B, and say factor-set A is more likely to be involved in accidents than factor-set B.

4 Results

Results of the analyses were reviewed with domain experts, some results were consistent with previous research findings and some were interesting in the sense that previous studies had not identified them. Highlights of the results are discussed below.

Company factors - factors such as mistakes by the company (or airline) personnel, and inadequate or lack of procedures by the company for performing a task – were consistently the highest ranked category of factors associated with accidents among the eight high-level categories of factors. This was an interesting result. Although previous studies had shown these factors contributed to accidents, their significance relative to other factors was not shown. Our research conducted a holistic study of the factors across multiple databases and in addition to identifying the factors associated with the accidents we could rank the factors in the order of their significance.

The next highest ranked accident factors were *Air Traffic Control (ATC)* followed by the *pilot* factors. Among the *ATC* factors, *communications* sub-factor had the highest rank of association with accidents. And among the *pilot* factors, *visual lookout* had the highest rank. Identification of *ATC communications* and *pilot visual lookout* as accident factors was consistent with previous findings. The interesting finding was that *ATC* factors which are less frequent than *pilot* factors were ranked higher. This implies that although *ATC* factors are less frequent but once they occur there is a high risk of having an accident (as opposed to an incident). *Pilot* factors are more frequent than other factors in accidents but they are also more frequent in incidents, which makes their Support Ratio lower and ranks them after the *company* and *ATC* factors.

Another interesting finding was association of *aircraft* factors with incidents. *Aircraft* factors are mechanical problems with the aircraft system or components, such as landing gears and flight control systems. The results showed these factors are more likely to be involved in incidents except when combined with other factors such as severe weather or pilot errors.

In Table 4 we show the results grouped by the factor category. These results are associations that were consistently identified by multiple analyses. Additional associations were identified by each individual analysis.

Table 4. Selected results of the analyses

Factor Category	***Associations***
Pilot	(pilot, airport, other) → accident (pilot, weather)→ accident (pilot) → accident
ATC	(ATC, pilot, airport, other) → accident (ATC, airport, company) → accident (ATC) → accident
Aircraft	(aircraft, weather) → accident (aircraft) → **incident**
Company	(company, maintenance, other) → accident (company, maintenance) → accident (company) → accident

Ranking of the results also showed that likelihood of a factor being involved in an accident rises as more factors co-occur with it. This means when multiple factors are present, there is a higher likelihood of having an accident (as opposed to having an incident). Tables 6 and 7 show some examples. For example in Table 6, the Support Ratio for combination of *pilot+airport* factors is 7.2 compared to the Support Ratio of 3.9 for the *pilot* factors, signifying that *pilot* factors combined with *airport* factors are more likely to result in accidents than the *pilot* factors alone.

Table 5. Ranking of results from NASA database analysis

factor-sets in NASA database	**Support ratio**
pilot, aircraft, company, other	3.7
pilot, company, other	3.6
pilot, aircraft, weather	2.9
pilot, airport, other	2.3
pilot, weather	1.9

Table 6. Ranking of results from FAA database analysis

factor-sets in FAA database	Support ratio
pilot, airport, other	14.3
pilot, aircraft	9.7
pilot, airport	7.2
pilot, weather	4.3
pilot	3.9

Note that the Support Ratio measure cannot be used for cross-database comparison of factor-sets. Factor-sets within a dataset can be compared using their Support Ratios since total numbers of accidents and incidents are the same in calculation of the Support Ratios.

5 Summary and Future Work

By applying contrast-set mining to the aviation safety data, we were able to analyze aircraft accident data in contrast to the incident data and identify patterns of factors which are associated with the accidents. Our ranking measure, the Factor Support Ratio, allowed ranking of the findings and identification of relative significance of the factors in contributing to accidents, compared to other factors.

This work analyzed aircraft accidents and incident pertaining to commercial flights within the United States. The methodology used here could be applied to the general aviation as well. The analysis could be extended to include world-wide safety events. In a future work, other data attributes, such as severity of the event, phase of flight, and type of aircraft could be included in the study to obtain more specific results.

Acknowledgements

Zohreh Nazeri wishes to thank the following individuals for their help: Christina Hunt, manager, at the FAA Aviation Safety Information Analysis and Sharing (ASIAS), provided much of the data needed for the study and patiently answered any questions about the data. Linda Connell, director, and her team, at the Aviation Safety Reporting System (ASRS) program at NASA, helped with downloading the ASRS data from their on-line query interface and explained the data. My colleagues at the MITRE Corporation generously shared their knowledge with me in multiple discussions.

References

1. Federal Aviation Administration, Air Traffic Organization, Aircraft Accident and Incident Notification, Investigation, and Reporting. Order 8020.16, `http://www.faa.gov/airports_airtraffic/air_traffic/publications/at_orders/media/AAI.pdf`

2. Ladkin, P.: ATM Related Accidents. Eurocontrol (2006), `http://www.eurocontrol.int/corporate/public/standrd_page/cb_safety.html`
3. Ladkin, P.: Causal Reasoning About Aircraft Accidents. In: 19th International Conference on Computer Safety, Reliability and Security (SAFECOMP), Rotterdam, The Netherlands (2000)
4. Dimukes, K.: The Limits of Expertise: The Misunderstanding Role of Pilot Error in Airline Accidents. ASPA/ICAO regional seminar (2005)
5. Van Es, G.: Review of Air Traffic Management-Related Accidents Worldwide: 1980-2001. In: Fifteenth Annual Aviation Safety Seminar (EASS), Geneva, Switzerland (2003)
6. Majumdar, A., Dupuy, M.D., Ochieng, W.O.: A framework for the Development of Safety Indicators for New Zealand Airspace: the Categorical Analysis of Factors Affecting Loss of Separation Incidents. In: Transportation Research Board (TRB) annual conference (2006)
7. Hansen, M., Zhang, Y.: Safety Efficiency: Link between Operational Performance and Operation Errors in the national Airspace System. Transportation Research Record, journal of Transportation Research Board 1888, 15 (2004)
8. National Aeronautics and Space Administration, Air Traffic Management System (2007), `http://quest.arc.nasa.gov/aero/virtual/demo/ATM/tutorial/tutorial1.html`
9. Bay, S.D., Pazzani, M.J.: Detecting Change in Categorical Data: Mining Contrast Sets. In: Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. The Association for Computing Machinery, New York (1999)
10. Agrawal, R., Srikant, R.: Fast Algorithm for Mining Association Rules. In: Twentieth International Conference on Very Large Databases, VLDB (1994)