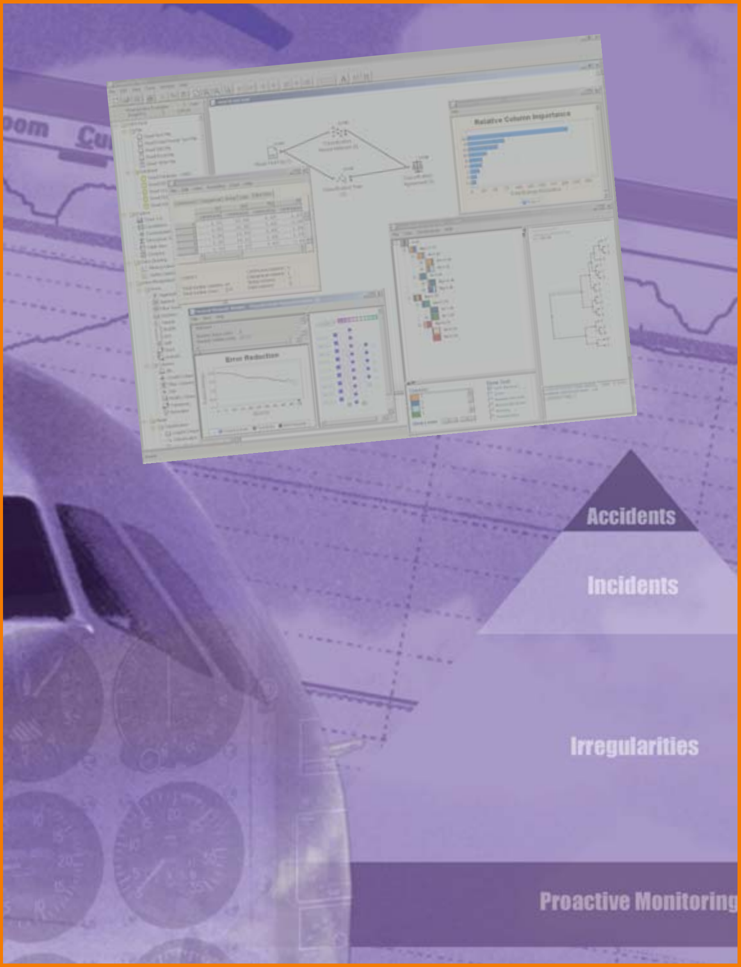


Application of Insightful Corporation®'s Data Mining Algorithms to FOQA Data at JetBlue Airways

A Technology Demonstration

In Partnership with the
Federal Aviation Administration and the
Global Aviation Information Network
(GAIN)

Report Prepared by:

Dr. Bob Treder
Senior Consultant
Insightful Corporation

Dr. Bill Craine
Pilot and FOQA Analyst
JetBlue Airways

jetBlue
AIRWAYS

Insightful®
intelligence from data

December 2004

Global Aviation Information Network
ENHANCING AVIATION SAFETY THROUGH SHARING
GAIN

Disclaimers; Non-Endorsement

INSIGHTFUL CORPORATION AND GLOBAL AVIATION INFORMATION NETWORK, ON BEHALF OF THEMSELVES AND THEIR SUPPLIERS AND CONTRIBUTORS TO THIS REPORT, HEREBY DISCLAIM ANY AND ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. ALL DATA AND INFORMATION IN THIS DOCUMENT ARE PROVIDED "AS IS."

The views and opinions expressed in this document do not necessarily reflect those of the Global Aviation Information Network or any of its participants, except as expressly indicated.

Reference in this document to any commercial product, process, or service by trade name, trademark, servicemark, manufacturer, or otherwise, does not constitute or imply any endorsement or recommendation by the Global Aviation Information Network or any of its participants (e.g., FAA) of the product, process, or service.

Notice of Right to Copy

This document was created primarily for use by the worldwide aviation community to improve aviation safety. Accordingly, permission to make, translate, and/or disseminate copies of this document, or any part of it, is freely granted provided each copy or partial copy clearly contains the following legend:

> Copyright © 2005 Insightful Corporation and JetBlue Airways, Inc. All rights reserved. Reprinted by permission from the Global Aviation Information Network.

Permission to make, translate, and/or disseminate derivative works of this document, or any part of it, is freely granted provided each such work acknowledges this work as the source of the derivative work, using the following citation and legend:

Citation:

> Insightful Corporation and JetBlue Airways, Inc., *Application of Insightful Corporation®'s Data Mining Algorithms to FOQA Data at JetBlue Airways: A Technology Demonstration In Partnership with the Federal Aviation Administration and the Global Aviation Information Network (GAIN)* (2005).

Legend:

> This work is created and distributed by permission from the Global Aviation Information Network.

If the document is translated into a language other than English, the notice must be in the language to which translated.

For further information on this Technology Demonstration

Dr. Bob Treder
Senior Consultant
Insightful Corporation
1700 Westlake Ave. N., Suite 500
Seattle, WA, 98109 USA
bob@insightful.com
+1 206-802-2231
www.insightful.com

Mr. Andy Muir
GAIN Program Office
Federal Aviation Administration / FSAIC
800 Independence Avenue, SW
Washington, DC 20591 USA
andy.muir@faa.gov
+1-202-267-9180
www.gainweb.org

Table of Contents

Acknowledgements

This project was funded by the US Federal Aviation Administration, Office of System Safety (known as "ASY" within the FAA) to facilitate the application of advanced methods and tools in the analysis of aviation safety data with the goal of improving aviation safety industry-wide. The project also involved the support and guidance of the Global Aviation Information Network (GAIN), an industry-led international coalition of airlines, manufacturers, employee groups, governments (including FAA) and other aviation organizations formed to promote and facilitate the voluntary collection and sharing of safety information by and among users in the international aviation community to improve aviation safety. Specifically, the project was guided by principles developed by GAIN's Working Group B, "Analytical Methods and Tools" (WG B), which has been tasked with fostering the use of existing analytical methods and tools and the development of new tools that elicit safety information out of aviation data.

Insightful Corporation® acknowledges that this report could not have been completed without the valuable support of: a number of staff from JetBlue including Bill Craine, Vince Zaccardi, Randy King, Tim Tatem, Cindy LoCicero, and Gemith Gemparo; FOQA software supplier SAGEM, in the person of Laurent Bloch; staff at Insightful Corp. including Peter Shaw, Jagrata Minardi, Lou Bajuk and Bob Treder; and Andy Muir, from the FAA's Office of System Safety (ASY) and a member of GAIN's WG B. Special thanks goes to Bill Craine who provided many comments and suggestions, reviewed analyses and documents and answered numerous questions between flying, vacations and house painting; to Laurent Bloch who spent countless hours extracting data from the AGS database and re-extracting it when data validity issues surfaced; and to Peter Shaw whose dexterity with Insightful Miner™ and S-PLUS® software helped tease out results that significantly enhance this report.

This page intentionally blank

Executive Summary

Insightful Corporation® applied data mining software (both Insightful Miner™ and S-PLUS®) to the analysis of JetBlue's flight operations quality assurance (FOQA) digital flight data to provide guidance on tools which may be useful in enhancing the current analysis of airline digital fight data. This report summarizes the findings. An additional report, the Analysis Report, provides a more thorough report of analyses completed. The findings are summarized as follows:

1. Current FOQA analysis methods provide useful analysis information that has improved management of flight safety, operations, and training programs. However, current methods typically require single-issue queries and some analyst formatting of outputs. Data mining techniques show promise in greatly improving efficiencies in the current methods by automating the query and output process.
2. All data should be thoroughly screened for errors and unrecorded and missing values before any serious analysis begins. Useful tools include summary statistics (e.g., minimum, maximum, mean, and standard deviation), histograms, and sequential observation plots. Outliers should be carefully examined to verify values.
3. Trellis (conditional) graphics (e.g., barplots, boxplots and scatterplots) are powerful tools for comparing groups.
4. When there are many points on a scatterplot, hexbinned scatterplots provide additional detail about point (observation) density and can reveal bimodal distributions completely missed in scatterplots.
5. Custom graphics provide a powerful context for understanding "hidden" relationships. The ability to graph how you want it and when it (with respect to data context) can help answer difficult questions and verify data in the context discovering a new relationship.
6. Principal components may be a useful tool for defining events which are composites of continuous parameters. Using principal components to characterize correlation structure can aid in understanding how fight parameters relate to each other and atypical flights.
7. Clustering methods coupled with heatmap displays provide a visual tool for understanding correlation structures in data.
8. Tree-based models provide an intuitive modeling structure for understanding how flight events or actions relate to flight parameters. They also provide a way to quantify variable importance.
9. Neural networks are not a very useful for understanding which parameters are associated with uncommon events or actions. For example, a neural network model could not even classify one go around correctly when go arounds constituted only about 1.5% of the flights.

This page intentionally blank

1 Introduction

One of the goals of the FAA Office of System Safety (ASY) is to identify and evaluate methods and tools not previously applied to aviation safety data with the goal of improving aviation safety industry-wide. One of the most interesting developments in information management is the proliferation of "data mining" analysis tools. One definition for data mining is "the process of discovering hidden patterns and relationships in data." Emphasis placed on automated learning from data, as the mining tools find interesting patterns without humans asking the initial queries. However, subject matter expertise from a human is always needed to review the initial results from a data mining project, to determine which of the patterns and relationships are of real value, and which are "commonly understood" (e.g. increased stopping distance in wet conditions) or otherwise not valuable.

This subject matter expertise is particularly valuable when designing and setting up the data mining procedure, through the definition of suitable "target" variables, and by gathering appropriate data. For example, one might wish to design a data mining work flow to determine whether the failure of certain mechanical parts are preceded by any early warning signs, or are associated with particular maintenance or operator activities.

Data mining includes various types of analysis, such as the following:

Classification	*Predicting a category for an example*
Regression	*Predicting a numeric value for an example*
Cluster Analysis	*Grouping examples with similar characteristics*
Association/Sequence Discovery	*Finding examples that occur together frequently*
Anomaly Detection	*Identifying unusual examples*
Exemplar-Based Learning	*Classifying based on similarity to previous*

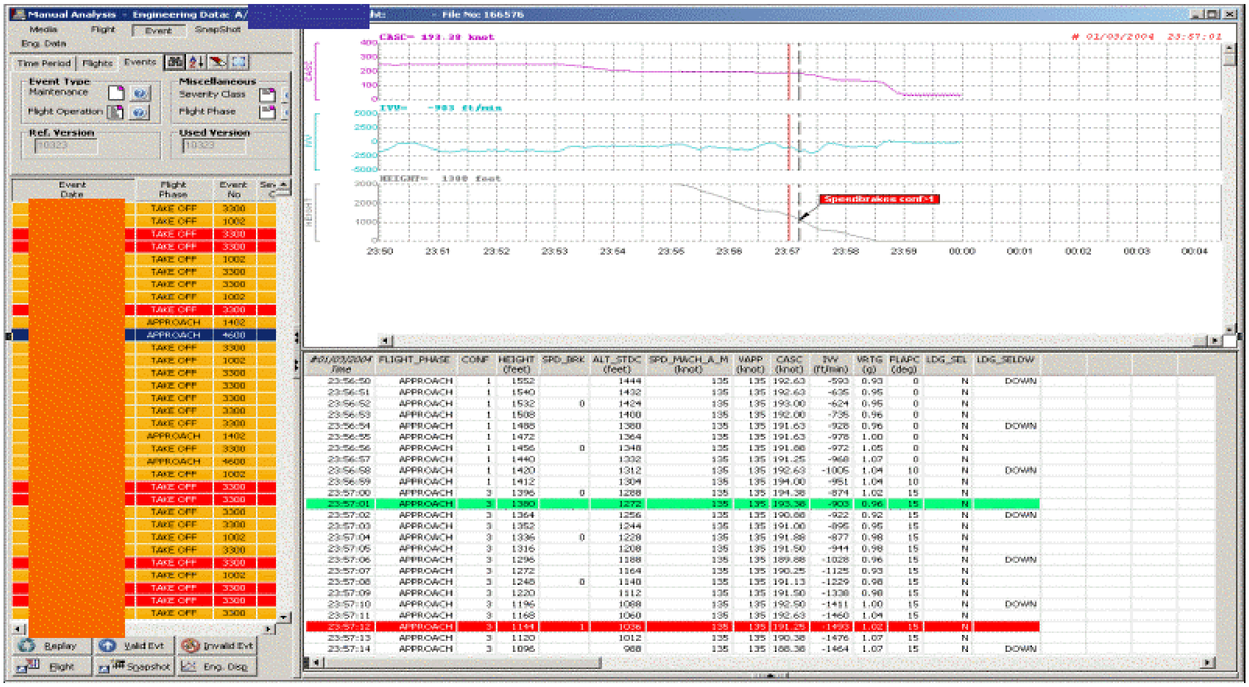
ASY believes that airlines could learn more about their operations by expanding the types of analyses provided by their current FDM programs. If data mining algorithms could be applied to this digital data, safety analysts might identify patterns and trends in the operation of their aircraft that were not previously known.

1.1 Background -- Current Airline Analysis of Digital Flight Data

This section describes the typical kinds of analysis of digital flight data currently performed at many airlines, including JetBlue. It describes the basic data set and two types of analysis currently performed: event exceedance analysis and flight performance distribution analysis.

1.1.1 The Basic Data Set

On each flight, over 300 parameters are recorded as time series. Parameters may be either numeric (airspeed) or categorical (autopilot on/off). The AGS system can display the series either as charts (3 traces shown in fig 1) or as tabular data (13 columns shown in Figure 1). Files are initially identified by flight number, city pair, and date. The system assigns a file number to each flight. After a period of time the flights are de-identified to mask the day and time of each flight.



Event Date	Flight Phase	Event No
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	1002
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	1002
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	1002
[redacted]	TAKE OFF	3300
[redacted]	APPROACH	1402
[redacted]	APPROACH	4600
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	1002
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	APPROACH	1402
[redacted]	TAKE OFF	3300
[redacted]	APPROACH	4600
[redacted]	TAKE OFF	1002
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	1002
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	1002
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300
[redacted]	TAKE OFF	3300

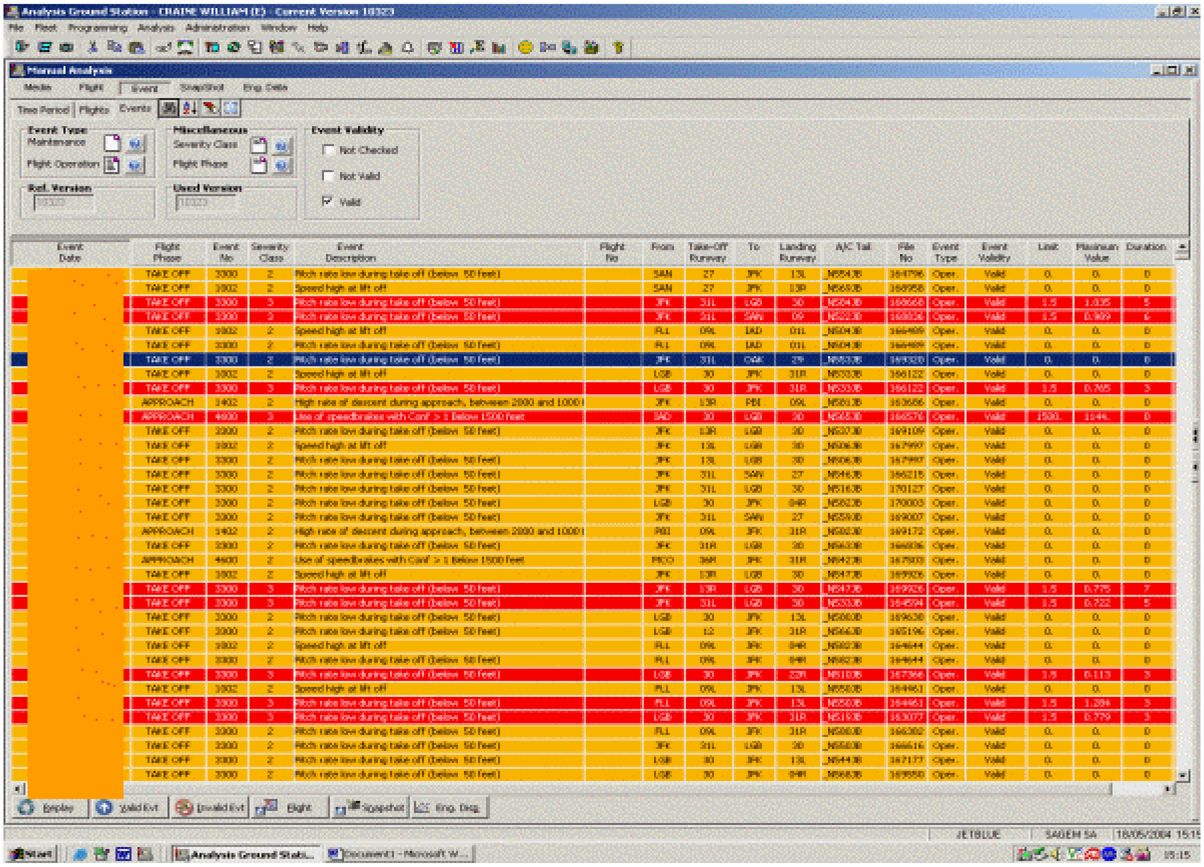
#01/03/2004 Time	FLIGHT_PHASE	CONF	HEIGHT (feet)	SPD_BRK	ALT_STDC (feet)	SPD_MACH_A_M (knot)	VAPP (knot)	CASC (knot)	IVV (ft/min)	VRTG (g)	FLAPC (deg)	LDG_SEL	LDG_SELDW
23:56:50	APPROACH	1	1552		1444	135	135	192.63	-593	0.93	0	N	DOWN
23:56:51	APPROACH	1	1540		1432	135	135	192.63	-635	0.95	0	N	
23:56:52	APPROACH	1	1532	0	1424	135	135	193.00	-624	0.95	0	N	
23:56:53	APPROACH	1	1508		1400	135	135	192.00	-735	0.96	0	N	
23:56:54	APPROACH	1	1488		1380	135	135	191.63	-928	0.96	0	N	DOWN
23:56:55	APPROACH	1	1472		1364	135	135	191.63	-978	1.00	0	N	
23:56:56	APPROACH	1	1456	0	1348	135	135	191.88	-972	1.05	0	N	
23:56:57	APPROACH	1	1440		1332	135	135	191.25	-968	1.07	0	N	
23:56:58	APPROACH	1	1420		1312	135	135	192.63	-1005	1.04	10	N	DOWN
23:56:59	APPROACH	1	1412		1304	135	135	194.00	-951	1.04	10	N	
23:57:00	APPROACH	3	1396	0	1288	135	135	194.38	-874	1.02	15	N	
23:57:01	APPROACH	3	1380		1272	135	135	193.38	-903	0.96	15	N	
23:57:02	APPROACH	3	1364		1256	135	135	190.88	-922	0.92	15	N	DOWN
23:57:03	APPROACH	3	1352		1244	135	135	191.00	-895	0.95	15	N	
23:57:04	APPROACH	3	1336	0	1228	135	135	191.88	-877	0.98	15	N	
23:57:05	APPROACH	3	1316		1208	135	135	191.50	-944	0.98	15	N	
23:57:06	APPROACH	3	1296		1188	135	135	189.88	-1028	0.96	15	N	DOWN
23:57:07	APPROACH	3	1272		1164	135	135	190.25	-1125	0.93	15	N	
23:57:08	APPROACH	3	1248	0	1140	135	135	191.13	-1229	0.98	15	N	
23:57:09	APPROACH	3	1220		1112	135	135	191.50	-1338	0.98	15	N	
23:57:10	APPROACH	3	1196		1088	135	135	192.50	-1411	1.00	15	N	DOWN
23:57:11	APPROACH	3	1168		1060	135	135	192.63	-1460	1.04	15	N	
23:57:12	APPROACH	3	1144	1	1036	135	135	191.25	-1493	1.02	15	N	
23:57:13	APPROACH	3	1120		1012	135	135	190.38	-1476	1.07	15	N	
23:57:14	APPROACH	3	1096		988	135	135	188.38	-1464	1.07	15	N	DOWN

Figure 1: Selected AGS parameter time series from the training database.

Airline operations procedures tend to be written in terms of maneuvers that have target parameter values defined. For example pilots are taught to initiate takeoff at a calculated speed (called Vr) and to rotate the pitch at about 3 degrees per second up to about 18 degrees nose high to achieve a calculated airspeed of about V2 plus 10 knots. The procedures may also define acceptable variation from these targets (what "about" means). Performance outside the target range can be defined as an "exceedance" with generates a FOQA "event".

A primary goal of FOQA programs is to determine the extent to which operations procedures are being followed. Thus the operator defines events (portions of maneuvers) and associated severity classes (exceedance levels). The key metrics are event rates. The FOQA software typically automates the identification and statistical reporting of event exceedances. Figure 1 shows a class 3 exceedance of the event "speedbrakes with configuration > 1".

Secondary data sets exist to record the identified exceedances. Sometimes data corruption makes an exceedance invalid. Sometimes a flight maneuver (i.e. sidestep to land on a different runway) makes an exceedance invalid. It is sometimes necessary to review the parameter time series associated with each exceedance in order to validate the exceedance. Depending on the number of events monitored and the degree of manual validation, event validation can be an extremely time consuming process. The Figure 2 screen capture shows an example recall of validated Class 2 and 3 event exceedances. An AGS window will open the time series data for review (Figure 1).


Analysis Ground Station - CRAINE WILLIAM (E) - Current Version 10323
File Fleet Programming Analysis Administration Window Help
Manual Analysis
Media Flight Event SnapShot Eng. Data
Time Period Flights Events
Event Type
Maintenance
Flight Operation
Miscellaneous
Severity Class
Flight Phase
Replay
Valid Evt
Invalid Evt
Flight
Snapshot
Eng. Diag.
Start
Analysis Ground Stati...
Document1 - Microsoft W...
15:15

Event Validity

☐ Not Checked

☐ Not Valid

☑ Valid

Ref. Version: 10323

Used Version: 10323

Event Date	Flight Phase	Event No	Severity Class	Event Description	Flight No	From	Take-Off Runway	To	Landing Runway	A/C Tail	File No	Event Type	Event Validity	Limit	Maximum Value	Duration
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		SAN	27	JFK	13L	_N554JB	164796	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3302	2	Speed high at lift off		SAN	27	JFK	13R	_N565JB	168958	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		JFK	31L	LGB	30	_N584JB	168668	Oper.	Valid	1.5	1.035	5
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		JFK	31L	SAN	09	_N522JB	168026	Oper.	Valid	1.5	0.909	6
[redacted]	TAKE OFF	3302	2	Speed high at lift off		FLL	09L	IAD	01L	_N504JB	166469	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		FLL	09L	IAD	01L	_N504JB	166469	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	31L	OAK	29	_N553JB	169320	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3302	2	Speed high at lift off		LGB	30	JFK	31R	_N533JB	166122	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		LGB	30	JFK	31R	_N533JB	166122	Oper.	Valid	1.5	0.765	3
[redacted]	APPROACH	1402	2	High rate of descent during approach, between 2000 and 1000 f		JFK	13R	PBI	09L	_N561JB	163686	Oper.	Valid	0.	0.	0
[redacted]	APPROACH	4600	3	Use of speedbrakes with Conf > 1 Below 1500 feet		IAD	30	LGB	30	_N565JB	166576	Oper.	Valid	1500.	1544.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	13R	LGB	30	_N537JB	169109	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3302	2	Speed high at lift off		JFK	13L	LGB	30	_N506JB	167997	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	13L	LGB	30	_N506JB	167997	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	31L	SAN	27	_N546JB	166215	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	31L	LGB	30	_N516JB	170127	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		LGB	30	JFK	04R	_N562JB	170003	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	31L	SAN	27	_N559JB	169007	Oper.	Valid	0.	0.	0
[redacted]	APPROACH	1402	2	High rate of descent during approach, between 2000 and 1000 f		PBI	09L	JFK	31R	_N582JB	169172	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	31R	LGB	30	_N563JB	166836	Oper.	Valid	0.	0.	0
[redacted]	APPROACH	4600	2	Use of speedbrakes with Conf > 1 Below 1500 feet		MCO	36R	JFK	31R	_N542JB	167503	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3302	2	Speed high at lift off		JFK	13R	LGB	30	_N547JB	169926	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		JFK	13R	LGB	30	_N547JB	169926	Oper.	Valid	1.5	0.775	7
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		JFK	31L	LGB	30	_N533JB	164594	Oper.	Valid	1.5	0.722	5
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		LGB	30	JFK	13L	_N580JB	169636	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		LGB	12	JFK	31R	_N566JB	165196	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3302	2	Speed high at lift off		FLL	09L	JFK	04R	_N582JB	164644	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		FLL	09L	JFK	04R	_N582JB	164644	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		LGB	30	JFK	22R	_N510JB	167366	Oper.	Valid	1.5	0.113	3
[redacted]	TAKE OFF	3302	2	Speed high at lift off		FLL	09L	JFK	13L	_N550JB	164461	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		FLL	09L	JFK	13L	_N550JB	164461	Oper.	Valid	1.5	1.284	3
[redacted]	TAKE OFF	3300	3	Pitch rate low during take off (below 50 feet)		LGB	30	JFK	31R	_N519JB	163077	Oper.	Valid	1.5	0.779	3
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		FLL	09L	JFK	31R	_N500JB	166302	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		JFK	31L	LGB	30	_N550JB	166616	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		LGB	30	JFK	13L	_N544JB	167177	Oper.	Valid	0.	0.	0
[redacted]	TAKE OFF	3300	2	Pitch rate low during take off (below 50 feet)		LGB	30	JFK	04R	_N568JB	169550	Oper.	Valid	0.	0.	0

JETBLUE | SAGEM SA | 18/05/2004 15:15

Figure 2: AGS event search result

1.1.2 Event Exceedance Analysis

Exceedances are thus defined and recorded to gain an understanding of when aircraft performance fails to meet target value ranges during specific maneuvers. The most natural questions that arise are baseline risks (how often exceedances occur) and trend (if the risk increasing or decreasing). Another issue is how consistent the risk is. That is, if the month-to-month variation in exceedance rates is consistent with a stable operations environment or if changes in the environment create increased risk of an exceedance.

A typical statistical method for analyzing baseline risk, trend, and variation is the control chart. This is a time series (usually monthly) of exceedance rates as a percentage of flights. The baseline risk is the overall average. This average is used to define upper and lower control limits that represent the predicted statistical variation one would see in a stable operations environment where only random probabilistic variation is present. Trend and excessive variation are easily identified. These charts are used to identify maneuvers warranting training attention and to monitor training program effectiveness.

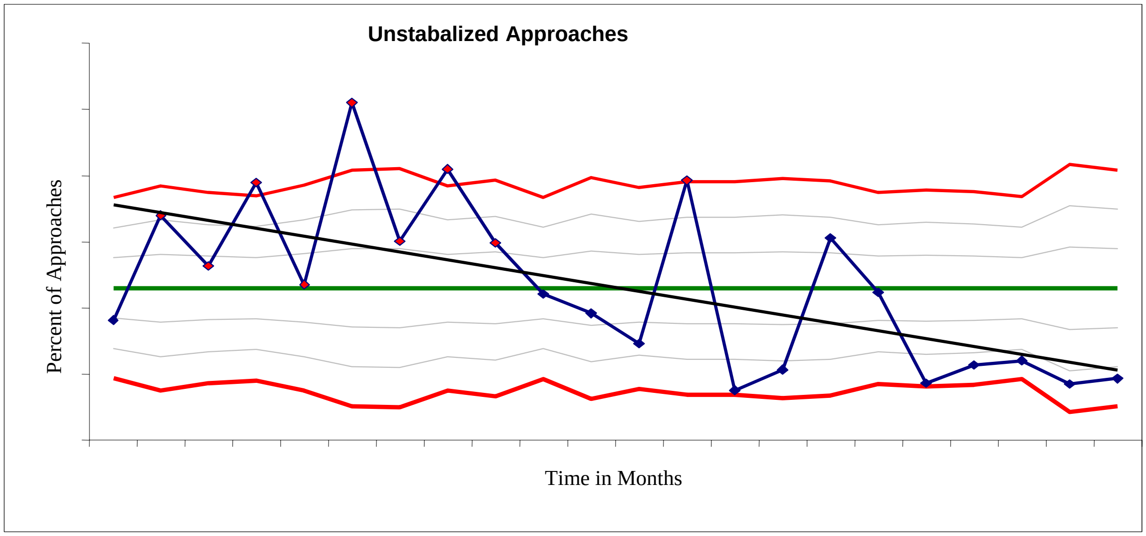


Figure 3: Increasing trend followed by an effective training program

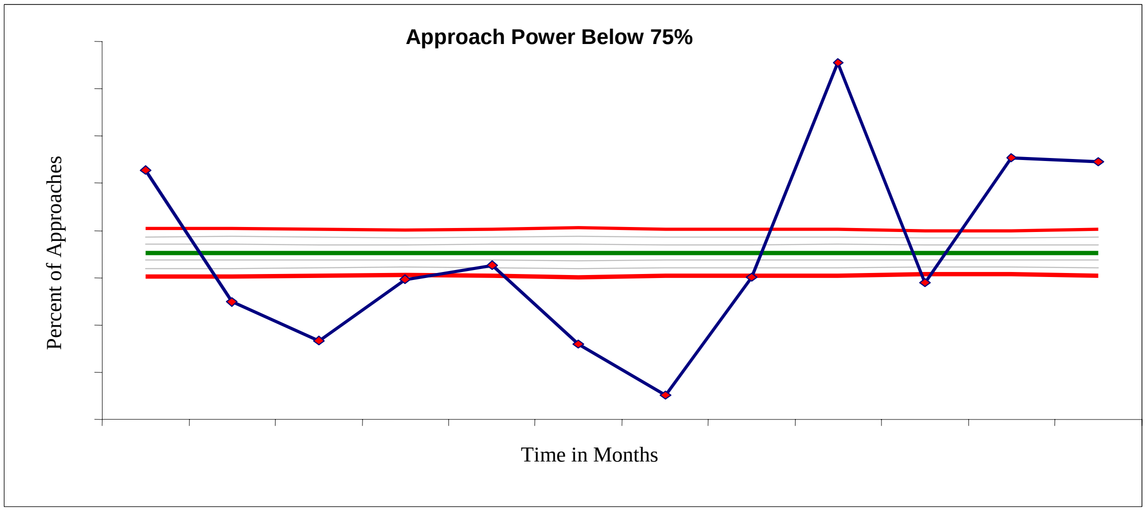


Figure 4: Excessive variation indicates an unstable operations environment

Exceedance rates are often correlated with one or more categorical variables, most often location. Analysis of exceptionally high or low exceedance rates may give insight into root causes of the exceedance. Figure 5 shows rates of “high speed at low altitude”. Some of the locations have special procedures where the high speeds are acceptable (the exceedance does not indicate a violation of procedures) while other locations’ high exceedance rates do indicate procedure deviations.

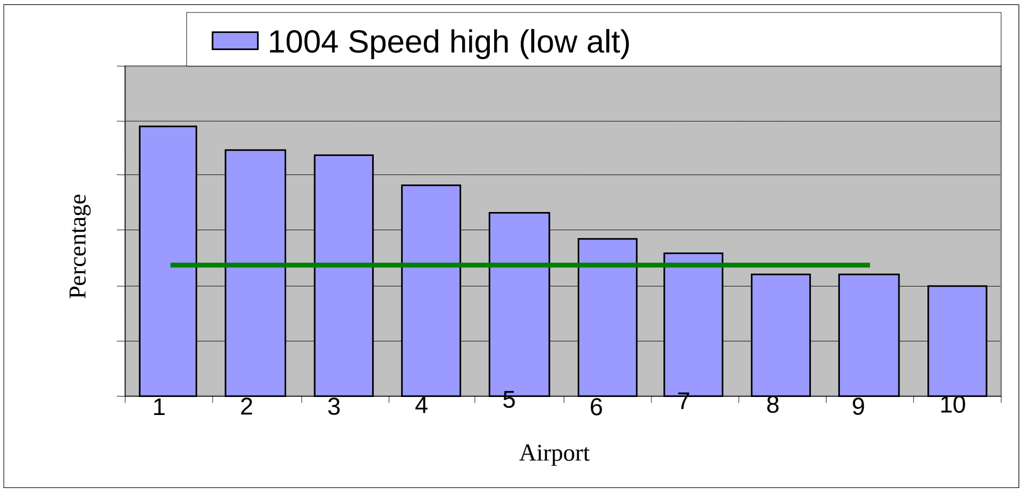


Figure 5: Event rates by arrival airport

A variation of the bar chart would include statistical variation bars to indicate confidence intervals (sampling error). Figure 6 came from an ASAP program database rather than a FOQA database. The common analysis method requires computing an event rate (percentage) for the values of a parameter. The goal of this study was to see if reported rudder malfunctions occurred at rates similar to other aircraft systems.

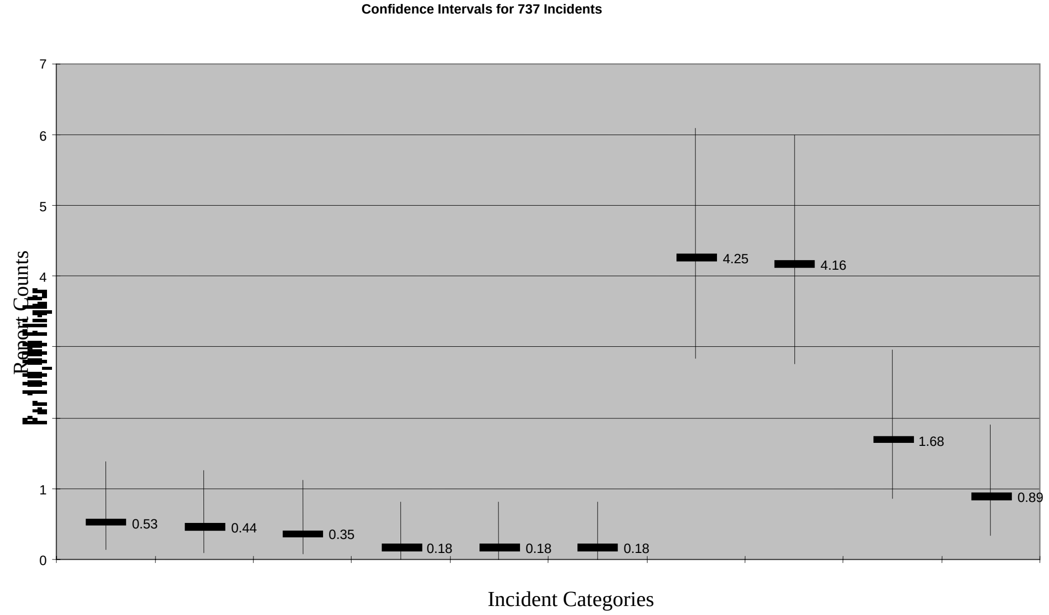


Figure 6: Comparison of ASAP reporting rates with confidence intervals

1.1.3 Flight Performance Distribution Analysis

As an alternative to event based analysis, JetBlue has increased distributional analysis of maneuver performance. Identifying event exceedances often requires the creation and recording of "snapshot parameters" that record aircraft data at a specific point in flight. The snapshot parameters can also be created independent of events. Two examples are "during initial takeoff" and "on approach at 1000 ft". They can capture either instantaneous values or max/min values during an interval. This set of data also stored in a database and is available for analysis. JetBlue takeoff procedures call for rotation rates of 3 degrees per second to achieve an initial climb speed of V2 plus 10 knots. Figures 7 and 8 show the example distributions that can be compared to performance targets.

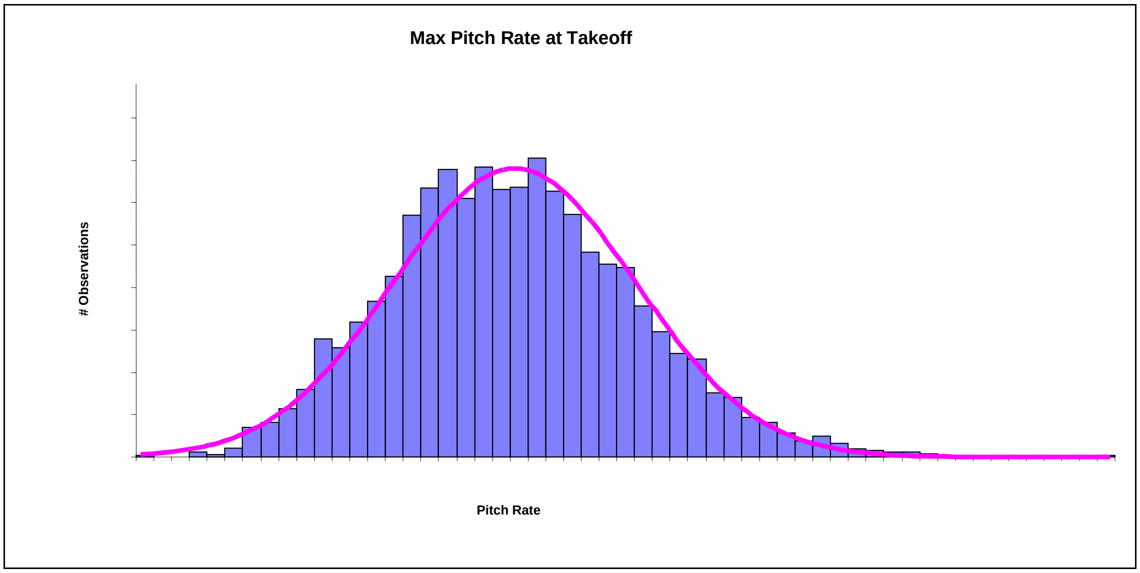


Figure 7: Max pitch rate on takeoff

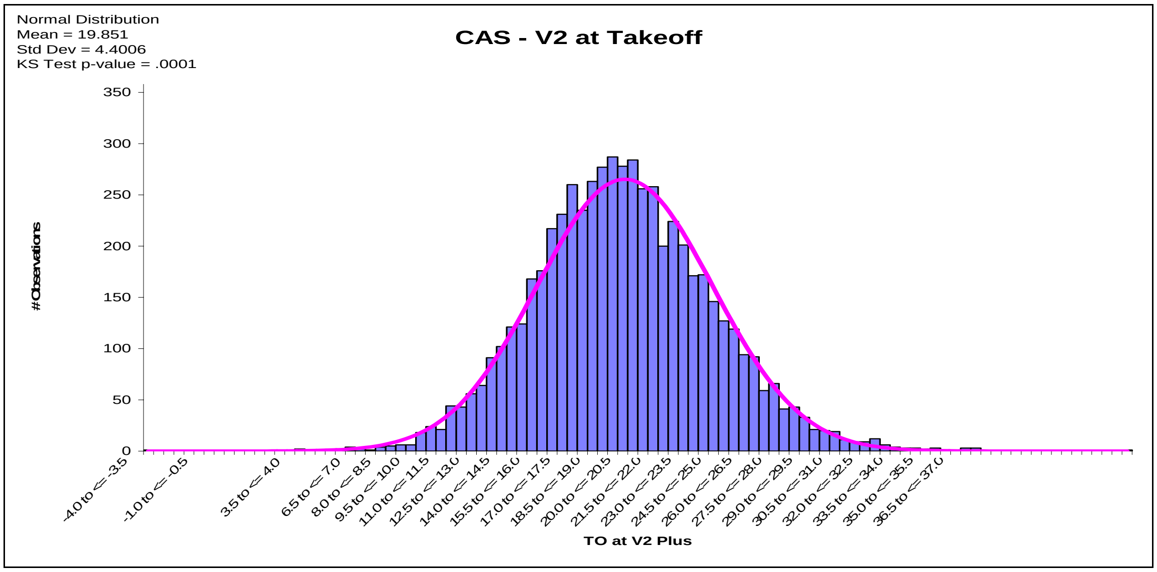


Figure 8: Airspeed above V2 during takeoff

The snapshot parameters also allow multi-variable statistical methods like regression analysis or analysis of variation (ANOVA). Figure 9 shows a scatter plot and linear regression that demonstrates a degree of correlation between two performance parameters. Some correlations are well understood. For example on takeoffs reduced rotation rates are correlated with increased takeoff speeds. On approach, low power is highly correlated with both high descent rates and high airspeeds on approach.

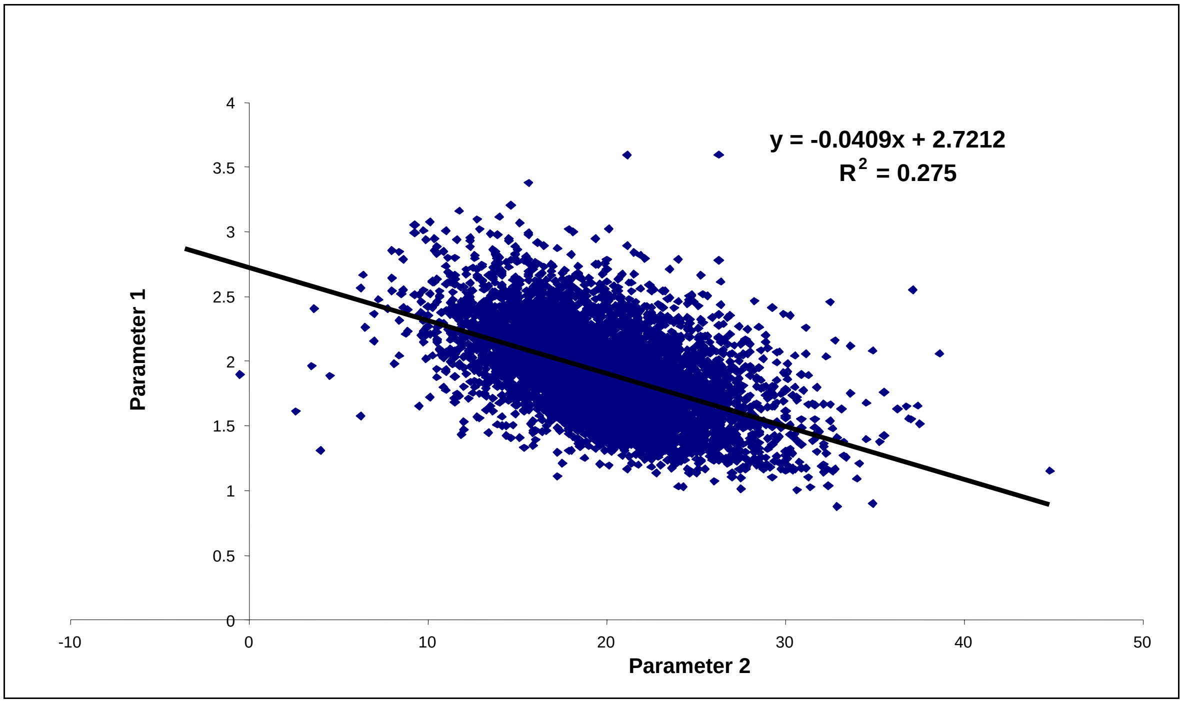


Figure 9: Linear Regression of Parameter 1 VS Parameter 2

1.2 Objectives of Project

The objective of this effort is to evaluate data mining techniques and their potential to enhance the current analysis of airline digital fight data. FAA Office of System Safety (ASY) hopes to facilitate the widespread application of data mining techniques to airline safety management, by (1) helping data mining companies and/or vendors of FDM systems understand the needs of airlines and develop appropriate data mining capabilities and (2) showing airlines the value of these techniques and inspiring them to integrate such techniques in their operations.

To support Objective (1), this project has attempted to document the steps required to merge a data mining software tool with an airline's FOQA system; the airline's perceptions about the mining tool and how it could be improved, such as what needs to be speeded up or made more understandable or automated; what features should be added to a data mining tool to improve its usefulness; and similar findings.

To support Objective (2), this project has attempted to evaluate value of the data mining tool (at least in a qualitative sense) in addressing things like identification of previously unknown safety issues and possible time savings over other analysis approaches currently in use, with some comparison to FOQA analysis capabilities without the data mining tool.

This report presents analysis summaries that produce interesting results primarily in support of Objective (2). These analyses result, in part, from extensive discussions with pilots and managers of JetBlue and SAGEM.

1.3 Insightful Miner™: Scalable Data Analysis Workbench - Overview

Insightful Miner™ is a highly scalable **data analysis workbench** which supports the entire lifecycle of data access, data preparation, exploration, modeling and reporting. Its key capabilities include:

- Insightful Miner's **visual programming interface** reduces the complexity of data analysis. By linking together analytic components (called "nodes") using a simple drag-and-drop interface, data analysts create sophisticated data analysis applications while the self-documenting visual workmap documents every step taken in an easy-to-understand visual depiction of the process.
- Insightful Miner includes **out-of-memory algorithms** to reliably scale to the largest data sets. The user interface is a visual metaphor for an underlying **pipeline architecture** that uniquely enables scalability by operating on large data sets in an incremental fashion, rather than reading the entire data set into memory at once.
- All Insightful Miner components (such as those for data access, exploration, manipulation, cleaning, statistics & machine learning, assessment and deployment) exploit the pipeline architecture for **scalability** in data size and processing speed.
- All Insightful Miner components can operate on in-memory and out-of-memory datasets.
- Insightful Miner works with the Insightful's S-PLUS® data analysis environment to provide a rich programming environment.
- Insightful Miner is designed for **deployment** to production environments, with algorithms designed to be able to process the largest data sets, and tools needed to integrate them into existing infrastructure.

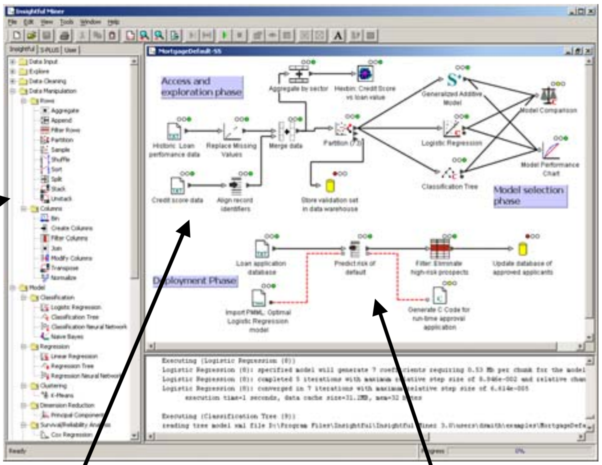
1.3.1 Visual Programming Interface

The Insightful Miner user interface is written in Java and features panes for the component library, workspace and engine reporting. Components are dragged from the library, placed in the workspace and wired together to form visual workmaps.

The visual programming metaphor is self-documenting and makes data mining more accessible and reproducible. Component properties are accessed and populated by double clicking on the component. Workmaps can be executed in their entirety, in subsets or as individual components. Once a model is built, a mouse click produces a prediction component for use in scoring a separate deployment stream. In addition, all the preprocessing steps necessary to prepare live data for scoring may be copied into deployment.

Scalable components support each step of the data mining process

Access
- Exploration
- Cleaning
- Manipulation
- Model building
  - Classification
  - Regression
  - Clustering
- Assessment
- Output


Insightful Miner
MortgageDefault-55
Access and exploration phase
Aggregate by sector
Hexbin: Credit Score vs loan value
Generalized Additive Model
Model Comparison
Historic Loan performance data
Replace Missing Values
Merge data
Partition (7:3)
Logistic Regression
Model Performance Chart
Credit score data
Align record identifiers
Store validation set in data warehouse
Classification Tree
Model selection phase
Deployment Phase
Loan application database
Predict risk of default
Filter: Eliminate high-risk prospects
Update database of approved applicants
Import PMML: Optimal Logistic Regression model
Generate C Code for run-time approval application
Ready
Progress
0%

```
Executing (Logistic Regression (8))
Logistic Regression (8): specified model will generate 7 coefficients requiring 0.53 Mb per chunk for the model
Logistic Regression (8): completed 5 iterations with maximum relative step size of 8.946e-002 and relative chan
Logistic Regression (8): converged in 7 iterations with maximum relative step size of 6.614e-005
      execution time=1 seconds, data cache size=31.2MB, mem=32 bytes

Executing (Classification Tree (9))
reading tree model xml file D:\Program Files\Insightful\Insightful Miner 3.0\users\dsmith\examples\MortgageDefa
```

Build Powerful Models by Programming ***Visually***

Wire components together to form self-documenting workflows, then quickly build and evaluate multiple models to choose the best.

Deploy Models into Enterprise Systems

Publish Web-ready graphics and reports to decision makers, or integrate your model and pre-processing steps into your production applications.

The visual workmap reflects the state of the underlying pipeline architecture and visually reflects the state of each node in your worksheet: unconfigured, ready to run, or completed. This makes it easy to spot exactly where you need to make changes if the entire workflow does not run to completion. In addition, since the results of each node are cached to disk (by default), the computations for preceding nodes are stored, and you only need to re-run those nodes downstream from the problem node. This makes it easy to develop workflows in an iterative and exploratory fashion, without needing to re-run the entire workflow whenever a change is made.

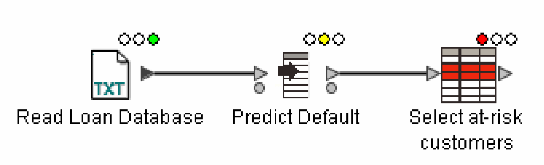


Figure 10: The visual pipeline interface indicates the state of all components: completed (green), ready to run (yellow), and requiring configuration (red). The results from completed nodes are automatically cached for dynamic review and to eliminate unnecessary re-computations.

The workflow is stored as a single "worksheet file" which may be distributed to other users. Also, the worksheet file is implemented as a single XML file with cross-platform and cross-edition compatibility, which allows easy reconfiguration and extensibility and ensures smoother integration with evolving IT infrastructures.

1.3.2 Pipeline Architecture

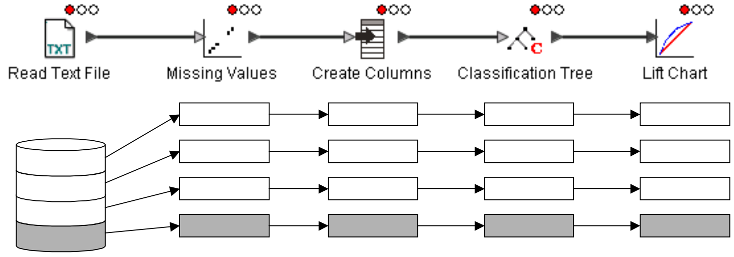
The key architectural feature of the Insightful Miner computational engine is the **pipeline**. This architecture solves the big-data problems related to reading, transforming, modeling and writing large data sets. The pipeline also enables several other important benefits:

1 flexible caching for balancing interactivity with resource use,
2 straight-forward integration of new data mining components and
3 support for parallel processing.

Insightful Miner's architecture partitions large datasets into rectangular structures of rows and columns called blocks. The number of columns in a block is limited to 2 billion, and the number of rows is adjusted so the block fits into RAM. Block size is set at a reasonable default, but can be adjusted by the user. Components process the blocks as they stream through memory-resident buffers. In this way, a practically unlimited number of rows can be processed allowing Insightful Miner to interact with datasets that are much larger than available RAM.

A series of operations is defined by linking components together to form a "pipeline". The pipeline supports numerous data types including numeric, date, categorical & string.

Pipeline Architecture



The pipeline breaks large datasets into blocks and processes them through a series of components

The pipeline system creates networks of components & buffers that process the blocks in sequence. While there are close similarities between the assembled components in a workmap and the low-level pipeline networks, they are not always equivalent. To perform a particular operation Insightful Miner may construct and execute multiple pipelines. For instance, the data cache option (discussed below) will construct pipelines that are not visually represented to the user.

During processing, the pipeline collects and stores metadata with each component that speeds downstream computations. It can optionally create intermediate cache files stored in a compressed binary format. Caching is critical to providing a highly interactive user experience with large datasets. By retaining cached representations at intermediate stages in the pipeline, users can revisit previous exploratory views and perform further processing without re-running previous steps.

1.3.3 Insightful Miner Components

The pipeline provides a block-wise, row-oriented framework for interacting with large data. All Insightful Miner components, whether built-in or user-defined, take advantage of the pipeline architecture and block framework to deliver scalability throughout the data analysis process.

While there are numerous data analysis and data mining methodologies, they all agree on the core processes involved: data access, preprocessing (exploration, cleaning and manipulation), model building, model evaluation/assessment and deployment. Referring to this as the data mining lifecycle, Insightful Miner provides a full range of components to support the data analysis lifecycle.

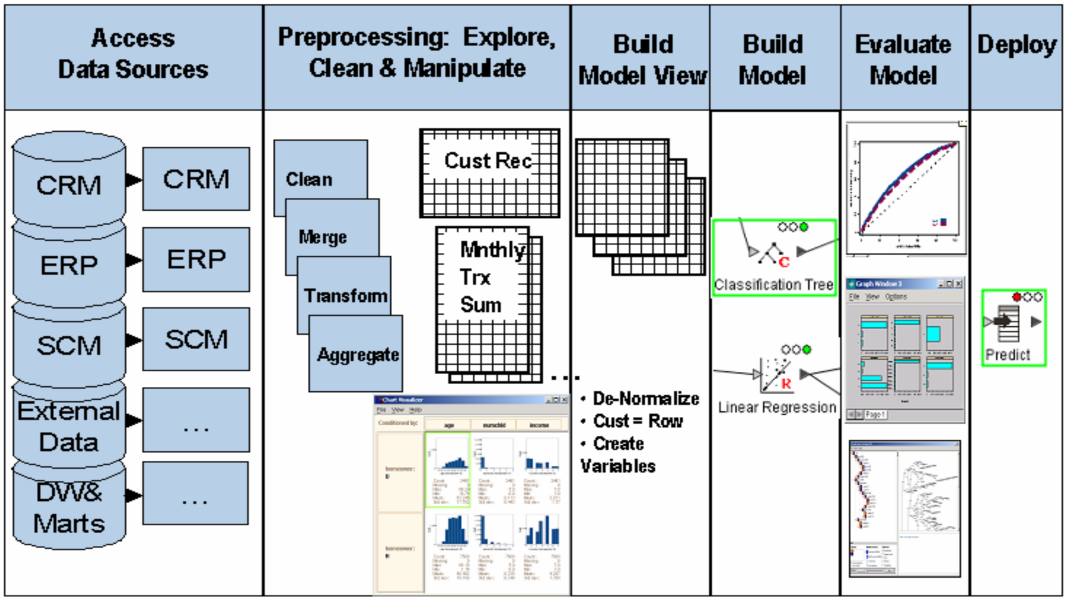


Figure 11: Insightful Miner supports each step of the data analysis lifecycle

1.3.3.1 Data Reading & Writing Components

Scalable processing for read/write components is straightforward and well understood. All components can recognize and coerce data types (e.g. treat numeric variables as categorical variables). The GUI also includes a preview pane to eliminate time-consuming misfires.

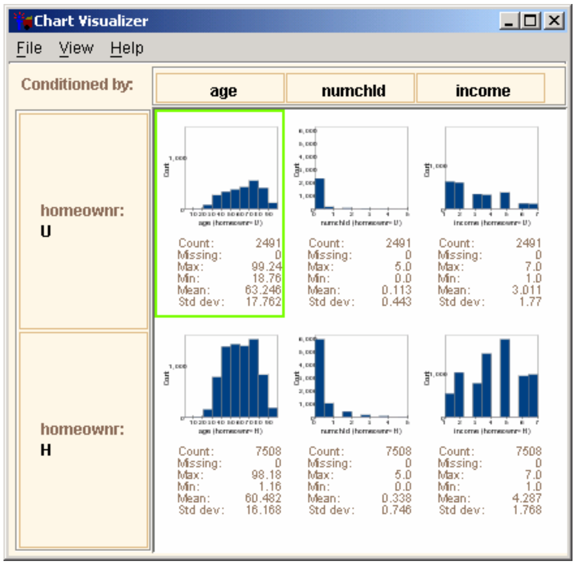
The basic palette includes the ability to read and write:

- Delimited and fixed-format text, with support for data dictionaries
- SAS, SPSS, Excel, S-PLUS and 25 additional file formats
- Relational databases including Oracle, MS SQL Server, DB2 & Sybase (accessed via a native interface for optimal performance) and all ODBC-compliant databases on Windows

Custom data access components may also be created.

1.3.3.2 Data Exploration Components

Data Exploration helps users understand their data and uncover important relationships, such as which variables have predictive power and which do not. Insightful Miner components use scalable techniques that leverage metadata, data caching and counts resulting in good performance even on gigabyte-sized datasets. For example, the *Table View* component provides very fast scanning through thousands of rows by using the data cache. The *Descriptive Statistics*, *Crosstabs* and *Charting* components provide text and graphic views from metadata and count information. Basic options include barcharts, piecharts & dotcharts for categorical data, and histograms & boxplots for continuous data. Advanced capability includes Trellis conditioning on categorical variables.



```
homeownr: U
Count:       2491    Count:      2491    Count:      2491
Missing:        0    Missing:       0    Missing:       0
Max:        99.24    Max:         5.0    Max:         7.0
Min:        18.76    Min:         0.0    Min:         1.0
Mean:      63.246    Mean:      0.113    Mean:      3.011
Std dev:   17.762    Std dev:   0.443    Std dev:    1.77

homeownr: H
Count:       7508    Count:      7508    Count:      7508
Missing:        0    Missing:       0    Missing:       0
Max:        98.18    Max:         5.0    Max:         7.0
Min:         1.16    Min:         0.0    Min:         1.0
Mean:      60.482    Mean:      0.338    Mean:      4.287
Std dev:   16.168    Std dev:   0.746    Std dev:   1.768
```

Figure 12: Trellis Graphs are a powerful visualization technique for unlocking hidden relationships in high dimension data.

When Insightful's S-PLUS data analysis environment is installed on the system, additional charting components become available. This includes 1-D charts such as pie charts, bar charts and density plots; 2-D charts including Hexagonal Binning "scatterplots" for visual display of very large data sets; 3-D charts such as perspective plots and contour maps; multivariate displays including the scatterplot matrix, and time-series charts including high-low and stacked bar charts. Because all of these charting components are based on S-PLUS's Trellis graphics design, they can be "conditioned" on a categorical variable such as age or region, to visually detect differences between sub-groups.

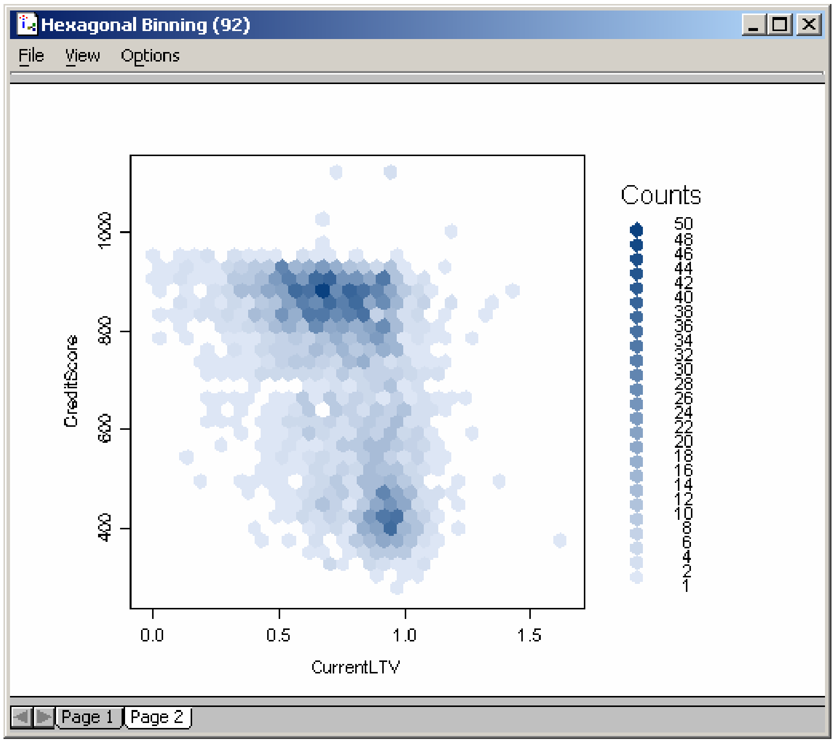
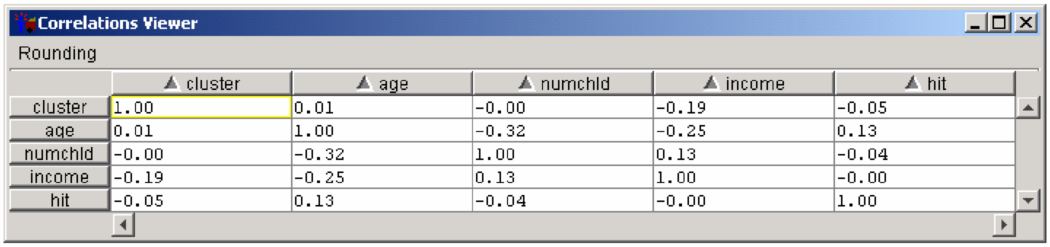


Figure 13: Insightful Miner's Hexagonal Binning chart provides a unique way of visualizing relationships between continuous variables, even for very large data sets. By representing the intensity of points falling within a small hexagonal region by depth of color, unusual clusters in the data can be detected visually.

1.3.3.3 Data Cleaning Components

Data cleaning components are used to repair data quality problems and identify outliers. The *missing values* component includes variance-preserving techniques and imputation methods, to repair missing values given available data. The *outlier detection* component uses Robust Statistics to spot outliers in high-dimension multivariate data. Screening for outliers can locate important, high value segments in data — not just miscoded variables — making this an important capability for modelers.



Correlations Viewer

Rounding

	cluster	age	numchld	income	hit
cluster	1.00	0.01	-0.00	-0.19	-0.05
age	0.01	1.00	-0.32	-0.25	0.13
numchld	-0.00	-0.32	1.00	0.13	-0.04
income	-0.19	-0.25	0.13	1.00	-0.00
hit	-0.05	0.13	-0.04	-0.00	1.00

Correlations and Covariance components give valuable criteria for discriminating between important and unnecessary variables.

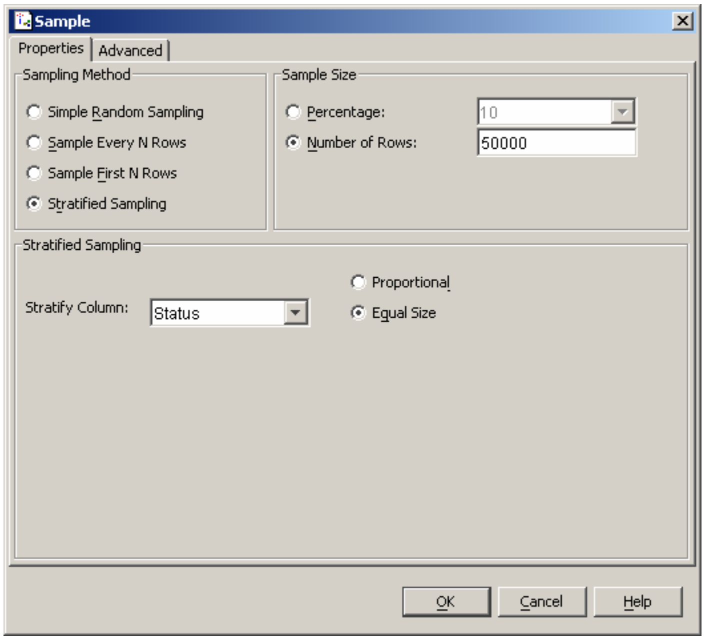
1.3.3.4 Data Manipulation Components

Data manipulation is critical for transforming data from its original formats into forms compatible with building models. For instance, flight safety models are typically built from a flat structure where each flight is represented by a single row. For many of the analyses completed in this study the numerous observations for each flight had to be summarized into a single row. All basic SQL manipulations are included, providing the ability to select, aggregate and de-normalize data to create good predictive views. Insightful Miner's data manipulation components support column-based and row-based operations including:

Row: Aggregate, Append, Filter, Shuffle, Sort, Split, and Stack/Unstack

Column: Create, Filter, Join, Modify, and Transpose.

Also important is the *Sampling* component that incorporates stratified techniques and includes the ability to up-sample or down-sample. This is particularly useful when training a classification algorithm against a categorical column with an underrepresented factor level.



Sample

Properties | Advanced

Sampling Method
○ Simple Random Sampling
○ Sample Every N Rows
○ Sample First N Rows
● Stratified Sampling

Sample Size
○ Percentage: 10
● Number of Rows: 50000

Stratified Sampling
Stratify Column: Status
○ Proportional
● Equal Size

Figure 14: The Sample node provides simple random sampling as well as stratified sampling techniques

1.3.3.5 Modeling Components

Components that train models are the most specialized components of a data analysis system. It's here where many scalability problems are encountered. Algorithms that assume all of the data is available in memory are inherently not scalable. Some common machine learning algorithms and products suffer this limitation, making them memory-bound. Insightful Miner provides an innovative solution (Block Model Averaging) which scales many inherently memory-bound algorithms.

Insightful Miner features scalable modeling components for classification & regression (supervised learning, or prediction), clustering (unsupervised learning) and dimension reduction:

Model	**Model Type**
Linear Regression	Regression
Logistic Regression	Classification
Trees	Classification & Regression
Neural Networks	Classification & Regression
Cox Regression	Regression
Naïve Bayes	Classification
Principal Components	Dimension Reduction
K-Means	Clustering

Many tools force users to sample large data sets in order to make them fit into RAM prior to modeling. This can be problematic because 1) Sampling requires expertise to produce good results, and 2) it forces the user to balance the ideal sample size against their system's memory capacity in both modeling and production environments. Under this scenario, users must ask themselves: How large a sample should be used to be statistically valid? Which sampling technique is best? What power is gained at a particular sampling level? Moreover, how does all this relate to my need to operate within given memory limitations?

Unfortunately, these computations are complex, time consuming and error-prone. The result is that modeling remains an expert-only activity and requires excessive effort to avoid incomplete runs and under-performing predictions. While these efforts are easier to justify when the model is built for a one-time prediction, in production environments, models must be regularly maintained and modified. Many tools hide sampling inside the model. While this reduces the burden on the user, it takes control away and can misrepresent results.

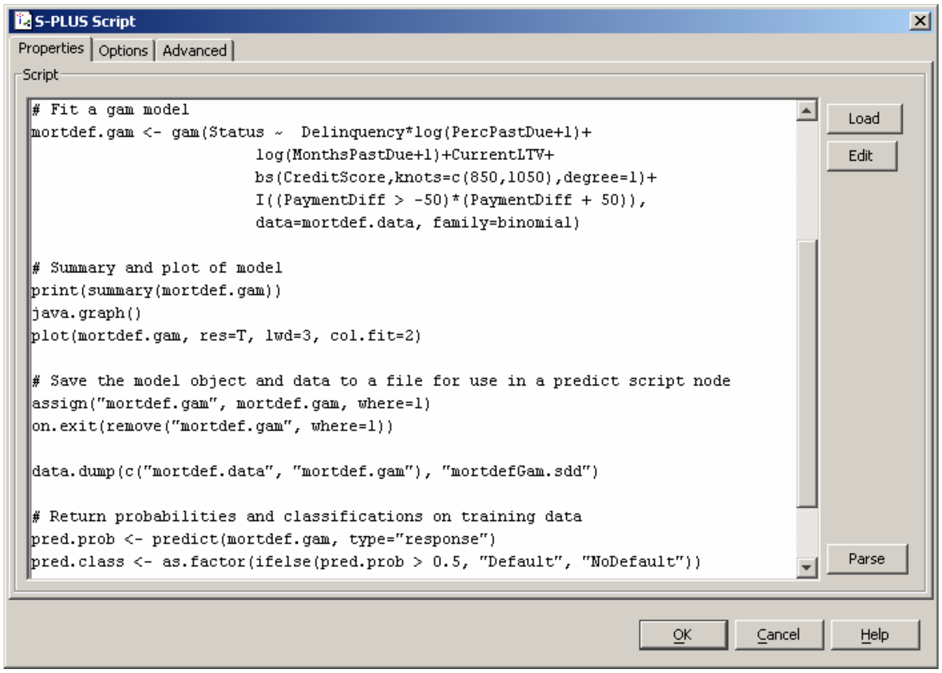
The point is not that sampling is universally bad. Sampling can reduce processing time and is justified in many cases. But sampling solely to overcome memory limitations is not good practice. As business analysis becomes increasingly sophisticated, the fractional differences in model accuracy that was once an acceptable margin of error have instead become the margin of profit. This is why advanced analytics became important in the first place, to deliver a more granular understanding of our customers, products and business processes. Insightful Miner's scalable components support sampling but don't require it or hide a rigid sampling scheme inside the model.

1.3.3.6 Adding S-PLUS to Insightful Miner

By adding S-PLUS to an existing Insightful Miner installation, additional functionality becomes available from within the Insightful Miner interface. This includes not only the additional charting components, but also several components that allow Insightful Miner to interface with the S programming language available within S-PLUS.

Three of these components, "S-PLUS Filter Rows", "S-PLUS Split" and "S-PLUS Create Columns" are analogues of similar components available in the default Insightful library. By using these components, all 4000+ functions of the S programming language are available for data selection and data creation tasks, greatly enriching the data processing capabilities of Insightful Miner. The S-PLUS Create Columns node, in particular, allows Insightful Miner users familiar with the S programming language to perform complex SQL-like transformations of existing columns, and create new random columns (for simulations) based on a rich set of probability distributions built into the S language.

Also, the S-PLUS Script node allows complete S scripts to be processed on data within the Insightful Miner pipeline. With this node, models charts and reports created by an S-PLUS programmer can be made available within the visual workflow interface by adding new custom components.



```
# Fit a gam model
mortdef.gam <- gam(Status ~  Delinquency*log(PercPastDue+1)+
                        log(MonthsPastDue+1)+CurrentLTV+
                        bs(CreditScore,knots=c(850,1050),degree=1)+
                        I((PaymentDiff > -50)*(PaymentDiff + 50)),
                        data=mortdef.data, family=binomial)

# Summary and plot of model
print(summary(mortdef.gam))
java.graph()
plot(mortdef.gam, res=T, lwd=3, col.fit=2)

# Save the model object and data to a file for use in a predict script node
assign("mortdef.gam", mortdef.gam, where=1)
on.exit(remove("mortdef.gam", where=1))

data.dump(c("mortdef.data", "mortdef.gam"), "mortdefGam.sdd")

# Return probabilities and classifications on training data
pred.prob <- predict(mortdef.gam, type="response")
pred.class <- as.factor(ifelse(pred.prob > 0.5, "Default", "NoDefault"))
```

Figure 15: The S-PLUS Script node brings the power and flexibility of the S language to Insightful Miner

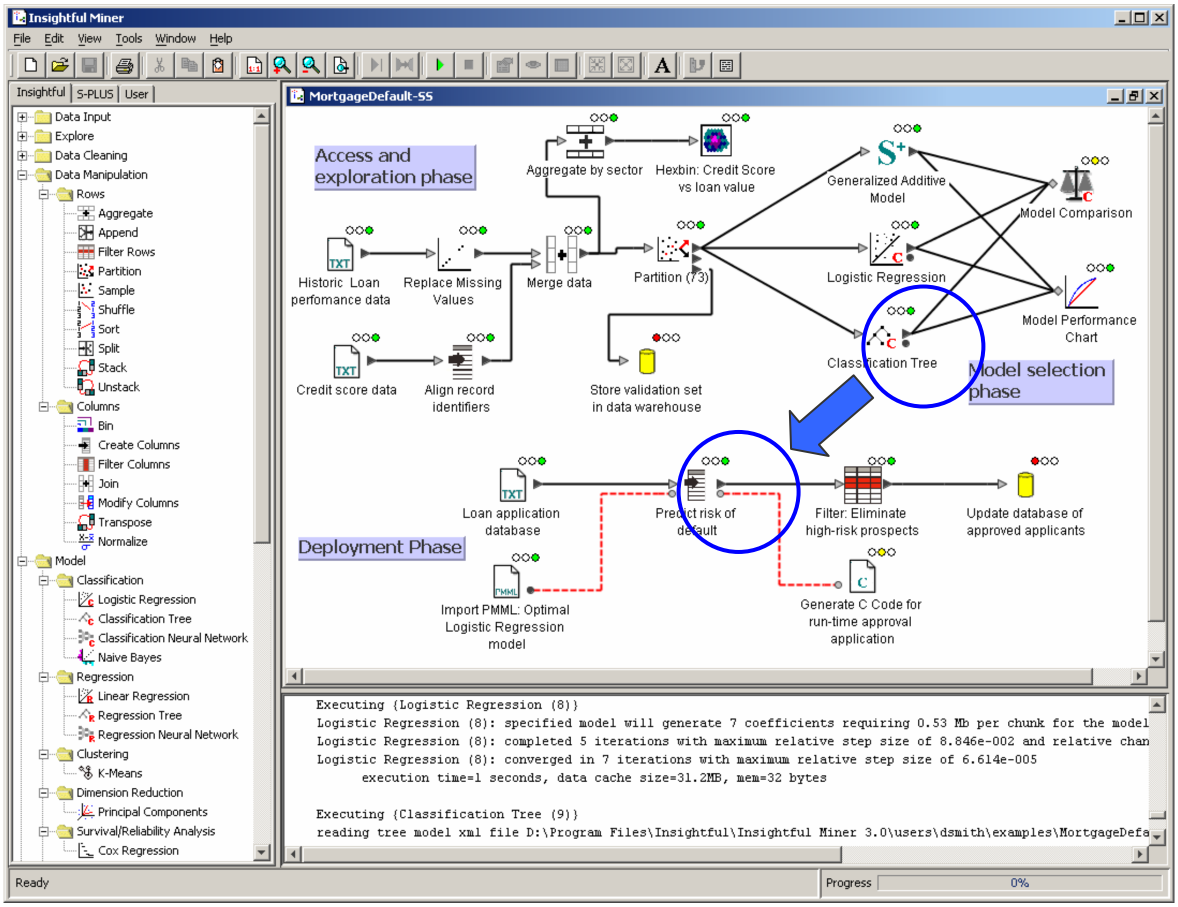
1.3.4 Deployment

Insightful Miner supports deployment into production in several ways:

1. Within-worksheet model scoring using the predict node
2. General process deployment in batch mode
3. Run-time model scoring with C-code generation
4. PMML deployment

1.3.4.1 Deployment with the Predict node

Once a model has been trained within Insightful Miner, scoring the model on another database is done by creating a new pipeline. The source model is linked to a **Predict** node in Insightful Miner, and that Predict node is linked to the source scoring data[1]. The "model ports" feature of Insightful Miner 3 provides the flexibility to store a trained model permanently, or to have the model automatically update when new training data is processed.



```
Executing {Logistic Regression (8)}
Logistic Regression (8): specified model will generate 7 coefficients requiring 0.53 Mb per chunk for the model
Logistic Regression (8): completed 5 iterations with maximum relative step size of 8.846e-002 and relative chan
Logistic Regression (8): converged in 7 iterations with maximum relative step size of 6.614e-005
      execution time=1 seconds, data cache size=31.2MB, mem=32 bytes

Executing {Classification Tree (9)}
reading tree model xml file D:\Program Files\Insightful\Insightful Miner 3.0\users\dsmith\examples\MortgageDefa
```

Figure 16: A Tree model is used for scoring by creating a Predict node and linking it into a separate scoring pipeline.

[1] Automatic Predict node generation is supported for the native Insightful Miner models listed in Section 1.3.3.5. For custom S-PLUS models, an S-PLUS script node calling the `predict` function is used.

1.3.4.2 Batch mode deployment

Any Insightful Miner worksheet can be saved as a single worksheet file in XML format. This worksheet can then be processed non-interactively using Insightful Miner's batch mode[2]. This is typically used for regular production tasks, where the source data files or database is updated by an external process, and then Insightful Miner's batch mode is initiated to process the data and output the results to files or tables in specified locations. Large-scale scoring applications are particularly suited to this type of deployment.

1.3.4.3 Run-time model scoring

Once a model has been trained in Insightful Miner, it may be of interest to score that model on a case-by-case basis in an interactive or high-frequency environment. For example, a desktop application may be needed for on-the-fly approval or denial of loans in a retail bank, or an online trading system may need to update buy or sell recommendations second by second. For applications like these, embedding the model as a C function within a larger application is a good solution.

Insightful Miner 3 allows you to export any trained model[3] as a function call in C. Insightful Miner generates the source code, which may then be embedded in a C-based application. The C function provides an interface for supplying the new data for scoring, and the value of the function provides the prediction based on the trained model. This is particularly useful for independent deployed applications where low latency of response and a minimal footprint for the scoring application is a key requirement. There are no restrictions on how the C code generated from Insightful Miner is used, and no additional licenses are required to use the C code in deployed application.

1.3.4.4 PMML deployment

Insightful Miner allows you to export trained models as a PMML (Predictive Modeling Markup Language) file. PMML is a new industry standard which records all the parameters of a trained model (for example, coefficients in a regression model, or breakpoints in a tree model) as an open XML document. The PMML standard is developed and maintained by the vendor-led standards organization The Data Mining Group.[4]

By saving a trained model in the PMML format, you can keep a permanent record of models in use (or which have been used in the past) in an open format. Standard XML tools can be used to view PMML files or to extract out relevant information (such as model parameters) for further processing. Insightful Miner also allows for the import of PMML files to describe a model for scoring.

[2] Insightful Miner Server Edition with Production Pack is required for batch mode operation.

[3] This applies only to native Insightful Miner models listed in Section 1.3.3.5. Custom models created using S-PLUS are not supported for C-code generation. Insightful Miner Server Edition with Production Pack is required for C code generation.

[4] For more information on PMML and the Data Mining Group visit www.dmg.org.

PMML is a relatively new standard, but as the importance of predictive models in industry grows, we expect the importance of PMML – and the number of applications that support the standard – to similarly grow with time. Some databases already allow certain models to be scored within the database from a PMML specification. In addition to the benefits of open access to model results, the potential of interoperability with other applications is one of the key reasons why Insightful Miner supports the PMML standard.

2 Preparation of Data and Tool

2.1 Description of input data

Data consisted of three different fundamental extractions. All data extractions were moved via ftp from SAGEM computing resources to Insightful computing resources as ASCII flat files. Large files were compressed before transfer.

1. **Event data:**
   Data included all events for all flights from June 2003 through June 2004. Data included variables of observed File No (equivalent to flight no), Month-Year, Event No, Event Short Name, Flight Phase, Severity Class, Exceedance Duration and Departure and Arrival Airports.
2. **Snapshot data:**
   Data consists of a selection of flight parameters observed at the time of an event, signaled by the exceedance of some flight parameter(s) beyond their SOP threshold(s) for all flights during the period January 2004 through June 2004. Data included File No, Event No and Flight Phase for matching with the event data plus parameters measuring altitude, speed, engine thrust, power settings, flap configuration, roll angle, pitch angle, angle of attack, etc.
3. **Fullflight data:**
   Data consists of a (large) selection of flight parameters observed at 1 Hz frequency for the first three minutes and the last eight minutes of each flight for all flights during the period January 2004 through June 2004. Data included time sequences for all the snapshot parameters plus additional parameters like Glideslope Deviation, Exhaust Gas Pressure and Severity of Ice Detected.

2.2 Data Capture, Cleansing and Transformation

As described in detail in section 1.1.1, for each flight over 300 parameters are recorded as time series. Parameters may be either numeric (e.g., airspeed) or categorical (e.g., autopilot on/off). In-flight *events* are triggered when predefined parameter envelopes containing normal operations were surpassed.

All data were extracted from SAGEM's AGS database. The event data and snapshot data were relatively small and easily extracted. Extracting the fullflight data was more difficult. Because of the volume of data (660 observations on over 125 parameters for thousands of flights) extractions were tedious and slow. Once analysis began on the fullflight data, any data issues or desire to extract additional parameters based on initial results, were not easily resolved. Long extraction times constrained the exploratory nature of this study cutting down on the iterations typical of exploratory studies. In retrospect, selecting more parameters, even redundant parameters, would have given this project more working room. Furthermore, doing a thorough analysis on a small set of flights, more easily extracted, would have uncovered any data issues prior to doing the full extraction.

Insightful received separate monthly ASCII files for events, parameter snapshots at event times and full-flight parameter data for the first three minutes and last eight minutes of each flight. All data were imported using Insightful Miner native ASCII file import nodes.

Data were examined for recording errors by computing summary statistics such as minimum, maximum, mean, standard deviation and missing value counts and plotting histograms for visual inspection. Improperly recorded data were either extracted from the AGS database again and reexamined before proceeding or eliminated from the analysis.

An assortment of new variables were created by a combination of logical operations and mathematical transformations using both I-Miner and S-PLUS. Two important new variables in the analyses which follow are 1) a categorical variable specifying the *approach class* of a flight and 2) a continuous variable estimating the *total energy* of an aircraft during the final approach phase of a flight.

2.2.1 Events Coincident with Unstable Approaches

Basic event counts are useful on a grand scale, but to understand which events occur more frequently during composite-event approaches, resulting in go-arounds or *unstable* approaches, we need to dig a little deeper. A new variable, **ApproachClass**, was defined to classify each flight with the following scoring:

0 when no events occurred during approach, final approach, taxi-in or go-around,

1 when maximum severity of any event during approach, final approach, taxi-in or go around is 1,

2 when maximum severity of any event during approach, final approach, taxi-in or go around is 2

3 when maximum severity of any event during approach, final approach, taxi-in or go around is 3 and at most three unstabalized approach indicator events are severity 3. (to partition the flights and avoid double counting)

4 when severity of any unstable approach is exactly 4

5 when severity of any unstable approach is exactly 5

The taxi-in phase was included in the definition for **ApproachClass** because unstable approaches are identified and quantified with a severity rating during the taxi-in phase. In general, severity classes 1 and 2 indicate values approaching but not above exceedance levels.

Once the **ApproachClass** variable is constructed we can look at a crosstabulation of event counts by approach class and maximum severity class as well as distributions of parameter values for different approach class categories.

2.2.2 Data Reduction Based on Correlation

Frequently, with large sets of correlated parameters and constructed variables, some correlations are so high (0.98 or higher in absolute value) that some variables can be considered redundant. That is certainly the case with the full flight parameter set and variables we constructed from them for this study.

We developed a screening method based on simply running pairwise correlations and removing redundancies with at least a correlation of 0.98 in absolute value.

2.3 Customization of Tool

A number of "customizations" were implemented by creating custom S-PLUS nodes inside I-Miner. By using the S-PLUS programming language, a number of specialty plots and data manipulations were created that no traditional data mining tool can do. These customizations could be helpful to other FDM data mining projects using I-Miner especially if data from a SAGEM system is being used.

3 Overview of Analysis Plan

After three days of intensive onsite discussions with JetBlue staff and a round of feedback on the task list we decided to focus on the following analyses to demonstrate data mining tools:

1. Enhancement of the current monthly report
2. Analysis of unstable approaches
3. Analysis of landing energy profiles
4. Data validation.

3.1 Monthly Report

Purpose is to inform non-specialists. Strategy is to keep them simple but informative. Traditional approach is event based. Distribution plots (i.e., histograms) are a recent addition to the report.

The project team developed the following possibilities for ways that I-Miner could supplement information currently captured by the monthly report:

- Principal Components Analysis (PCA) loadings. These statistics demonstrate relationships amongst variables (i.e., parameters) related to the event(s). By looking at correlation structure of parameters, which are highly correlated to events, we can provide details on event scenarios or constellations. These analyses can identify flights which may not generate event signals but are so near the boundaries throughout the flight phase under consideration that they are worthy of further study.

- Conditional (Trellis) graphics. These may be useful for multi-dimensional displays and are better than 3-D because they extend beyond 3 dimensions in an easily comprehensible way. Parameter histograms (e.g., speed high conditional on flap configuration) rather than simple event counts will provide a more complete picture of parameters that define critical events.
- Correlation of different events. Specifically focus on go-arounds (discontinuing an approach and instead adding power to climb to higher altitude, retract the gear and flaps as in an initial takeoff, and maneuver for another approach).
  - A Trellis graph example with the first branch Go VS No Go, second branch Stable VS Not Stable (Class 1,2) VS Not Stable (Class 3, 4, 5), third and lower branches other events, principle parameters, etc.
  - An unstable approach should result in a go-around, but rarely does. When (and at what severity level) does an unstable approach produce a go around? Which component events are most predictive? Are there other events highly correlated but not part of the unstabilized approach definition?
  - A go-around often is causal in other events such as flap over speeds. What events are correlated? For what parameter values just prior to the go around?
  - Go-arounds are often directed by the control tower. Can externally caused go-arounds be distinguished from pilot initiated go arounds?
  - What parameters are principal in go-arounds?
  - Reducing go-arounds would improve both safety and costs and are a high interest item at all airlines.

3.2 Unstable Approaches

Landing approach deviations occur that are a normal part of the dynamic element of flying and are controllable but outside of the limits of expected performance. The term "unstable approach" as used in this report refers atypical (compared to the average) parameter values and *not* the stability of the aircraft.

We investigate approach and final approach landing events and associated parameters and investigate multiple-event (multiple occurrences of single events) approaches using event data and full parameter data.

a) Correlate events (analysis of contingency table)
b) Correlate events with snapshot parameter values
c) Correlate events with summaries of 2-3 minutes of full–flight parameter values.

This study was designed to characterize unstable landings in terms of *event* profiles and parameter correlations. PCA, tree models and neural net models were used to quantify relationships between parameters and the likelihood of an event or constellation of events during approach or final approach phases of a flight.

All three analyses will be completed, time permitting. If there is insufficient time to do all three they will be completed in the order listed until time is exhausted.

3.3 Landing Energy Profiles

The project plan was to define a nominal or ideal energy profile for an approach (which may be airport runway specific) and measure deviations from the profile. The total energy should decrease as altitude is reduced until level over the runway with only kinetic energy from airspeed. The "safe" deviation limits will get narrower as the aircraft gets closer to the ground. At 2000' an aircraft can safely recover from a wind shear that produces a 1000' loss of altitude before achieving an on-speed climb. At lower altitudes, this same event might be beyond the capabilities of the aircraft and aircrew to recover.

The strategy for this part of the analysis was to:

1) Eliminate "visual" approaches which are not on-line from 1000' altitude.
2) Develop a measure for total energy. (The confidential Analysis Report provided to JetBlue contains more details.)
3) Develop an average ("typical") approach by averaging approach profiles for all non-event approaches.
4) Develop profile confidence bands – 90%, 95%, 99%.
5) Examine characteristics of extreme profiles, > 95%, > 99%
   Develop a total energy measure for the approach, then look at
   - summary stats
   - graphical displays tied to events
   - parameter correlates
   - clustering on integrated total energy
   - tree models?
   - other procedures?
6) Examine relationship between multiple-event (multiple occurrences of single events) approaches using event data, full parameter data and calculated energy variable. See Analysis 2.

3.4 Data validation

The task discussed is to examine correlated parameter sets and quantify correlation, regression slope coefficients and variation. The strategy is to analyze full-flight parameter summaries (e.g., max, min, average, median, value at given altitude) for various phases of a flight. Parameters, which are highly correlated, should have predictable behavior as a function of other parameter(s) to which they are strongly correlated. For highly correlated parameters, knowing the values of one parameter will provide target and acceptable range of values for the other correlated parameters. This study could lead to automated procedures for data validation of use for any of the full flight data sets.

4 Analysis

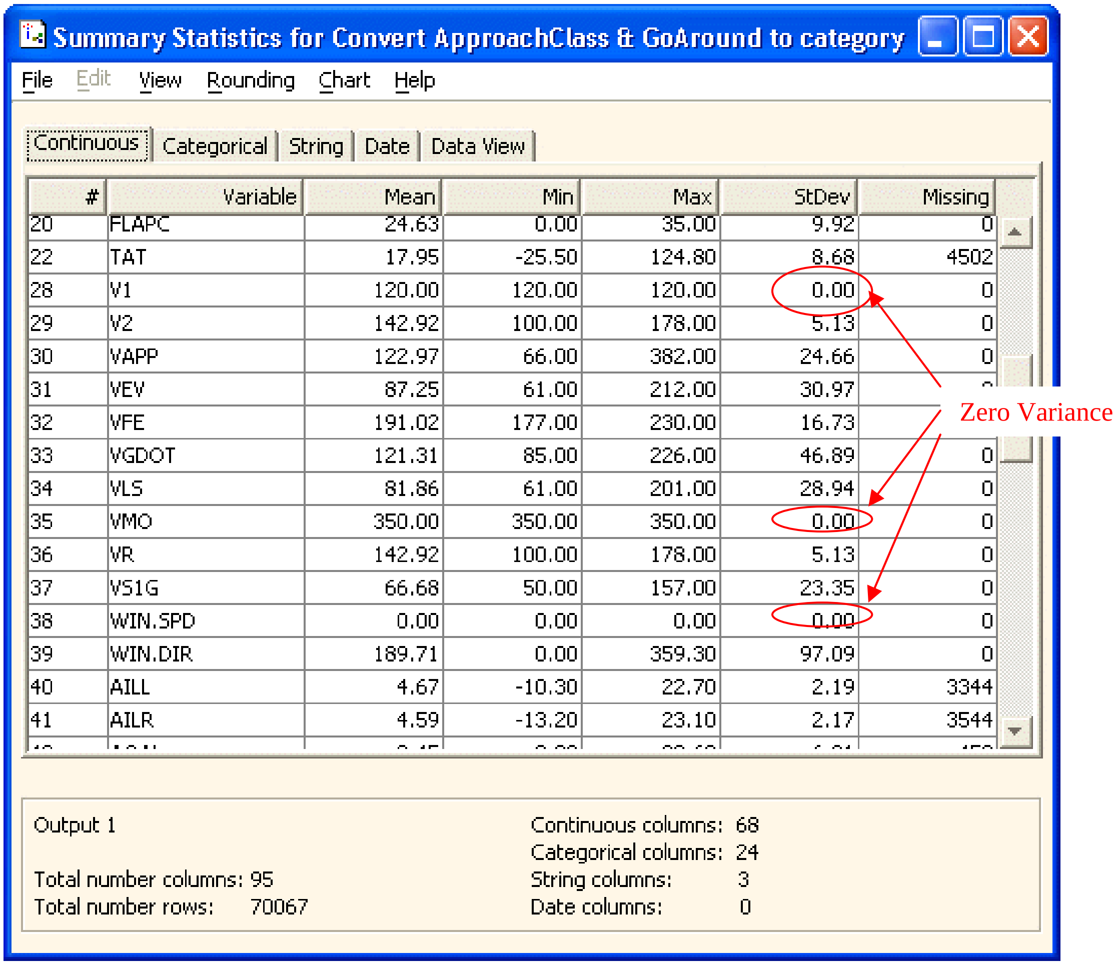
This section presents an overview of the results from this analysis. (A more detailed report on the analysis has been provided to JetBlue in confidence, as it contains various data that may be considered company-sensitive.)

4.1 Summary Statistics

4.1.1 User inputs/manipulations/decisions

All Insightful Miner (I-Miner) nodes provide summary statistics including missing value counts without any extra effort. Summary statistics for all data available for calculation at any given node are immediately available by opening the Table Viewer displayed in Figure 17. The Table Viewer was also used to discover which parameters had high numbers of missing values. Given the volume of data we decided to eliminate rows with missing values rather than substitute or impute values to replace missing data. Consequently variables were deleted when they had large numbers of missing values which, when deleted, substantially reduced the number of go arounds.

4.1.2 Example results



Summary Statistics for Convert ApproachClass & GoAround to category
File Edit View Rounding Chart Help
Continuous Categorical String Date Data View

#	Variable	Mean	Min	Max	StDev	Missing
20	FLAPC	24.63	0.00	35.00	9.92	0
22	TAT	17.95	-25.50	124.80	8.68	4502
28	V1	120.00	120.00	120.00	0.00	0
29	V2	142.92	100.00	178.00	5.13	0
30	VAPP	122.97	66.00	382.00	24.66	0
31	VEV	87.25	61.00	212.00	30.97	0
32	VFE	191.02	177.00	230.00	16.73	
33	VGDOT	121.31	85.00	226.00	46.89	0
34	VLS	81.86	61.00	201.00	28.94	0
35	VMO	350.00	350.00	350.00	0.00	0
36	VR	142.92	100.00	178.00	5.13	0
37	VS1G	66.68	50.00	157.00	23.35	0
38	WIN.SPD	0.00	0.00	0.00	0.00	0
39	WIN.DIR	189.71	0.00	359.30	97.09	0
40	AILL	4.67	-10.30	22.70	2.19	3344
41	AILR	4.59	-13.20	23.10	2.17	3544

Zero Variance

Output 1

Total number columns: 95
Total number rows: 70067

Continuous columns: 68
Categorical columns: 24
String columns: 3
Date columns: 0

Figure 17: Data Viewer of I-Miner. This summary statistics page shows some variables have standard deviation (and variance) of zero indicating that the values for those variables did not change over the entire set of 70,067 events.

4.1.3 Findings

Some variables were not recorded throughout the flight producing zero variance (equivalent to zero standard deviation). These variables were eliminated from further study. Additionally, some parameters had many missing values. When row-wise deletion of missing values substantially reduced the number of go arounds, the variable was dropped rather than dropping missing values row-wise.

4.2 Basic One- and Two-Dimensional Graphics

4.2.1 User inputs/manipulations/decisions

Data were checked for validity using simple histograms like those displayed in Figure 18 and Figure 19. Graphics such as these make it very easy to check for anomalies like those displayed by **GW** (gross weight) and **AOAR** (angle-of-attack).

Bad data like that displayed in Figure 20 are harder to spot because the constant (bad) values are in the center of the distribution of values. However, the bad values stand out when used in the total energy calculations. Once total energy was computed, the flight depicted in Figure 20 stands out as displayed in Figure 21. The decision to create Figure 20 and Figure 21 was inspired by trying to understand why the distribution for the class 3 approach at elevation 200 feet was so much broader than at other altitude. See Figure 24 for relevant detail.

4.2.2 Example results

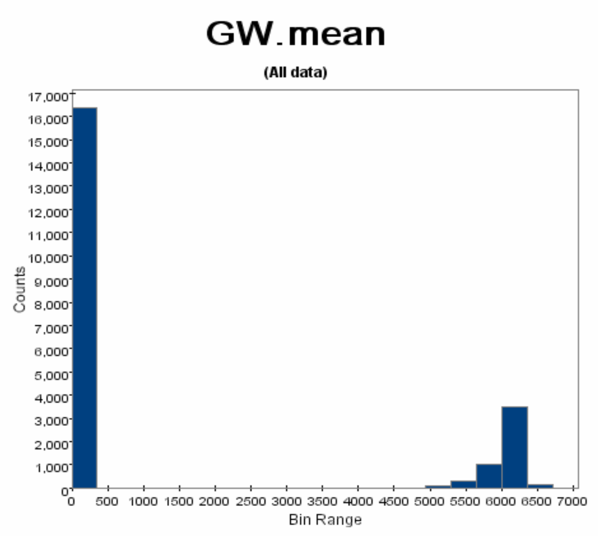


Figure 18: Histogram for gross aircraft weight averaged over all events during the approach phases of each flight. The large spike at zero indicates many flights had zeros recorded for gross weight and stored in the GW variable.

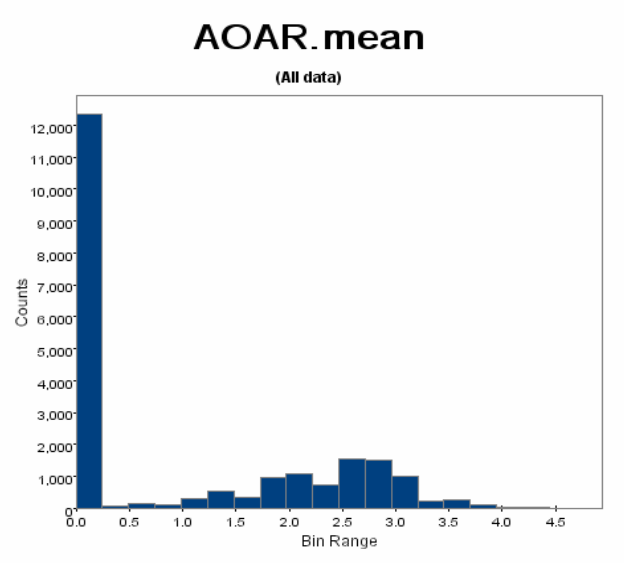


Figure 19: Histogram of indicated angle of attack averaged over all events during the approach phases of each flight. The large spike at zero indicates that many flights did not have indicated angle of attack recorded and stored in the AOAR variable.

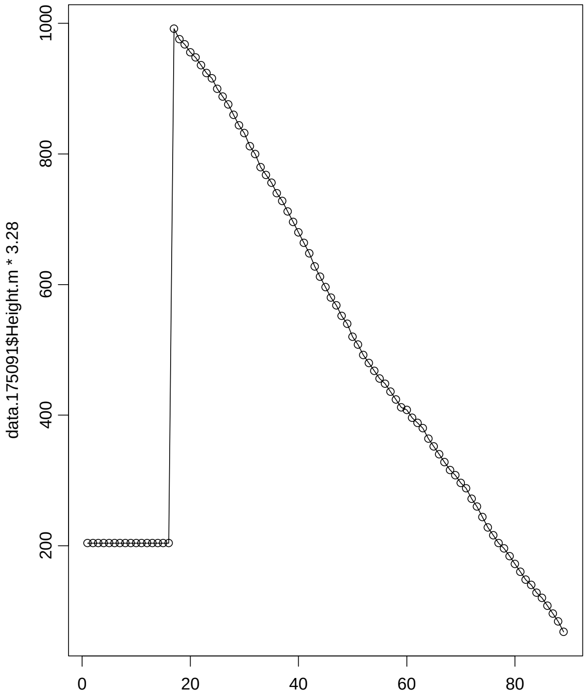


Figure 20: Plot of altitude vs. sequence of observation showing bad initial values for altitude.

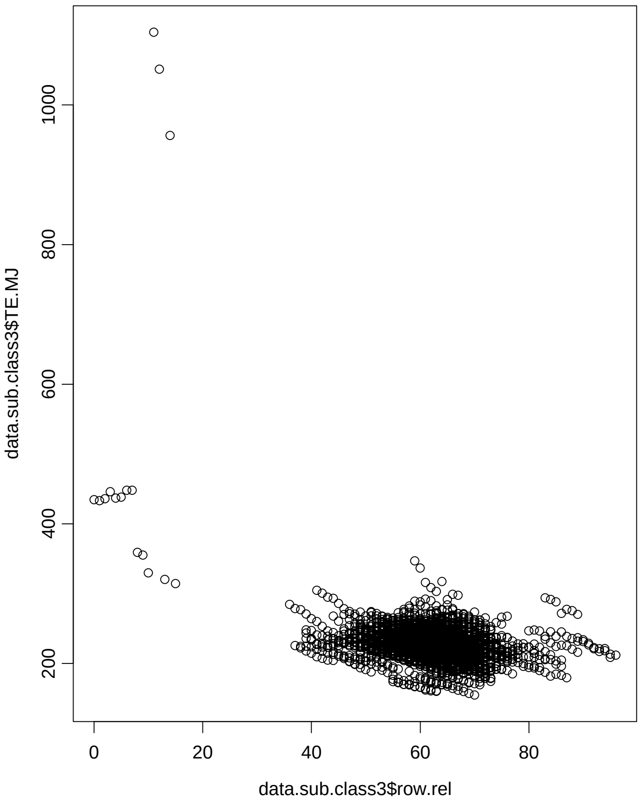


Figure 21: Total energy for all class 3 flights at specific altitude (feet elevation) versus sequential row number. All the energy values for row.rel less than 20 belong to the improperly recorded flight data displayed in Figure 20.

4.2.3 Findings

The data were not as clean as originally claimed. Being able to do quick checks and variable deletion or follow-up graphics as necessary allowed the project to proceed at a reasonable pace.

4.3 Trellis Graphics

4.3.1 User inputs/manipulations/decisions

Figure 22 displays percentages of events by approach classification. This is a Trellis graph that displays the five groups side by side for easy comparison. Nodes were created to aggregate flights by **ApproachClass** and by **EventShortname** and to convert counts to percentages. Percentages for events in approach class 5 were sorted in decreasing order prior to complete this plot. A similar plot was produced for the **GoAround** classification variable. Trellis graphs allow easy comparison between groups.

4.3.2 Example results

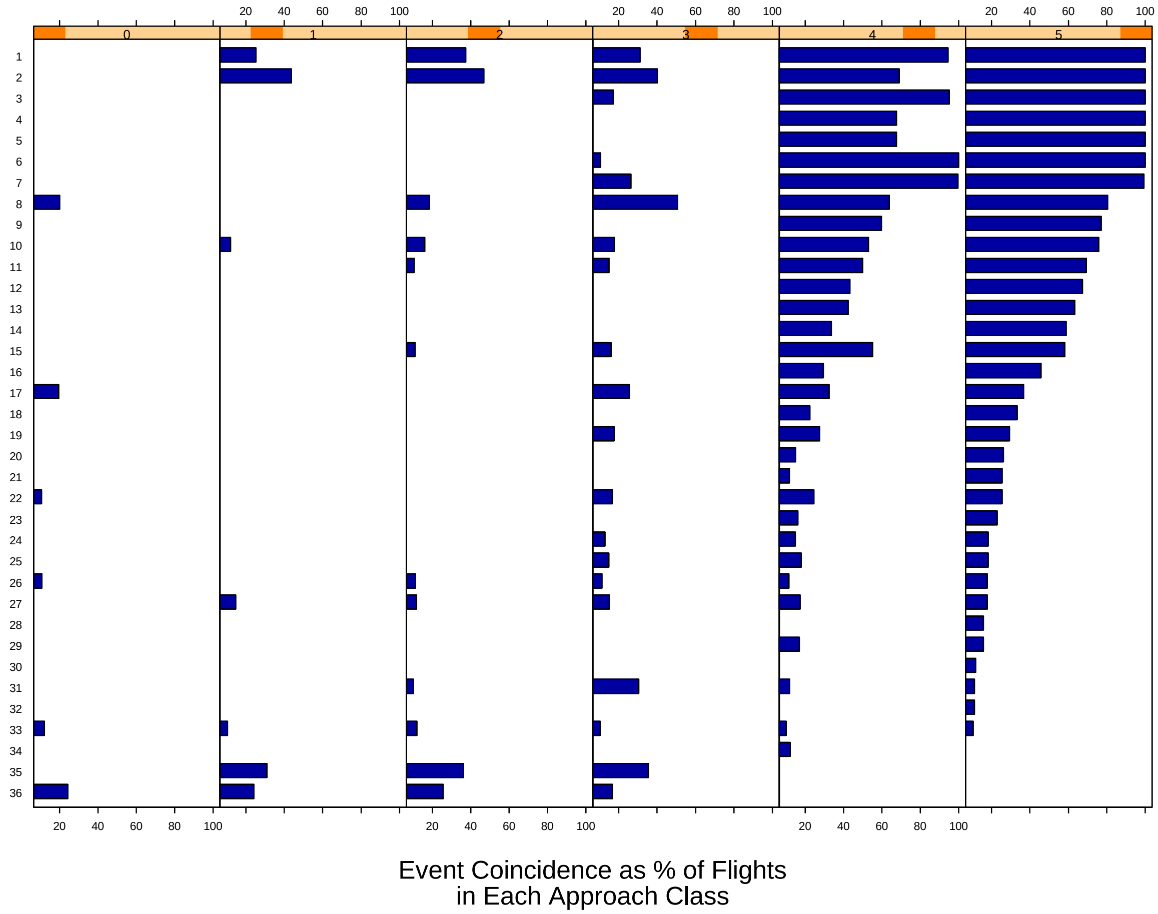


Figure 22: Coincidence of events with Approach Class. Graphics like this help create an event "profile" for approach classes.

Figure 23 displays a trellis graph of two principal components (PC2 vs. PC3) conditional on two other principal components (PC1 and PC4). This graph is used to show that extreme values on some principal component axes clearly separate go arounds from no go arounds.

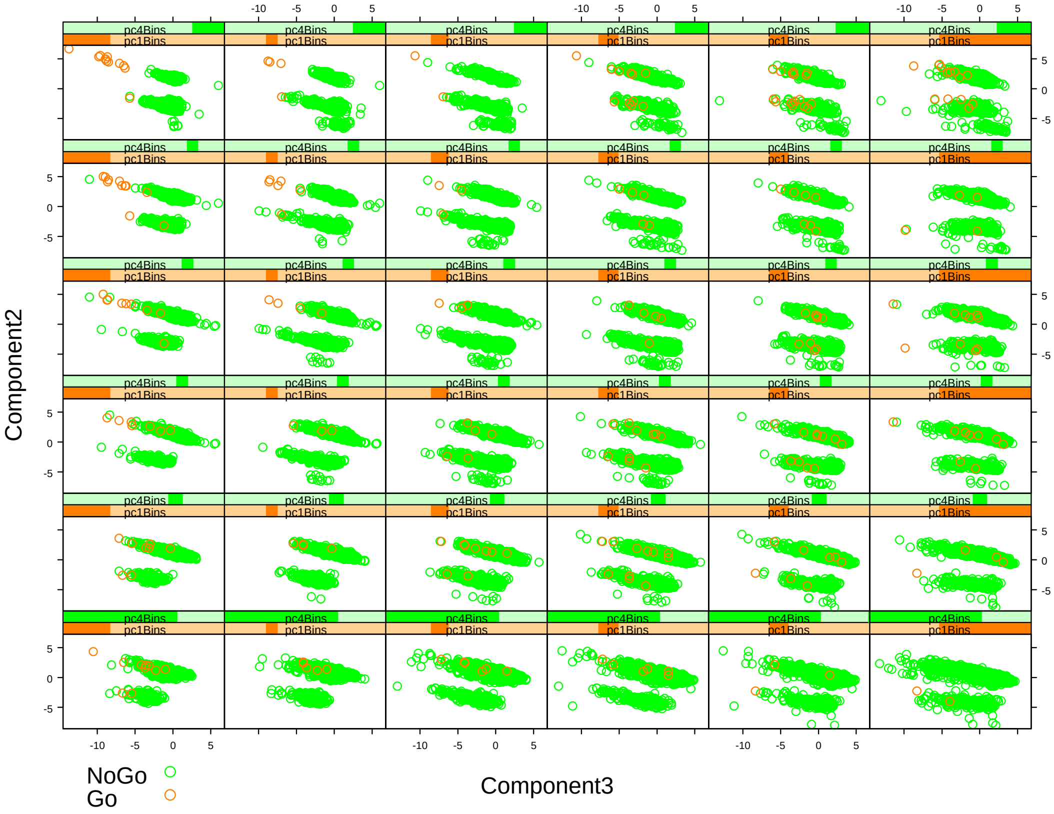


Figure 23: Trellis scatterplot of PC 2 vs. 3 conditional on bins for PC 1 and 4. Green and orange points represent no go around and go around respectively. Note the very nice separation of go arounds from no go arounds in the upper left panels. These panels correspond to low PC1, high PC4, low PC 3 and high PC 2.

4.3.3 Findings

Trellis graphs are powerful for 1) comparing groups and 2) for locating extremes where subgroups are denser. Figure 23 displays a region in principal components space where go arounds are much denser.

4.4 Custom Graphics

4.4.1 User inputs/manipulations/decisions

The nature of this study necessitated explorations in non-standard ways. By coupling S-PLUS with I-Miner, S-PLUS could be used to create graphics of subsets and data summaries generated in I-Miner. This provided the flexibility of S-PLUS in the context of the scalability of I-Miner. Figure 24 shows one example of a sequence of custom graphs using subsets and aggregation generated in I-Miner and a custom S-PLUS graphic. The chart sequence shows density estimates of total energy by altitude above the runway and by approach class within each panel. For energy conservation, JetBlue procedures allow for higher energy approaches in day visual flight conditions. The 900' chart clearly shows nominal and higher energy approach profiles. At lower altitudes the chart sequence shows a rapid energy reduction in the high-energy profiles to match the nominal profile distribution.

4.4.2 Example results

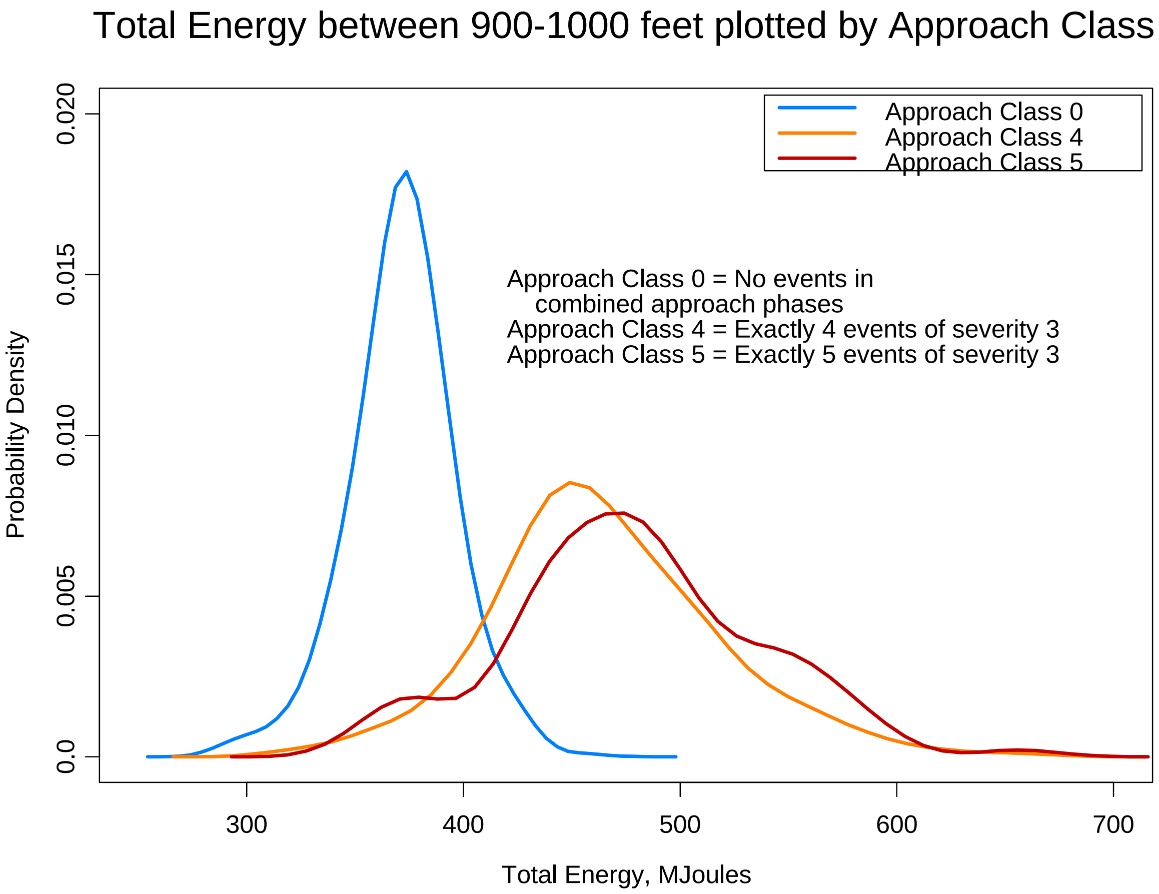


Figure 24: Estimated probability density distributions of total energy at 900 feet in altitude above the runway.

Figure 25 is another custom graphic displaying the differences in distributions for the approach classes in another way. The plot shows average total energy versus distance from touchdown for combined approach classes overlaid on percentile bands to give a sense of how extreme the approach class 4,5 is compared to the rest.

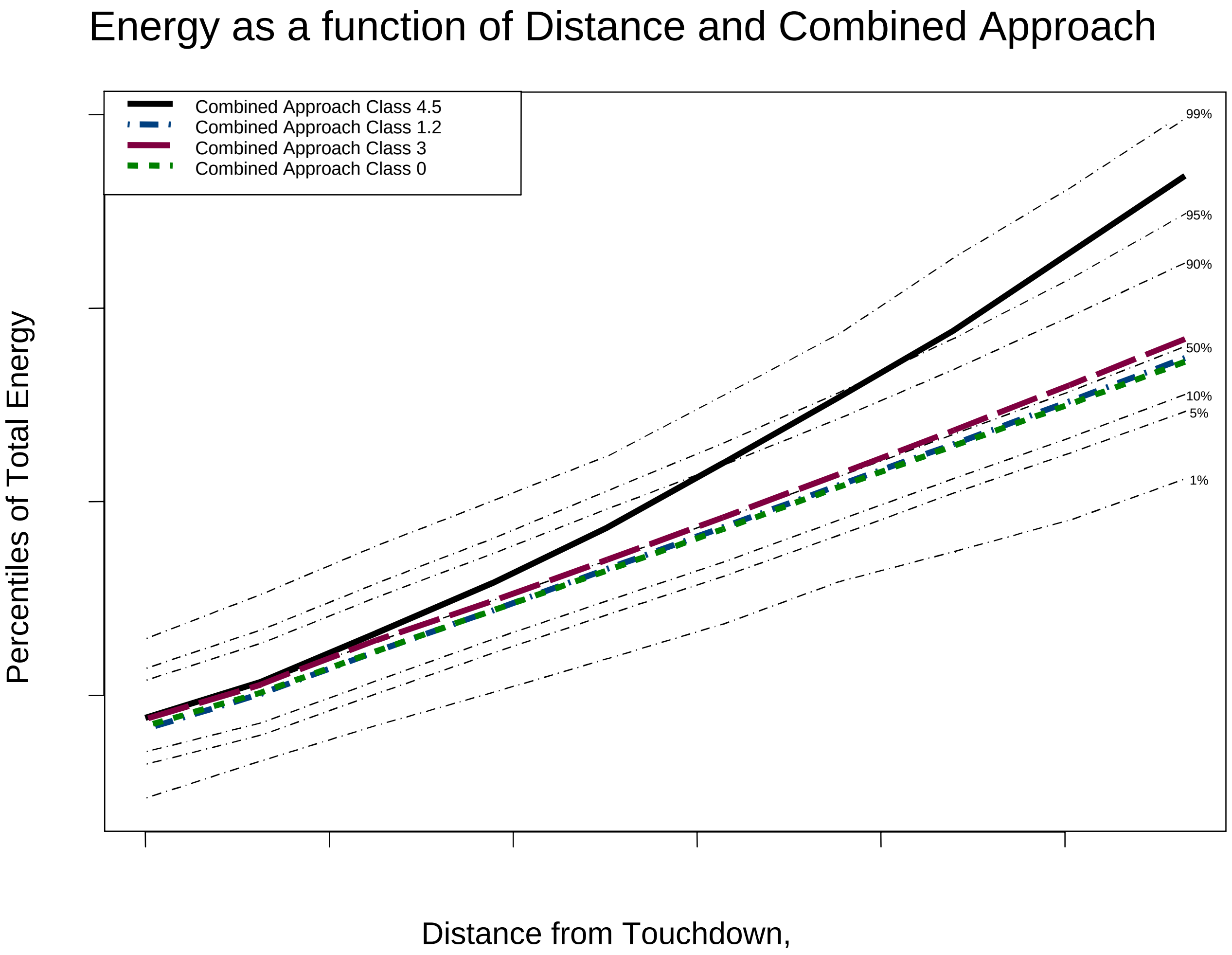


Figure 25: Average total energy by combined approach class with percentile bands representing all flights. The profile for the combined approach class, 4,5, starts above the 90% percentile but merges with the average energy profile sometime before landing.

Figure 26 displays a grayscale plot of total energy versus height above runway for approach class 4 with reference line for approach class 0 added to enhance comparison with no-event approaches.

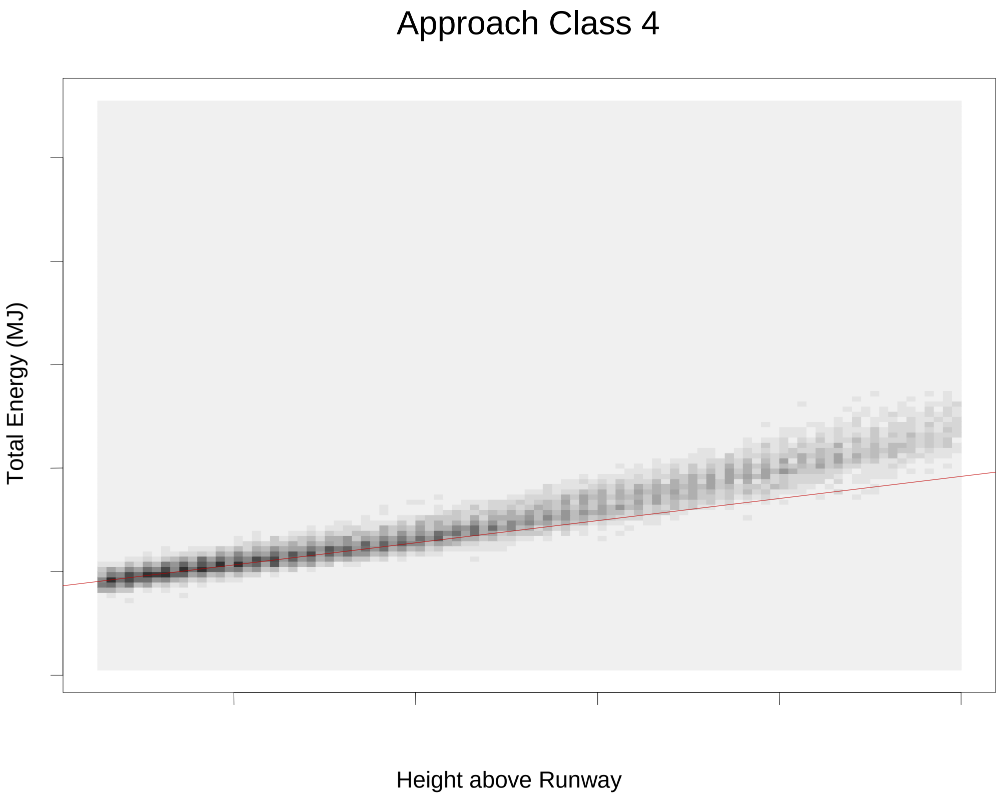


Figure 26: Energy vs. height for approach class 4. The red line is the average approach for approach class 0 assumed to be the ideal, no-event reference approach.

Figure 27 displays engine thrust versus height above the runway for approach class 3. Note the low thrust states early in the approach as the aircraft drop energy. Figure 27 also displays a sense of the spool-up time for engines to go from minimal (idle) thrust to typical thrust levels on approach.

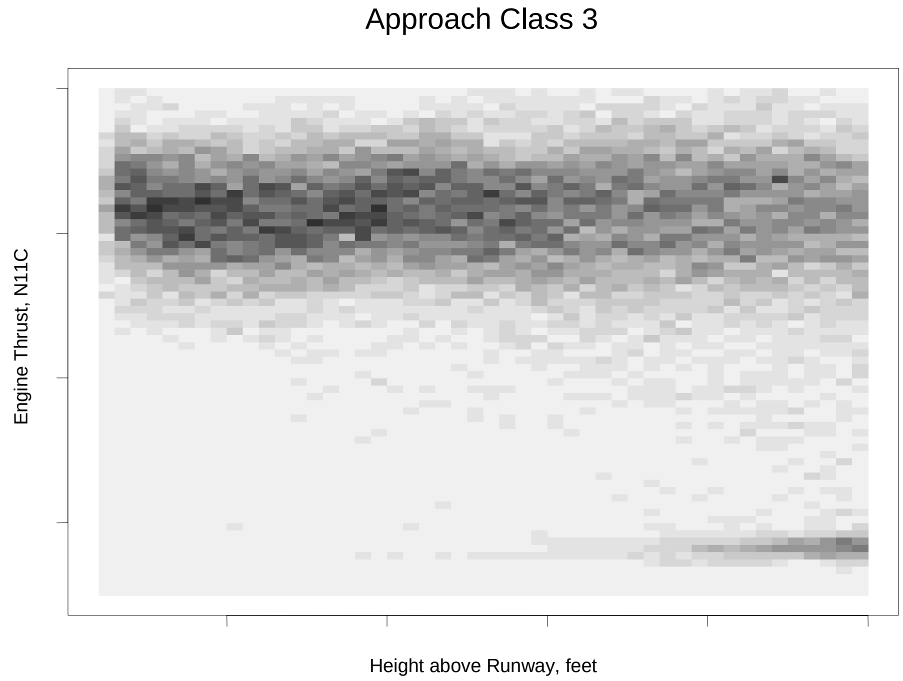


Figure 27: Plot of engine thrust vs. height above runway for approach class 3.

4.4.3 Findings

Custom graphics provide a way of teasing out displaying results that get to the heart of fundamental questions. The total energy plots in Figure 24, Figure 25 and Figure 26 dramatically show the higher energy levels of class 4 and 5 approaches. Figure 27 not only shows the low thrust states of class 4 and 5 approaches but also gives a sense, in practical terms, of the time (in terms of lost altitude) it takes for engines to spool up to approach thrust levels.

4.5 Correlation Filtering

4.5.1 User inputs/manipulations/decisions

In the context of aggregating data over time or some other dimension, a number of new computed variables, such as average, minimum and maximum, are typically created. These variables are often very highly correlated. Furthermore many of the flight parameters are highly correlated due to the way they are measured or the way flight operations are standardized. For example, one flight parameter measures thrust on the left engine and another one measures thrust on the right engine. Since engine thrust levers are usually moved in synchrony during operation, the thrust measurements are usually very highly correlated.

Very high correlation is synonymous with redundancy. For this study we created nodes for filtering out redundant variables. We set a threshold at correlation equal to 0.98 or higher in absolute value, keeping only one variable out of each redundant set. We used graphics similar to the one displayed in Figure 28 as a quick check on data redundancy.

We also used the graphic displayed in Figure 28 as a discovery tool to give a general sense of which variables and flights were most similar. Figure 28 shows a "clustered heat map" presentation of *scaled* data summaries for a number of flights. We have selected 1700 flights that represent a balanced cross section of the Combined Approach Classes 0, 1/2, 3, 4/5.

Data summaries for each flight are created in a two-step procedure.

- We first create six separate summaries for each raw parameter: minimum, mean, and maximum values in the two height windows 400-500 feet and 900-1000 feet above the runway. This process creates a large number of potential variables of interest, many of which are redundant number of columns in the data matrix increases from 131 to 385 in the process.
- We use the correlation structure of these variables to eliminate variables with correlations of 0.98 or greater, a process that cuts down the number of variables from 385 to 160.

4.5.2 Example results

The gray-shaded matrix in the central part of Figure 28 shows the data, where dark signifies large values and white small values. There are 1700 horizontal rows in the matrix; each row corresponds to a single flight. There are 160 vertical columns shown here; each of these corresponds to a single parameter summary.

One distinguishing feature of this figure is that the rows and columns of the figure have been re-arranged automatically using clustering methods, in order to place similar rows and similar columns close to one another. Vertical blocks of very similar columns are apparent.

Natural groupings occur in the parameters, for example the engine command and power data are quite similar to one another. We use statistical clustering methods to re-arrange the rows and columns of the data matrix, to group similar parameters and similar flights together and natural groupings to code the parameters for to aid in reading the graph. A legend for the parameter color coding follows immediately after Figure 28.

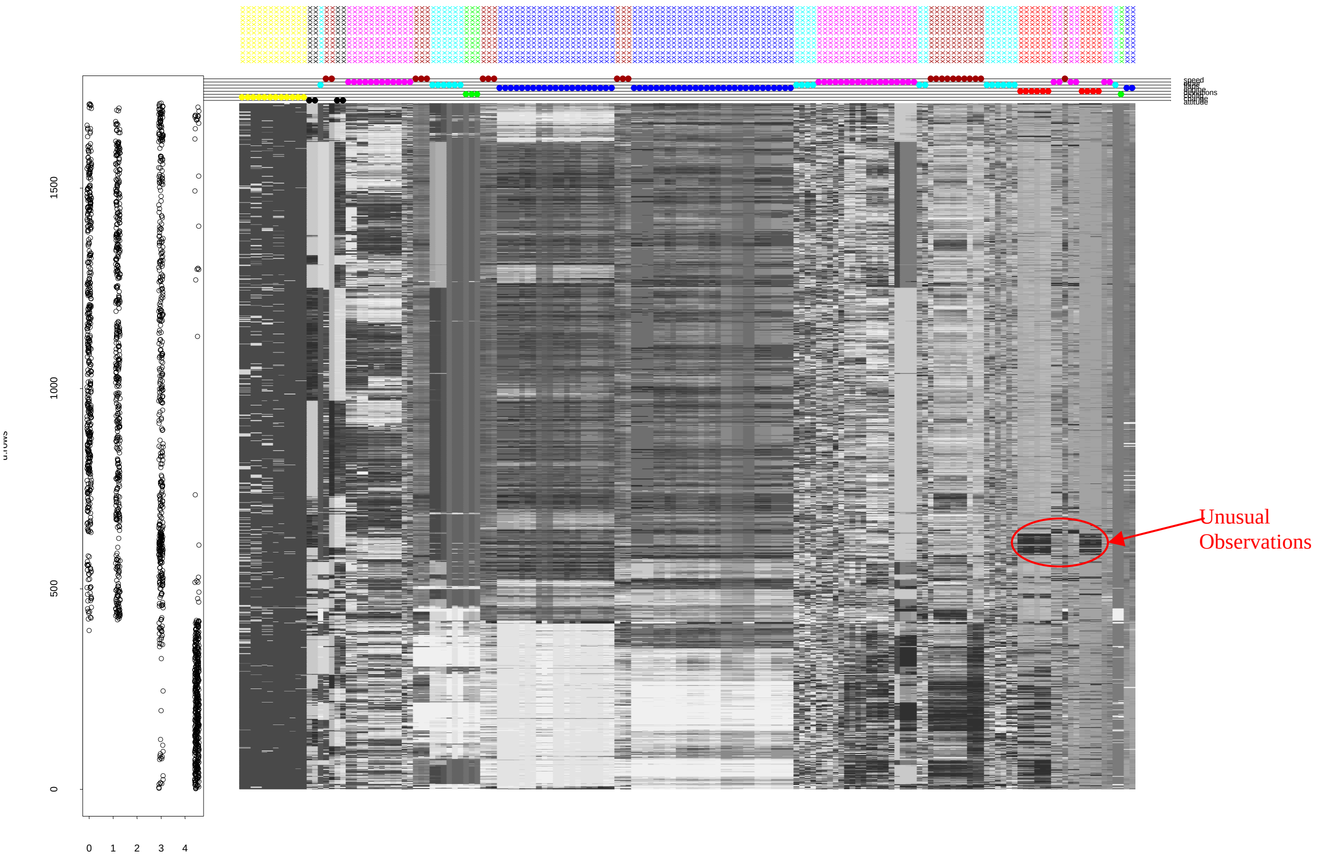


Figure 28: A clustered heat map presentation of data summaries for a 1700 flights that represent a balanced cross section of the Combined Approach Classes 0, 1/2, 3, 4/5.

- ● other
- ● speed
- ● flaps
- ● engine
- ● deviations
- ● config
- ● change
- ● attitude

Parameter Color Coding

In addition to the central matrix portion of the graph there is a vertical side panel on the left side of the graph. The side panel indicates which rows belong to each approach class by placing a plot symbol alongside the row and in the column corresponding to the approach class.

Figure 28 is used primarily as a *discovery* tool to understand how parameters relate to the approach classes and to identify unusual cases. Some obvious landmarks reveal information consistent with intuition. The color-coded variables across the top of the graph are fairly consistent within each color group, i.e., they cluster together. For example, one color group represents speed parameters and another represents power parameters. The graph also reveals that the combined approach class 4, 5 has quite different values for most of the parameters displayed than flights in the other classes. The white patch at the bottom of the graph corresponds to low engine power. Virtually all the power-setting parameters are low for approach classes 4 and 5. The heatmap confirms that class 4,5 approaches have low power settings compared to the other approach classes.

Figure 28 can also be used to find atypical flights that may be worth investigating when they behave different from the rest. For example, the circled region on the right side of the graph represents a small subset of flights that behave differently from the rest.

4.5.3 Findings

Correlation studies and filtering are valuable tools for

1. Eliminating and/or reducing redundancies in the data,
2. Understanding the correlation structure and,
3. Finding unusual relationships.

Graphics like that displayed in Figure 28 provide a quick graphical assessment of the correlation structure.

4.6 Principal Components Analysis

4.6.1 User inputs/manipulations/decisions

Principal components analysis provides a mechanism for both data reduction and characterization of correlation structures. With the bulk of the data being explained by only a dozen or so variables it may be easier to discover dimensions which discriminate subsets of interest.

4.6.2 Example results

Figure 29 shows the reduction in variance for the first twelve principal components (82%) computed from 45 variables. Over 50% of the variance is explained in the first four components.

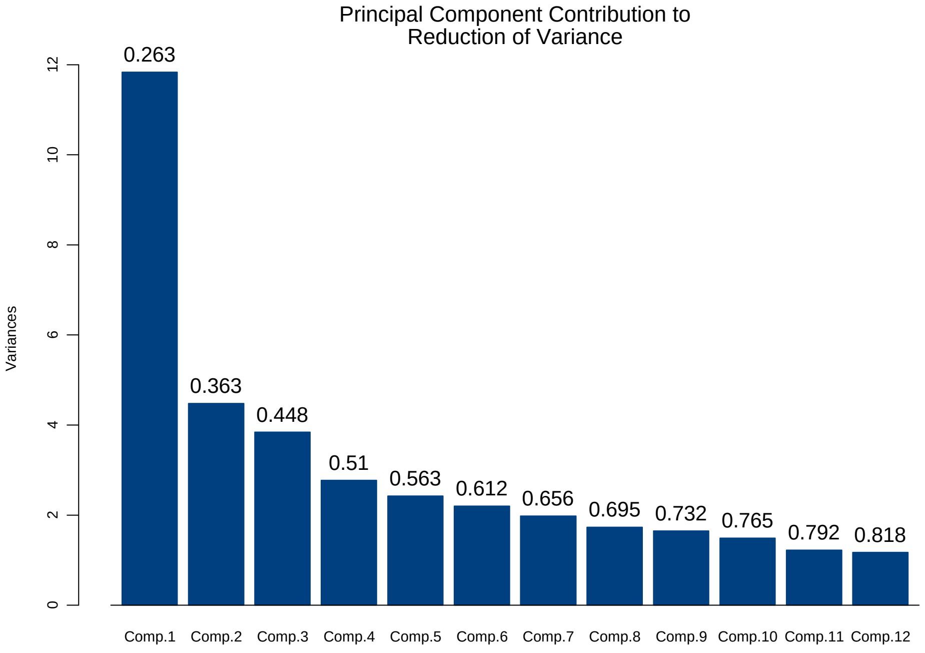


Figure 29: Overall variance reduction due to the first twelve principal components.

Figure 30 displays the principal component loadings (coefficients) for the first 12 principal components. Using the loadings and a bit of interpretation, we can characterize each component by the variables that dominate in each component definition. For example, the first principal is characterized as the component that contrasts power and speed (e.g., low power and high speed). Note that many of the higher components are dominated by only a few (less than five) variables.

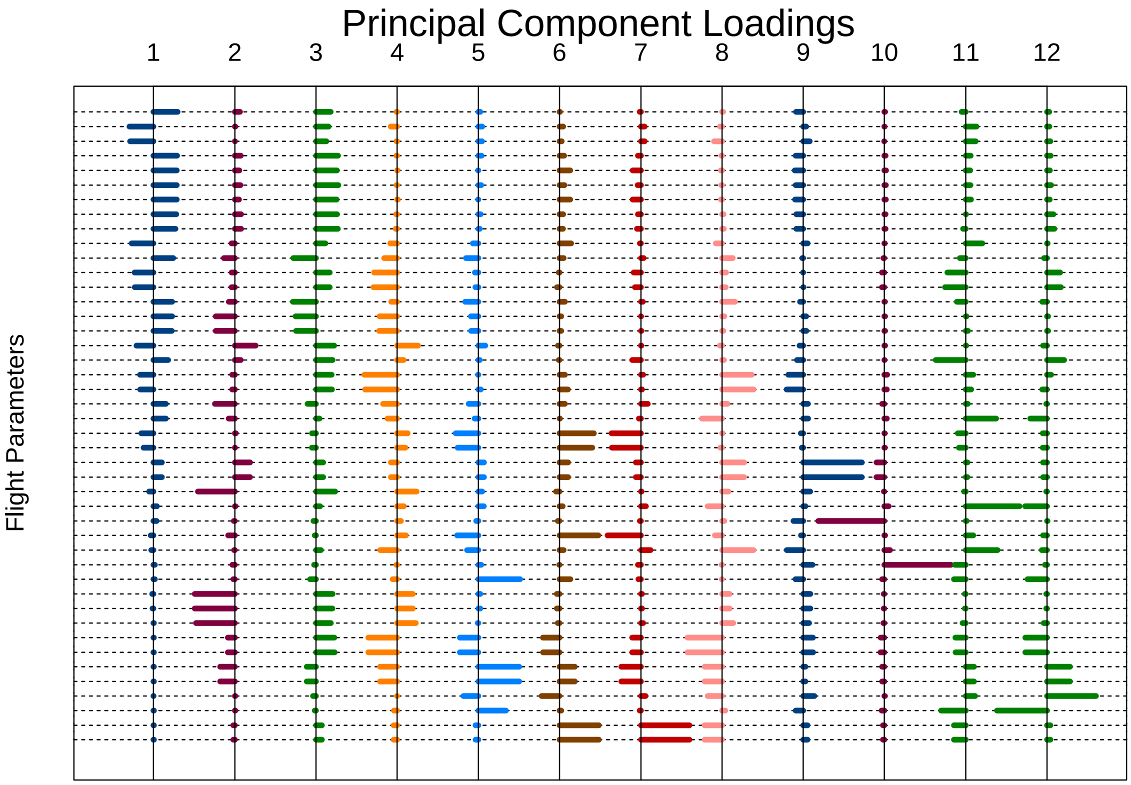


Figure 30: Principal component loadings plot which shows each principal component as a combination of flight parameters. The longer the bar (in absolute value) the larger the influence a variable has on a principal component. Consequently, the principal components can be "characterized" by the variables with the largest loadings. Bars pointing to the left have negative signs and bars pointing to the right have positive signs.

4.6.3 Findings

By using principal components in graphics like that displayed in Figure 23, it may be possible to find dimensions where atypical approaches (e.g., predominately go arounds) pop out.

4.7 Tree-Based Models

A tree-based model is a very useful tool for finding subsets of data which are most similar. It provides a natural way of finding variables, or rather, subsets of values of variables where, for example, go arounds are more frequent.

4.7.1 User inputs/manipulations/decisions

Figure 31 displays a tree modeling go arounds as a function of the snapshot parameters, both continuous and categorical (total of 51 variables). Note how cleanly many of the go arounds are separated from the no go arounds. In particular, the bottom two nodes on the tree correspond to the low, fast go arounds.

4.7.2 Example results

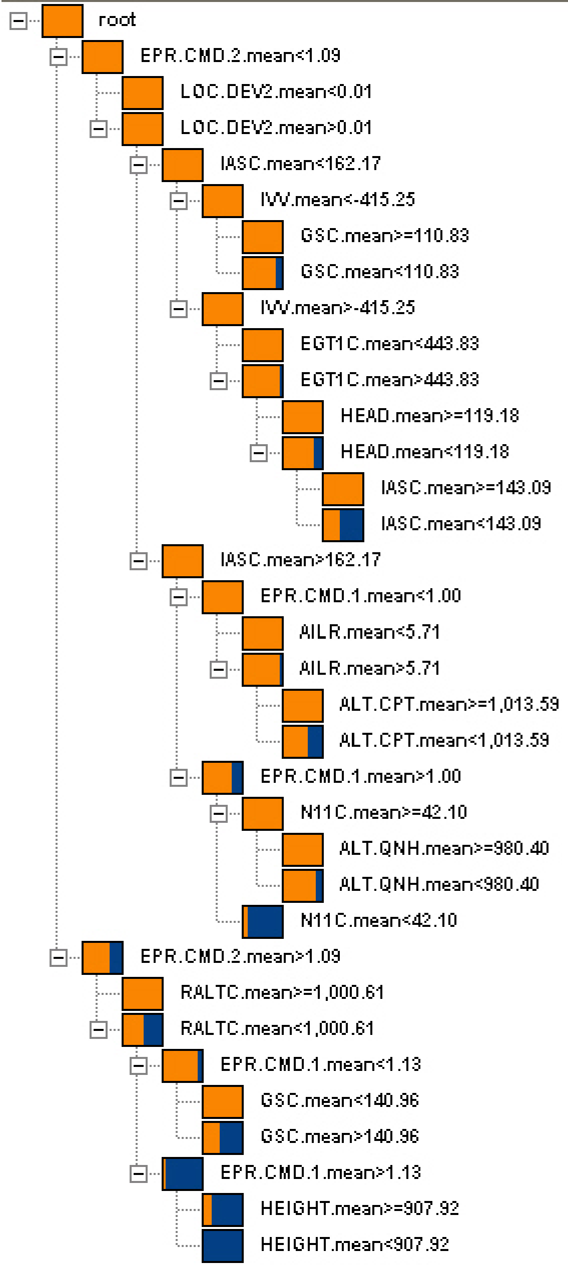


Figure 31: Tree model of GoAround response as a function of snapshot parameters. The blue and orange boxes are simple histograms indicating homogeneity of each node of the tree. Blue corresponds to go around and orange to no go around. Note that some nodes are perfectly homogeneous, indicating that those nodes are predicted perfectly. In particular high speed, low altitude approaches are perfectly predicted to go around.

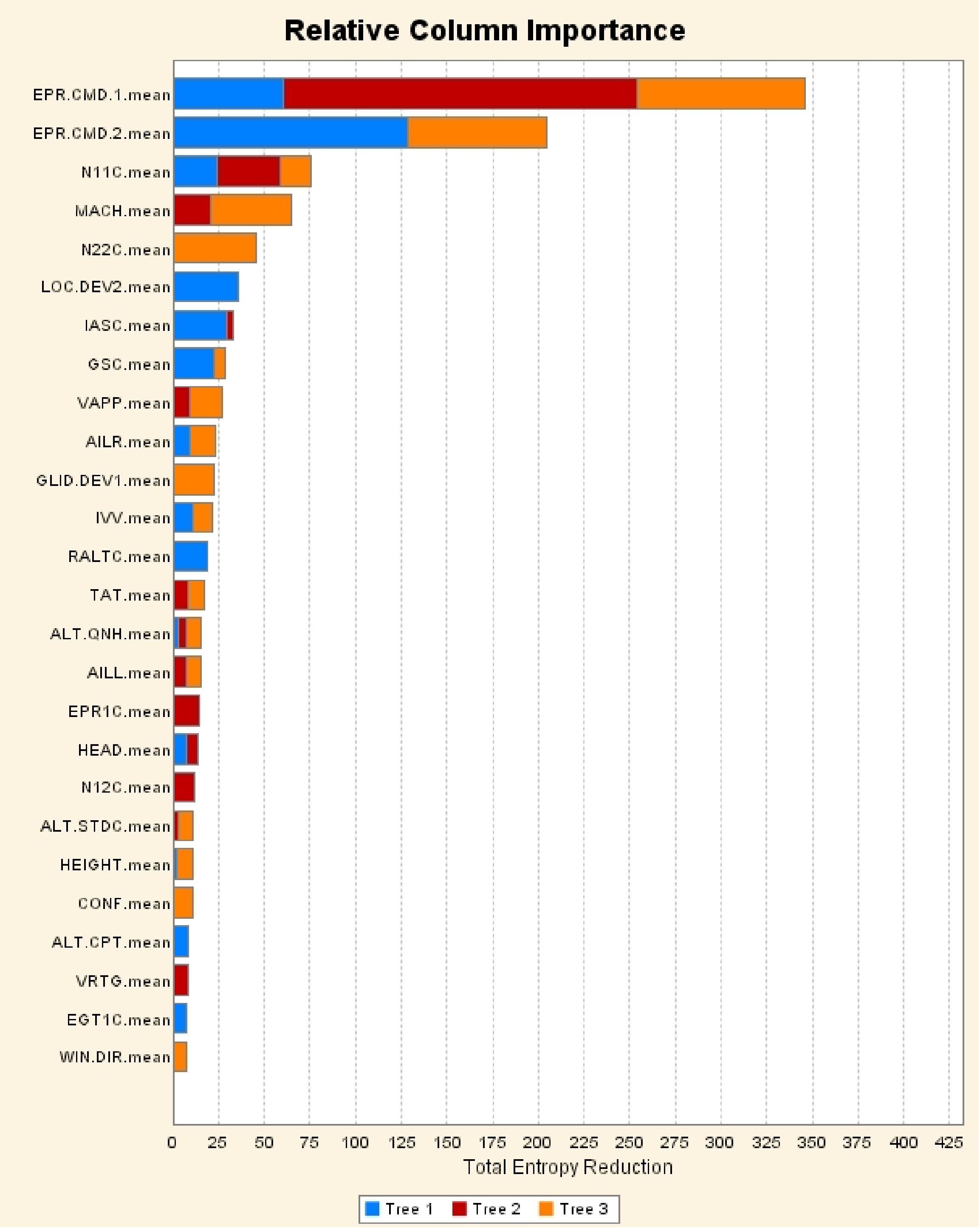


Figure 32: Combined relative importance of variables used to construct 3 different tree models from 3 mutually exclusive subsets of the data. This plot indicates that (high) power settings are the single biggest indicator of a go around.

In addition to finding regions where some event of interest in more likely to occur, tree ensembles can provide a quantitative measure of variable importance. Figure 32 displays a plot showing relative column importance. Note, that this figure can be somewhat misleading when the event of interest has a small probability of occurrence because the variable(s) (**HEIGHT** in our example) which finally separates out the (small) group of interest may have little impact on the overall performance of the model.

4.7.3 Findings

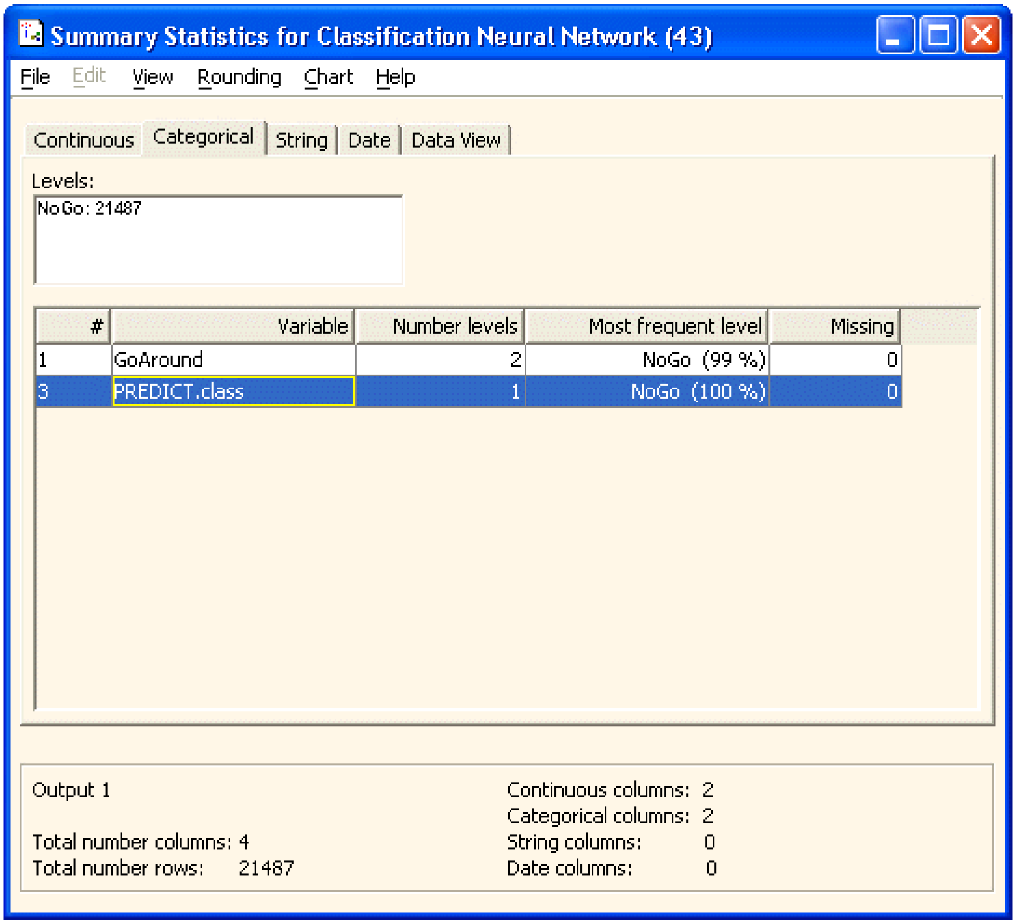
A tree model is a powerful tool for finding subsets of predictor values which have a higher prevalence of an event of interest. The example displayed in Figure 31 shows a number of homogeneous nodes that show commonalities in approaches that result in go arounds.

4.8 Neural Networks

4.8.1 User inputs/manipulations/decisions

Neural networks are commonly used data mining tools so they have been included here for comparison. We did not compare prediction accuracy between tree models and neural networks because neural networks failed to answer the most basic question of the study of go arounds – which variables (or subsets of variables) are more highly associated with go arounds.

4.8.2 Example results



Levels:

NoGo: 21487

#	Variable	Number levels	Most frequent level	Missing
1	GoAround	2	NoGo (99 %)	0
3	PREDICT.class	1	NoGo (100 %)	0

Output 1

Total number columns: 4
Total number rows: 21487

Continuous columns: 2
Categorical columns: 2
String columns: 0
Date columns: 0

Figure 33: Output of neural network model showing all predictions as NoGo.

4.8.3 Findings

Fitting a neural network model to the same data, using the same predictors as for the tree model, the neural network failed to distinguish the go arounds from the no go arounds. The best fitting models predicted all flights as no go arounds, the safe prediction, when over 99% of the flights did not go around.

By drawing a sample from the no go class equal in size to the go around class, a neural network model could be developed that does distinguish between the groups. However, neural network models are more difficult to interpret than tree models and doing preliminary sampling adds another layer of complexity to the modeling process.

5 Assessment of Results by Partner Airline

The analysis conducted during the proof-of-concept delivered useful and intriguing results. The detailed analysis report provided by Insightful both validated and expanded on previous JetBlue results and identified new areas for further study. In addition to the specific analysis provided by Insightful, it was valuable to have outside experts review and comment on JetBlue's current analytical methods and reporting process. Although the project scope was quite limited, the data mining methods show great potential for providing airline managers with clear, relevant insights that lead to improvements in flight safety, operations, and training programs.

The project also identified a number of barriers that other airlines or industries may encounter in attempts to expand their analysis capabilities through data mining techniques. Investigation of these barriers may prove as useful as studying how the methods can be applied in an aviation safety setting.

5.1 Observations on Mining Tool and Suggestions for Improvement

The proof-of-concept demonstration was aided by being staffed with subject matter experts with an unusually broad range of experience. Insightful is a successful consulting firm has applied data mining methods is a variety of industries and is very familiar with specialized software. JetBlue assigned to the team a line pilot with extensive academic and practical experience in mathematical analysis. SAGEM avionics contributed considerable time and resources to data extraction and validation issues. In all, the project staff was highly qualified and a continuation of the collaboration would require a significant budget.

Despite the specialized expertise however, the team spent a great deal of time and effort on framing the research questions and placing the analysis in context of the flight operations. The Insightful team was not familiar with relevance of many individual parameters nor the many complex procedures required in different phases of flight. The FOQA data is used for a variety of purposes:

- Monitoring adherence to aircraft manufacturer limitations.
- Monitoring compliance to FAA or company procedures
- Observing which of a variety of acceptable procedures are used
- Recording parameter statistics where target values are not specified

As a result the project became very interactive and it required a continued effort to keep individual issues in context and to communicate effectively between areas of expertise.

Having a consulting firm with numerous airline clients, having data mining tools embedded in FOQA vender software packages, or having training and software available to FOQA managers would reduce the implementation costs of similar projects.

Considerable effort was also spent on data export and validation. Both the event records and full flight data databases are very large it was extremely time consuming to export and reformat the information. A collaboration with a different but similar energy profile study was hampered by the other group's inability to export some data and results. It was apparent that maintaining an export capability is a critical component of FOQA data management systems.

5.2 Assessment of the Tool's Value to Flight Safety Analysis

JetBlue FOQA analysis currently uses most of the basic analysis tools that are common outputs of data mining algorithms. However the process of defining snapshot parameters or event sets, running queries, and formatting charts or graphs is time consuming. The data mining methods showed promise in improving efficiency by automating portions of the query and output process.

The Insightful algorithms proved quite powerful on data validation issues and identified several examples of large scale recording errors in the JetBlue database. Partly this was due to a focus on parameters that have not been studied fully before. Partly this was the result of correlation studies that produced clusters of atypical parameter values later found to come from invalid data. In any case data validation is a critical component of analysis and the data mining algorithms performed well

The advanced data mining methods go beyond current capability and represent new ways to organize the information.

- The ability to process full flight data independent of snapshot parameters and event sets was quite valuable. For example with current methods it is difficult to record whether an approach event occurs before or after a go-around. Using full flight data the process is straight forward.
- The Hexagonal Binning (fig 13) and Trellis graphs (fig 22) provide very compact representations of event correspondences and seem accessible to non-technical managers.
- The Principle Component Analysis and Trellis Scatter plots (fig 23, 29, 30) may provide insights to accomplished analysts, but seem more complex than most managers would accept.
- Some applications of the Tree Model (fig 31) charts were quite informative and accessible while others were less so. The usefulness to Tree Models appears dependent on an analyst's expertise in setting up the initial branching.
- Neural Networks did not produce any useable analysis.

- The energy profile study produced very intriguing results. The analysis was consistent with current event based approach definitions, however provided much greater depth. The ability to quantify "how unstable" a particular approach is both in terms of an energy index and in terms of percentiles within a distribution is new and informative.

5.3 Suggestions for Improvement

The proof-of-concept clearly demonstrated that there is value in analysis partnerships between airlines, FOQA venders, and specialist consulting firms. A clear direction to improve flight safety programs is to support and expand similar collaboration efforts. As data mining results begin to be reflected in manager decisions, the more useful techniques will become apparent.

The difficulty of course, is in providing a cost-benefit study that would justify the expense. The costs could be reduced by:

- Emergence of a consulting firm with specialized data mining in aviation expertise. If numerous client airlines collaborated with a small group of analysts the ramp up time could be considerably lessened.
- Incorporation of data mining algorithms and report formats in FOQA vender software packages. This would allow existing industry conferences and communications to serve as information sharing forums.
- Improved data export capabilities. This would allow an analyst to select software and external support that matches their company's particular needs.

6 Summary/Conclusions

The purpose of this proof-of-concept project was to evaluate practical usefulness of data mining tools applied to the analysis of airline digital flight data, and to develop methodologies for the analysis of de-identified data in the AGS database of JetBlue Airlines.

Event history data, parameter snapshot data and full-flight parameter data were analyzed with the help of Insightful Miner in combination with S-PLUS. Event history profiles and flight parameter correlation structures were utilized in the discovery process to construct new variables and subsets which helped identify flight safety issues. A variety of machine learning and visualization algorithms were utilized during this process.

The results proved to be useful to the JetBlue Flight Operations Quality Assurance (FOQA) personnel for identifying potential safety issues across different multiple attributes, such as approach class, event history, aircraft power settings, altitude and velocity. The ability to create specialized variables and graphics and to visually examine relationship from multiple viewpoints provided deep insight into data relationships. Cluster and principal components analyses coupled with large data modeling algorithms augmented the process by visually showing strong associations between attributes of interest.

The project demonstrated that significant value could be generated through:

- Careful data validation to uncover anomalies in data recording
- Discovery of general flight parameter relationships that lead to improved safety operations
- Discovery of atypical flight patterns that may present safety issues
- Efficient use of analyst's time
- Automation of repetitive processes
- Consistent and comprehensive utilization of *all* flight safety data (e.g., event histories, parameter snapshots and full-flight parameter data).

Additionally, it is likely, based on the proof-of-concept, that similar data mining analysis techniques can be applied to information derived from other data sources in an aviation organization, such as airport safety and aircraft maintenance, to deliver a more comprehensive picture of overall safety issues.