12-1-2014

Predicting Irregular Flight Operations Using a Binary Machine Learning Approach Based on National Meteorological Data

Martin Hellwig

PREDICTING IRREGULAR FLIGHT OPERATIONS USING A BINARY MACHINE LEARNING APPROACH BASED ON NATIONAL METEOROLOGICAL DATA

by

Martin D. Hellwig
Diplomkaufmann, Lüneburg University, 2004
Master of Business Administration, University of North Dakota, 2010

A Thesis

submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfillment of the requirements

for the degree of

Master of Science

Grand Forks, North Dakota
December
2014

This thesis, submitted by Martin D. Hellwig in partial fulfillment of the requirements for the Degree of Master of Science from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

_____
Dr. Travis Desell

_____
James Higgins

_____
Michael Poellot

_____
Dr. Kimberly Kenville

This thesis is being submitted by the appointed advisory committee as having met all of the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.

_____
Wayne Swisher
Dean of the School of Graduate Studies

_____
Date

PERMISSION

Title	Predicting Irregular Flight Operations Using a Binary Machine Learning Approach Based on National Meteorological Data
Department	Aviation
Degree	Master of Science

In presenting this thesis in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my thesis work or, in his absence, by the Chairperson of the department or the dean of the School of Graduate Studies. It is understood that any copying or publication or other use of this thesis or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my thesis.

Martin D. Hellwig
November 3rd 2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to the members of my advisory committee as well as Dr. Wei Ding from the University of Massachusetts in Boston for their invaluable guidance and support with this thesis. I further wish to particularly thank Dr. Kimberly Kenville for her continued support throughout the entire program.

ABSTRACT

Flight delays are caused by a multitude of external influences as well as revenue driven carrier decisions. Some factors are obvious while others remain inaccessible to the traveling public. Yet knowing of potential flight delays or cancellations in advance can significantly improve passengers' travel experience and empower them to make informed decisions when flight irregularities occur.

We combine a Naïve Bayes - based feature selection method with publicly available meteorological data and flight performance statistics to create a forecasting tool that provides passengers with an improved prediction of potential delays. After promising initial results we optimize our feature selection and weighting, yielding a 66% true positive rate paired with a 66.5% accuracy. This means that 66.5% of our forecasts are correct while the model manages to properly detect 66% of irregular flights. Compared to a probabilistic forecast based on historical data, this represents an improvement of 332% and 436% respectively.

CHAPTER I

INTRODUCTION

Statement of the Problem

Air Transportation Networks have long been known to be rather complex structures (Currie, Dickey, Duclos, & Price, 1974). This complexity has grown with the increasing traffic demands on the national airspace system (Meyer, Saghi, & Tarnai, 2008). In such a tightly knit network, both local and systematic delays can be caused by a large number of heterogeneous events throughout the entire system (Ball M. , et al., 2010). Reasons may include such obvious obstructions as adverse weather (Robinson, 1989) or malfunctioning aircraft, but may also include delayed crews (Rubin, 1977), traffic congestion (Nogami, 1995) and a large number of other factors. Any one of these factors may delay a single flight or a certain set of flights at any given station throughout the system. Such delays may then propagate throughout the system and may be amplified in the process (Li & Ding, 2008).

These irregular operations impose significant cost on passengers and the environment (Dray, Evans, Vera-Morales, Reynolds, & Schafer, 2008). Passengers are disproportionally affected in cases of cancellations irrespective of whether those disruptions were the direct result of operational requirements or whether the operator chose to cancel the flight based on economic considerations (Bruce, 2011).

Purpose of this Study

This study will analyze whether a binary machine learning approach may be used to accurately forecast flight delays and cancellations in spite of significant uncertainties. While such attempts could be made using a wide variety of algorithms and methods, we specifically test a Naive Bayes approach to feature reduction paired with a binary classification system of meteorological features derived from weather observations reported by weather stations throughout the 48 contiguous United States.

Throughout the study several aspects of the original approach may be adapted to optimize the final results that will be tested against a probabilistic forecast based on historic flight performance data. Ultimately we aim to answer the question whether the machine learning approach can consistently outperform the probabilistic base line.

This study does not aim to generate a forecasting method that can be immediately implemented in industry. Instead it is intended to lay the groundwork for future research that may ultimately lead to such an implementation. In order to be a viable as such, the machine will have to achieve an accuracy and a true positive rate of at least 60% each.

Literature Review

Causes of delay have been studied extensively, both from an operational (Ball M. , Barnhart, Nemhauser, & Odoni, 2006) and from a theoretical view (Sun, Clinet, & Bayen, 2011). Studies examine the entire flight period beginning with the preparation of the aircraft at the gate (Hebert & Dietz, 1997) and the delay-prone subsequent queuing process for departure from congested airfields such as Boston (Idris, 2002), Dallas and Atlanta (Mayer & Sprong, 2008).

To combat enroute delays, many studies propose numerous linear integer and mixed integer models (Dell'Olmo & Lulli, 2003). Others present complex forecast models aimed at helping decision makers in their efforts to optimize traffic flow (Lacher & Ball, 2002). However, researchers also note that enroute delays are largely influenced by factors outside traditional models (Sood, Mulgund, Wanke, & Greenbaum, 2007) including unexpected adverse externalia such as regional (Matus, Hudnall, Murray, & Krueger, 2010) and distant volcanic activity (Dacre, Grant, & Johnson, 2013) and even the psychological patterns among air traffic controllers (Gronlund, Dougherty, Durso, Canning, & Mills, 2005).

Separation requirements during approach - especially in congested airspace - can also exacerbate existing delays (Wang & Tsao, 2012) and even spread to previously unaffected aircraft en route (Slattery & Cheng, 1997). This especially holds true in reduced visibility situations (Pisano, 2008)

While delays are usually the direct or indirect result of an external influence, cancellations are more often the result of a conscious decision made by the operator (Seelhorst & Hansen, 2014). Such decisions may themselves be the result of insurmountable delays or obstructions to regular operations. They can, however, also be strategic decisions attempting to optimize overall operations (Shavell, 2001) or to improve purely economic factors such as yield or load factors (Wang & Regan, 2006).

Most externalities can – at least to a certain extent – be monitored, analyzed and used to predict flight delay generation and propagation as well as related cancellations. The strategic decisions made by operators are less transparent and not uncommonly considered proprietary tools of competitiveness (Seelhorst & Hansen, 2014). An air

carrier that makes the best strategic decisions can be expected to generate the highest overall yield which underlines the importance of good flight management including cancellations where opportune. However, it is quite difficult for the traveling public or even us as researchers to understand the opaque models on which carriers base their cancellation decisions. They must therefore be treated as unknown properties in the larger scheme of factors influencing the regularity of flight operations and cannot be used in any forecasting model or flight delay mediation approach.

While mere observation of past flight data allows interested parties to create simple statistics of any flight's on time performance that in turn may be used to generate a forecast of delays and cancellations, such a forecast will be inherently unreliable as it is purely based on past probabilities and does not take any current and future externalities into account (Lorentson, 2011). To create a more appropriate forecast, one needs to determine and define a sufficiently influential set of such external factors that may negatively affect flight operations.

Meteorological Externalities

Many of the most important externalities in aviation are meteorological phenomena. Aircraft operations are not only negatively affected by reduced visibilities arising from precipitation or fog (Black, 2010), but also by other factors including particularly high or low ambient temperatures (Federal Aviation Administration, 2014). Adverse weather conditions can affect departure and arrival operations (Mueller, 2002) and also prompt airlines to reroute aircraft in flight to avoid operational hazards (Zobell, Ball, & Sherry, 2001). We therefore propose that high resolution meteorological data

provides the best stand-alone basis for forecasting irregular flight operations irrespective of any other external variables.

The National Oceanic and Aerospace Administration (NOAA) maintains a database of such high resolution data with a wide breadth of weather measurements and observations from a large number of stations throughout the United States and its territories. Figure 1 provides a visual representation of the geographic distribution of included weather stations. In areas with greater station density, different intensities were used to improve the visual appearance. These varying intensities are not indicative of any variance in station quality, importance, availability or any other feature relevant to its validity for our further analysis.

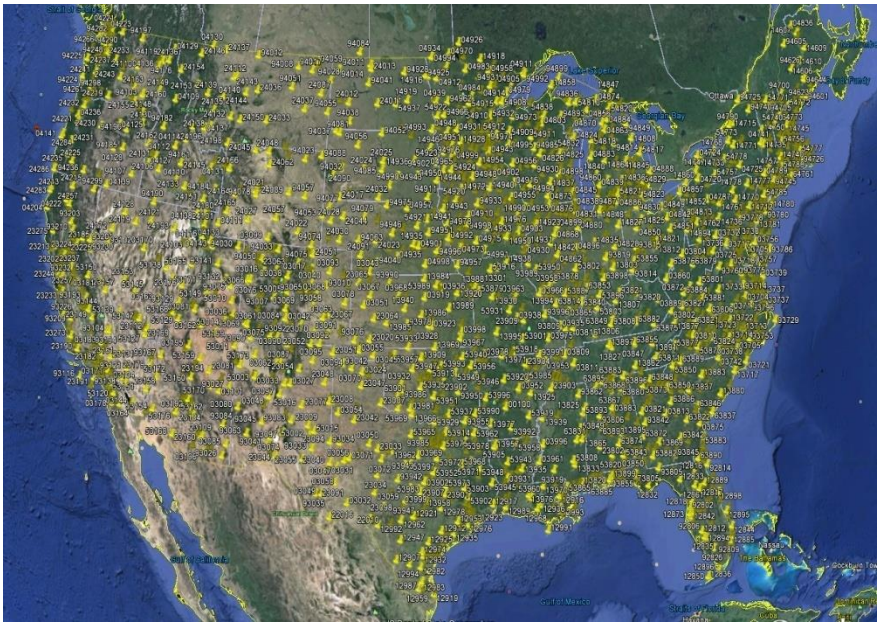

Figure 1: Weather Stations in the Contiguous United States

The main purpose of this study will be to determine whether delays and cancellations can be forecast using a binary machine learning approach based purely on current, recent and historic weather observations reported by stations throughout the continental United States. As meteorological systems vary greatly between the

continental United States and geographically separate areas including Alaska (Partain Jr, 2008) and Hawai'i (Sanderson, 1993), we will limit our scope to these contiguous 48 states.

It is apparent that weather at the departure airport, the destination airport or along the route will directly influence the flight's performance. However, as stated before, air transport networks tend to be rather complex and delays often propagate in a less obvious fashion that can only be forecast to a certain extent (Ding & Li, 2011). These propagations will vary between different geographic regions and across airline networks as different operators employ vastly different network structures (Seelhorst & Hansen, 2014). To create a universal forecast model, we need to analyze which weather conditions at which stations throughout the contiguous United States may increase the probabilities of a delay or cancellation for the flight under consideration.

We will accomplish this by training the machine for each new flight using historic data. Based upon this data, our model will decide which stations to consider for a given flight and what weights to assign to the individual measurements and observations. Once training has been completed for a given flight route, future forecasts for similar flights on other days can be based on this training and on updated current weather data.

CHAPTER II

METHODOLOGY

Data

The Bureau of Transportation Statistics (BTS) provides detailed departure statistics for most commercial domestic passenger flights. Data are publicly available and include specific information regarding various different types of delay incurred as well as the precise delay in minutes. While aggregate information is available, we choose to obtain and use raw data that allows us to build a consistent aggregate dataset that best suits our model.

One major drawback of this dataset is the BTS's reporting method. Data are provided specific to route, operating carrier, month and year. As a result, we need to obtain and consolidate at least twelve different reports for each year multiplied with the number of included operating carriers. This will provide us with the annual data for one specific route. Unfortunately each individual subset must so far be obtained manually in a rather time consuming process that is exacerbated by the BTS's transmission speeds. The typical time required to obtain just one subset averages about seven minutes. Interestingly, the HTML version of the data gets reported significantly faster which suggests that the issue lies not within the BTS database itself but in the delivery of structured CSV files. However, HTML interfaces change frequently and any automated retrieval system will need to rely on predictable, structured data.

Our request for faster access was unfortunately not granted by the BTS. Therefore the scope of this study will be limited to a small number of examples. We are, however, confident that these examples are representative of the entire population and that faster access could be negotiated with the BTS if a more comprehensive implementation of the suggested method should be desired.

While access to the BTS's flight data is somewhat limited, NOAA provides an excellent and accessible repository of detailed meteorological data in various stages of aggregation. Yet again we chose to create our own aggregates based on raw source data files which exceed a size of 6.5GB for each period of twelve months.

The model must be trained using relatively large long-term datasets for both weather and delays or cancellations. Processing this raw data will take considerable amounts of computing time, but weather data only needs to be preprocessed once as the results can be used to train any future flight routes. The smaller set of weather data used for the actual forecast generation will compute rather quickly and it is unlikely that any significant performance increase could be achieved by using aggregate data. This study does not use any proprietary datasets but relies solely on publicly available data sources.

In order to facilitate the forecasting routine and to avoid unnecessary noise, we will create binary variables for both the independent and the dependent realms. While the dependent variables describing flight behavior are easily defined, a significantly more complex definition procedure is necessary for the independent features describing the reported meteorological environment. We will describe this process in detail in the upcoming paragraphs.

Temporal Scope

We begin by training the machine on annualized data and subsequently proceed to refine the training procedure using monthly data. The latter option is much more expensive in terms of computing resources but could potentially improve the forecast quality as weather phenomena are inherently seasonal in nature and traffic patterns also change throughout the year to accommodate changing demand. These changes in network structure change delay propagation characteristics while the reduced slack during peak and shoulder seasons reduces the network's robustness and its ability to recover from irregularities.

Feature Selection

Flights

Our model aims to forecast both cancellations and delays. While cancellation is – by nature – a binary feature, our source data will provide the exact duration of the delay incurred in minutes as an integer value. We are, however, less interested in the actual delay duration and instead aim to forecast the probability of a severe delay. While various different delay measures exist throughout the industry, we will consider any delay in excess of ninety minutes to be severe, irrespective of flight route or scheduled duration.

Our dependent feature can hence be defined as follows:

$F_i = 0$ Flight departed earlier than ninety minutes after the scheduled departure time

$F_i = 1$ otherwise

Departure Time Clusters

Every iteration of our forecasting algorithm will aim to make a forecast for a particular flight route operated by a specific carrier on a given day. However, delays may propagate very differently during different times of the day.

Most morning flights, for example, are less prone to delays caused by inbound aircraft operating flights on the same day. They could, however, be affected by delays that built up on the previous day and ultimately resulted in an evening cancellation causing an aircraft shortage in the morning. Similarly, crews arriving late on the night before may not be available in the morning due to duty time limitations (Missoni, Nikolic, & Missoni, 2009).

At some airports, local weather also tends to vary greatly throughout the day. Issues such as morning fog can drastically impact operations at these stations. Local traffic patterns and rush-hour peaks are further influences on delay creation and propagation characteristics that may be very different between mornings and evenings.

One possible approach to this problem would be a finer resolution of the forecast. In this scenario we would only analyze one particular flight and equally also only create a forecast for that specific flight. This approach would work well for consistent schedules of carriers that operate the same flight at the same time every day. However, some carriers - regional airlines in particular - very often change flight times and flight numbers on any given route. This lack of consistency does not allow for unique identification of any particular flight for a sufficiently long time frame.

To refine our forecast to different parts of the day without relying on any particular flight identifiers, we will group the data into six clusters based on departure time:

- Flights departing between midnight and 7:59am
- Flights departing between 8am and 10:59am
- Flights departing between 11am and 1:59am
- Flights departing between 2pm and 4:59am
- Flights departing between 5pm and 7:59pm
- Flights departing between 8pm and 11:59pm

Different carriers

An additional problem arises when different carriers operate the same route in the same time slot. These carriers are likely to have quite different network structures that are vulnerable to different externalia and have significantly different delay propagation characteristics. Data from different carriers cannot safely be mixed without introducing unwanted and potentially detrimental noise. Hence our model must be refined based on the operating carrier. The BTS does report flight performance data specific to each carrier, so we need not include this in our preprocessing stage but simply take it into account at the data selection level.

Meteorological Data

Binary Enumeration

The operational performance has already been classified in a binary fashion. Now we need to establish a number of equally binary meteorological features to use as possible predictors of irregular operations. While it is easy to create binary descriptors for regular vs. irregular operations, the reported weather data is quite heteromorphic.

Table 1 provides an overview of the variables and data types included in the source data supplied by NOAA (National Oceanic and Atmospheric Administration, 2014). Some of the reported variables are suitable feature candidates that can easily be converted into a binary feature by thresholding. Others may require significant preprocessing in order to create applicable binary feature expressions.

Table 1: Data Reported in NOAA Files

Feature	Data Type
Sky Condition	Concatenated Strings
Visibility	Float
Weather / Precipitation Type	Concatenated Strings
Precipitation	Float
Temperature	Float
Dew Point	Float
Humidity	Integer
Wind Speed	Integer
Wind Direction	Integer
Pressure	Float
Altimeter	Float

High Temperatures

It is well known and documented that aircraft performance is adversely affected by the higher density altitudes induced by high temperatures (Federal Aviation Administration, 2008). Aircraft operators may choose to adjust payload, change equipment or even cancel flights in response to the changed operational requirements

presented by increased ambient temperatures. For our study we initially establish a relatively low threshold of 31 Celsius (approx. 88 Fahrenheit). Any temperature above that threshold will be considered to be a potential contributor to delays or cancellations.

Low Temperatures

Low temperatures by themselves are not known to negatively affect aircraft performance. They can, however, negatively affect flight operations in an indirect fashion. For example, temperature is a major factor in the decision whether to subject an aircraft to deicing procedures (Lindholm, Hage, Wade, & Rasmussen, 1997) which can cause significant delays. We will therefore consider temperatures below 5 Celsius (approximately 41 Fahrenheit) to be a contributing factor to irregular operations.

Sky Conditions

Sky Conditions are supplied as observational values describing each cloud layer present, if any. For each layer, the observation will specify the altitude and the coverage classified as few (FEW), scattered (SCT), broken (BKN) or overcast (OVC). Clear skies are reported as CLR while an empty value represents a missing observation.

The standards used for cloud reporting in our dataset (National Oceanic and Atmospheric Administration, 2014) are identical to the ones defined by the International Civil Aviation Organization (ICAO) for cloud reporting in the Aviation industry (Federal Aviation Administration, 2014). We will analogously apply these standards to enumerate the individual designations following the brackets shown in Table 2 while calculating the overall impact factor.

Table 2: Cloud Cover Enumeration

Condition	Impact Factor
FEW	1/8
SCT	3/8
BKN	6/8
OVC	8/8

Coverage is reported for each individual cloud layer. In aviation, lower layers have a much greater impact on airport operations than higher layers. The latter may be indicative of certain atmospheric disturbances that could prompt a carrier or air traffic control to preemptively reroute traffic. Individual flight crews may also request a different routing to avoid potentially negative impacts on passenger comfort or flight safety. However, the resulting delays will be less severe than those incurred during the departure and arrival phases. During these phases, increased separation requirements may produce significant delays.

While the different layers have very different impacts on flight operations, we need not specifically account for these differences. Instead, these different layers may be treated as cumulative and allow our algorithm to automatically learn which locations are generally more prone to delays caused by sky coverage.

Clear conditions may obviously be ignored and missing values must be equally discarded. As a result, one may calculate the total impact value of the n cloud layers as follows

$$Fc = \sum_{k=0}^{n} 0.1\ FEW + \ 0.3\ SCT_k + 0.6BKN_k + OVC_k$$

Cloud coverage in excess of 70% will be considered significant. This is represented by a threshold of Fc>= 0.7*n.

Visibility

Reduced visibility affects both departing and arriving traffic. The latter is disproportionately affected as landing in reduced visibility situations is considerably more challenging in spite of the variety of available support systems such as instrument landing systems (ILS) and increasingly the global positioning system (GPS). While departing aircraft are less affected by reduced visibility itself, they are subject to indirect delays. As separation requirements for arriving aircraft increase, runway capacity decreases and especially busier airports will incur delays as they must prioritize landing traffic to avoid technical diversions as a result of declining fuel levels on aircraft in holding patterns. As a result of this prioritization, departing flights are likely to incur significant visibility related delays as well. Our initial threshold for the visibility feature will be set at 3 miles of reported ground visibility.

Weather

Just like cloud conditions, weather is reported as a concatenated string that need to be interpreted in the preprocessing stage using the definitions provided by the NOAA (National Oceanic and Atmospheric Administration, 2014)

.Weather types fall into two major categories:

- Precipitation
- Obscuration

Precipitation includes rain (RA), drizzle (DZ), showers (SH), snow (SN), hail (GR), ice crystals (IC) and unknown precipitation (UP) while obscurations represent reduced visibilities due to fog (FG), haze (HZ), mist (BR), dust (DU) and sand (SA). We will aggregate snow, hail and ice crystals into one category representing frozen precipitation while rain, drizzle and showers will be maintained as a separate category of liquid precipitation.

"Unknown Precipitation" is a category used by automated stations and could describe a range of precipitation types – mostly rain and snow. As it is impossible to ascertain whether this precipitation is frozen or not, it must be categorized based on the likely effect on our forecast. Considering all "unknown" precipitation as rain, we might reduce the overall true positive rate (TPR) as a result of undervaluing the occurrence of frozen precipitation. Conversely, the true negative rate (TNR) would be negatively impacted by overvaluing non-frozen precipitation when adding UP to the frozen category. In order to maximize TPR, the second alternative is clearly preferable. Therefore, "unknown" precipitation shall be considered alongside frozen types.

Within the frozen precipitation group, the different types of precipitation will be weighted according to Table 3. These weights are based on the assumed effect on aviation operations. Snow is most likely to negatively impact airport operations as clearing an airfield from snow cover takes considerable time. Snow is also the only type that can be considered both as precipitation and as obscuration. Snow – especially blowing snow – can drastically reduce visibility and hence adversely affect operations. However, as our model considers visibility as an independent feature, we need not separately account for snow-induced visibility reductions.

However, blowing and drifting snow cause a variety of challenges on airports that may cause delays similar to those caused by falling snow. Both types of snow require increased airfield maintenance and aircraft deicing procedures. We will hence consider any mention of blowing or drifting snow on par with snowfall.

Table 3: Frozen Precipitation Weights

Precipitation Type(s)	Weight
Snow	4/8
Ice Crystals, Hail	3/8
Unknown Precipitation	1/8

Ice crystals and hail occur during different times of the year. Hail is most common during the summer months as thunderstorm activity is much greater during that season. However, ice crystals and hail may be classified as a separate subgroup within the frozen precipitation category. Our model will consider temperature and thunderstorm activity separately which eliminates the need for a particular distinction within this group. Finally, unknown precipitation may or may not be frozen and will hence be assigned a relatively low weight.

We are less interested in the particular cause of reduced visibility and will aggregate all types of obstructions into one major category.

Thunderstorms

Thunderstorms are reported alongside the aforementioned types of precipitation. However, especially in aviation, thunderstorms must be considered a particular threat to safety and hence regular operations. For example, the average delays caused by just one

thunderstorm at Frankfurt International amounted to 740 aircraft minutes (Hauf & Sasse, 2002).

As thunderstorms may have such a substantial impact on airport operations, we will consider any report of thunderstorms whatsoever to be critical enough to consider the "thunderstorm" feature to be true.

Wind

Steady winds do not necessarily negatively affect operations as long as wind direction does not significantly differ from runway orientation. However, many airports only have one or a small number of parallel runways resulting in potential crosswinds depending on wind direction. Strong crosswinds can mandate larger separations (van Es, van der Geest, & Nieuwpoort, 1999) and hence cause flight delays – particularly when wind speeds fluctuate. The amplitude of wind gusts heavily depends on the airfield's surrounding geography (Agustsson & Olafsson, 2009) but will generally be larger when average wind speeds increase. For our study we will therefore consider sustained wind speeds in excess of 20 kts to be potential causes for delays. Such a low threshold will aid in avoiding false negatives for this feature.

Wind Direction

Wind directions are reported in degrees indicating the direction from which the wind is approaching. This results in decimal readings between 1 and 360 degrees which do not easily convert into a binary structure. We use two separate binary feature values to describe wind direction – one that indicates whether air masses are moving southward

(coming from a northerly direction) and another indicating whether they move eastward (coming from a westerly direction).

Table 4: Wind Quadrants

Wind Direction	Northerly?	Westerly?	Binary Representation	Quadrant
360	Yes	No	10	2
45	Yes	No	10	2
90	No	No	00	0
135	No	No	00	0
180	No	Yes	01	1
225	No	Yes	01	1
270	Yes	Yes	11	3
315	Yes	Yes	11	3

As can be seen in table 4, these two features allow us to specify wind direction in four different quadrants: 0-89 degrees, 90 to 179 degrees, 180 to 269 degrees and finally 270 through 359 degrees. Both 0 degrees and 360 degrees will be considered valid definitions of true north. Using the binary representation, wind direction may be categorized into one of the four quadrants enumerated from zero to three.

Most airports tend to have strong predominant wind directions and hence relatively little variability in these two values should be expected. However, since even small changes in wind directions can significantly affect local climate (DeGaetano & French, 1991) and of course determine their further trajectory, wind direction is expected to have at least a certain influence on our model.

Pressure

Low pressure systems are generally less stable (Billet & Titlow, 2010) and cause more adverse weather events. Therefore, lower pressure is expected to be positively correlated with irregular operations. 1013mbar indicate normal average atmospheric pressure at sea level while in larger hurricanes pressures as low as 892 mbar have been observed on landfall (McCallum & Heming, 2006). As with temperatures, we also choose a highly sensitive value for barometric pressure. Any value below 99.5% of the regular atmospheric pressure shall be classified as representing a low pressure system. This results in a threshold of 1008mbar.

In total, we consider the following eleven binary features:

- Cloud Cover
- Reduced Visibility
- Larger Amounts of Rain
- Larger Amounts of Frozen Precipitation
- Strong Winds
- Thunderstorms Present
- High Temperatures
- Low Temperatures
- Northerly Wind Direction
- Westerly Wind Direction
- Low Pressure

Aggregation

While data resolution varies between stations, most stations provide a large number of reports on any given day. Processing our further analysis using this high resolution would be rather unfeasible on any affordable computing architecture. However, we can assume that aggregating the data to daily reports from each station will still provide us with a valid basis for the following calculations. Removing the varying resolutions may also reduce the risk of distortions caused by noise in the high resolution data (Nettleton, Orriols-Puig, & Fornells, 2010).

Table 5: Occurrence Thresholds

Feature	Occurrence Threshold
Reduced Visibility	15%
Liquid Precipitation	70%
Frozen Precipitation	20%
Strong Winds	30%
Thunderstorms	0% (any thunderstorm throughout the day)
High Temperature	20%
Low Temperature	20%
Northerly Wind	50%
Westerly Wind	50%
Low Pressure	50%

This aggregation can be accomplished by counting the individual occurrences of each feature throughout the entire day at a given station. For each feature, we define the occurrence thresholds shown in Table 5 and consider the feature to be true for the entire day if the cumulative individual feature observations exceed that threshold. The various different threshold levels reflect the individual feature's potential impact as well as the likely occurrence throughout the day. This is of importance as certain features are less likely to occur at certain times. High temperatures, for example, occur more often during the daytime than during nighttime.

Missing Values

Not all stations report data in constant, predictable intervals. A given station might report several times an hour and subsequently fall silent for several hours. This could be due to a system malfunction or transmission errors, but may also be the result of intentional system design. A station might be configured to only report data that the individual operator would consider interesting. This is particularly true for remote stations that report their observations automatically using expensive data links. Such custom reporting intervals may also cause the same station to be present in the data during certain seasons while it may be absent during the rest of the year.

Even if a station does report its observations several times each day throughout the year, not all values are necessarily present. Once again this could be a result of system design as a station might be configured to only report certain measurements at certain times. However, it is more likely to be the result of a malfunction as there is little incentive to a station operator to suppress individual observations from a report as the

added data volume is typically minimal. Missing values may be caused by transmission errors, system malfunctions or – most likely – by malfunctioning sensors. Sensors can be damaged by adverse weather conditions, wildlife and even vandalism.

These inconsistencies pose a serious challenge for further data analysis. We will address this challenge depending on the particular circumstances:

- During the training phase:
    - If a station does not report any data for a given day, it will be excluded from our analysis for that time frame.
    - If a station does report throughout a given day but does not report a certain value, that value will be ignored while the remaining report will be processed normally.
- During the testing and forecasting phase we are only concerned with those stations that were deemed sufficiently influential during training and that have been included in the forecasting model. In this phase missing values need to be handled differently:
    - If a station does not report any data for a given day but does report for both the day before and after, we will interpolate between these two days. This is accomplished by calculating the arithmetic mean of each value reported on the adjacent days and substituting the result for the missing value.
    - The same method will be applied to individual missing values.
    - If a station does not report data for any of the adjacent days, it must be ignored as any more complex interpolation could negatively affect the

model's validity (Li, Heap, Potter, & Daniell, 2011). In this case, proportionally higher weights will be assigned to the remaining stations in the forecast model.

Probability Calculation

As a result of the previously described preprocessing, our main analysis will use eleven binary features based on the data reported by a varying number of approximately eight hundred weather stations. The model will be trained on at least twelve months of consecutive data which results in a total number of 3.2 million features. Including each feature for the day of each flight as well as the preceding three days, the number of features increases to roughly ten million features. This would be rather inefficient and we hence need to determine which features are the best predictors of irregular operations and reduce the number of considered features to a manageable level.

Naïve Bayes

As indicated in the previous paragraphs, we are processing ten million independent features and only several hundred or a few thousand expressions of the dependent variable. The number of dependent expressions varies with the number of flights operated by the selected carrier within the analyzed time slot on a given date. Calculating the probability of the dependent features for each independent feature initially appears to require significant computing resources. Considering the large number of independent features in contrast to relatively few dependent features also raises concerns regarding the result's validity. However, the same reasons also make it

relatively easy to calculate the probability of each independent feature given that any dependent feature is true.

The Naïve Bayes classification method allows us to process that calculation and then use the result to easily compute the probability with which each independent feature may cause a dependent one to be true. In very simplistic terms, the probability of a delay given that feature X is present is equal to the probability of feature X given a delayed flight multiplied with the overall probability of a flight delay and divided by the overall probability of feature X (Jiang, 2012). Using the standard notation P(dependent variable | independent variable), we can write:

$$P(delay|feature\ X) = \frac{\text{P(feature X | delay)}*\text{P(delay)}}{\text{P(feature X)}}$$

Once the large set of weather stations is sorted by their impact factor based on the previously described Naïve Bayes calculation, we can create a visual representation displaying the flight route and the most influential stations. Figure 2 presents an example of such a visual representation for a flight from Chicago's O Hare to Knoxville, TN. Depending on the further analysis, a varying number of these stations will be used in further calculations.

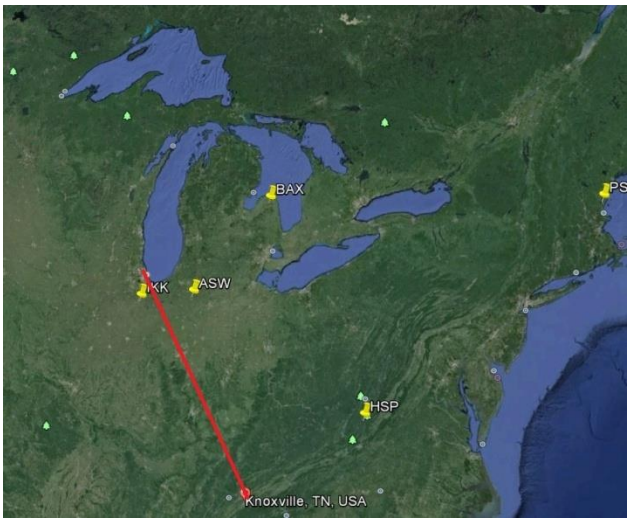

Figure 2: Stations Relevant to the Chicago-Knoxville Route

Scoring Function

After determining the particular weather stations to be included in our calculations, we initially create a forecast for a random day from the training set using the individual's relative importance determined during the Naïve Bayes calculation as their weights. This forecast will then be compared to the actual known outcome which will place the forecast into one of the following four categories:

- True Positive (TP): The machine expected an irregularity on a day that actually experienced delays or cancellations.
- True Negative (TN): The machine expected no irregularities on a day that did not incur any.
- False Positive (FP): The machine expected an irregularity on a day that did not incur any.
- False Negative (FN): The machine failed to forecast an irregularity on a day that actually experienced a delay or cancellation.

It is evident that true positives and true negatives are desired outcomes while the occurrence of false positives and false negatives should be minimized. Particularly the latter are of concern as we aim to specifically forecast irregularities.

In order to train the machine, each forecast must be evaluated in order to enforce tendencies that appear to improve results while avoiding such that seem to have an adverse effect on forecast quality. We have a strong preference for accurately predicting irregular operations over simply predicting normal operations and therefore score each forecast as follows:

True Positive: +4

True Negative: +2

False Positive: -3

False Negative: -9

Each training session will include multiple forecasts - one for each day within the training period. As a result, the overall score for the entire period is calculated using the following formula based on the counts of true positives, true negatives, false positives and false negatives that occurred throughout the training period:

$$S = 4*TP + 2*TN - 9*FN - 3*FP$$

This approach allows us to create a score for the initial feature weights and to subsequently adjust said weights in future iterations. Each such iteration will be scored and compared to the preceding one. If modifying the feature has improved the overall forecast score for the training period, we continue to incrementally amplify the modification that appears to improve the result. If at any time our modification of a

feature reduces the overall forecast score, we will reverse direction and reduce the step size by 50%, seeking the local maximum in that direction. Should no improvements be reached in either direction, the machine will use the weight computed using the Naïve Bayes calculation.

The same general logic allows us to decide the number of stations to use for the forecast. In some cases relatively few very influential stations may provide the best results while in other situations a larger set of relatively less influential stations will result in a better forecast score. The optimal number of stations may be different for each flight route, time period and operator as well as scheduled departure time. As figure 3 illustrates, the number of stations will also vary between different seasons. In this case the optimal number of five stations in February shrinks to just two stations in May.

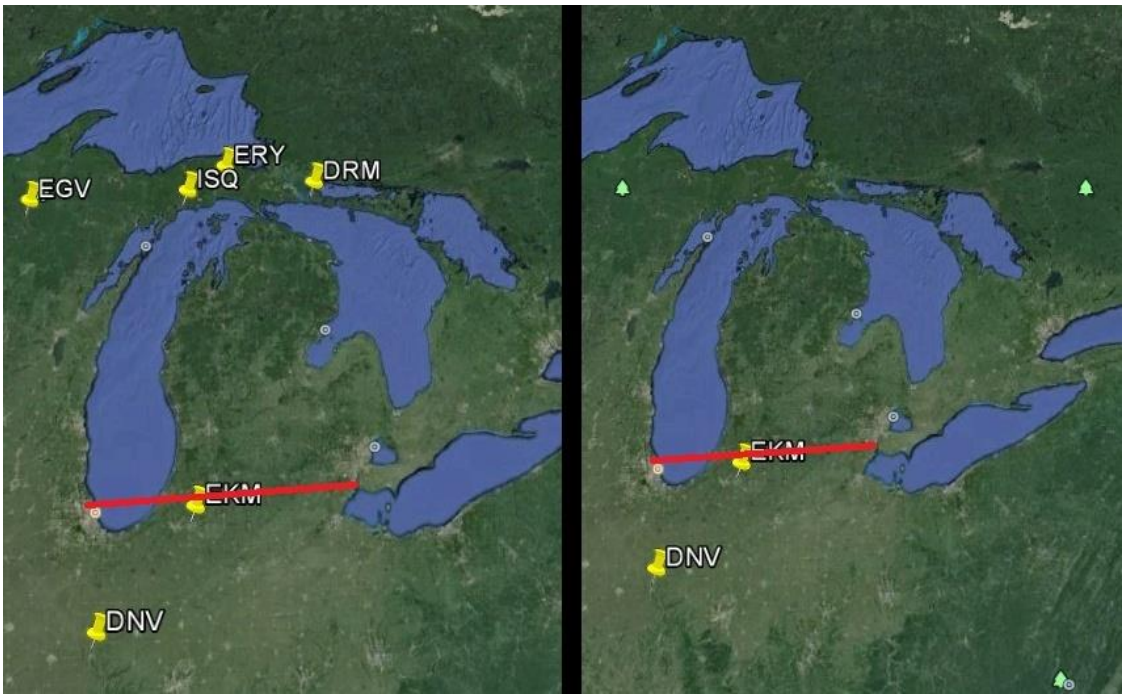

Figure 3: Stations Selected for Chicago-Detroit in February and May

Minimum Number of Stations

As illustrated in figure 3, our machine may on occasion select a very small number of stations as optimal. However, such a small number of independent features will make the forecast model substantially more vulnerable to distortions in the reported data. If data from one of these stations should be unavailable for any of the reasons described in an earlier paragraph, a forecast based on just two stations would become very unstable. We will therefore override the machine's decision and require a minimum number of four stations. At the same time, the number of stations allowed in the forecast model is capped at fifty. While the minimum requirement applied to several of our tests, the machine has never chosen a model with more than 27 stations.

CHAPTER III

RESULTS

Station Selection

Figure 4 shows the seventy most influential stations for a Fargo bound flight departing Chicago O'Hare after 8pm. Seventy stations would exceed the maximum cap on the forecast model and in fact the machine proposed to only use 13 of these 70 stations. We are showing 70 here to visualize the validity of our Naïve Bayes based selection method.

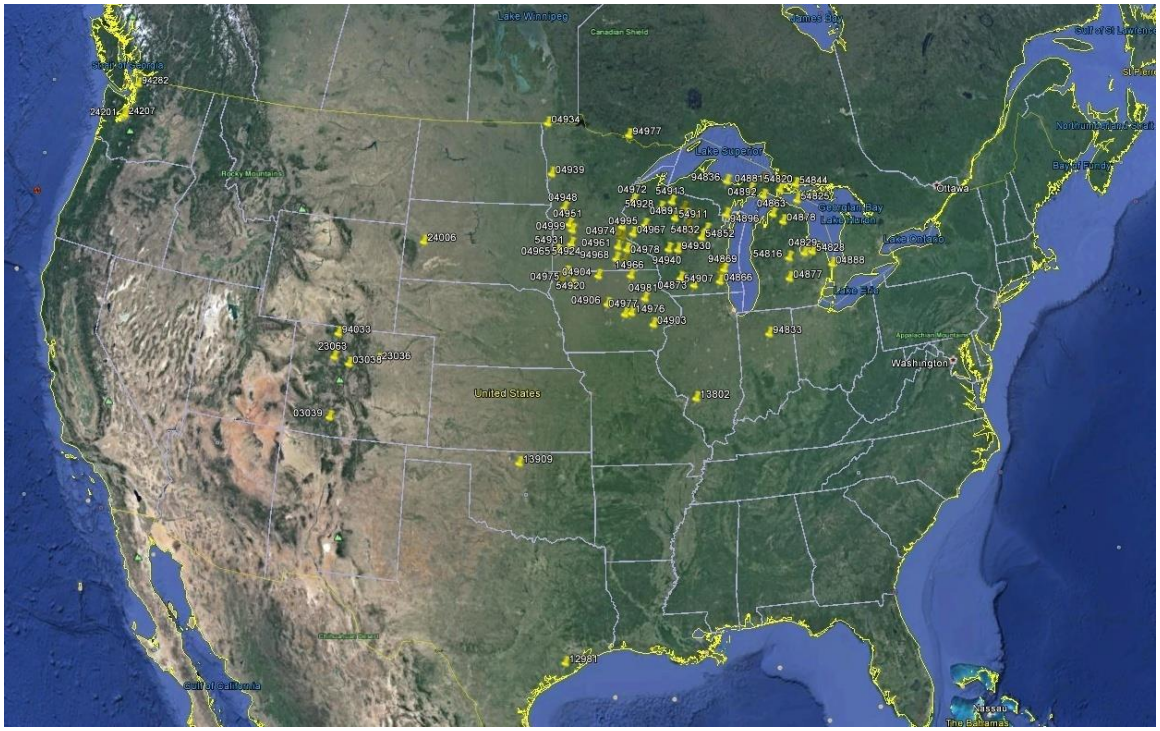

Figure 4: Stations Selected for Chicago-Fargo Route

The relative importance of stations along the flight route is not particularly surprising, but our model does assign relatively high importance to a collection of

weather stations in Colorado's Rocky Mountain area. This may at first seem unexpected and raise questions about the model's validity. However, the importance of those stations may be explained by taking the route network into account. The tested flight between Chicago and Fargo is operated by ExpressJet Airlines operating under the United Express brand. ExpressJet's United Express fleet serves several major United hubs including Chicago and Denver. In fact, ExpressJet serves Fargo from both of these hubs with the same equipment. It is quite apparent that delays occurring in one part of an airline's complex route network can easily propagate throughout it (Li & Ding, 2008).

While we create an impact factor value for each station, our model will only consider a certain number of them. For each flight route and time slot, the model attempts to achieve the optimal result by including enough stations to create a valid forecast while limiting noise that could be introduced by large numbers of stations. To limit required computing time, the optimization process is capped at a maximum of fifty stations. Beginning with a minimum of four stations, testing will proceed with increasing sizes of station sets as long as the overall accuracy score increases. Sets smaller than twenty stations are tested in every iteration, while larger sets up to fifty stations will be considered as long as the score continues to increase. For example, Figure 5 shows the analysis of a flight from Chicago O'Hare to Charleston, WV. In this case the model decided to use the top twenty stations for its further calculations. Again there are clearly visible clusters of stations in the vicinity of the flight route.

However, in addition to those clusters, the model includes a relatively large number of relevant stations along the Appalachian Mountain range in Pennsylvania, in the Chesapeake Bay region and in Michigan as well as one station in upstate New York.

As the Appalachian Mountains present a significant moisture barrier (Konrad, 1994), it is not surprising to see a number of relevant stations along this range extending from Virginia into Pennsylvania.

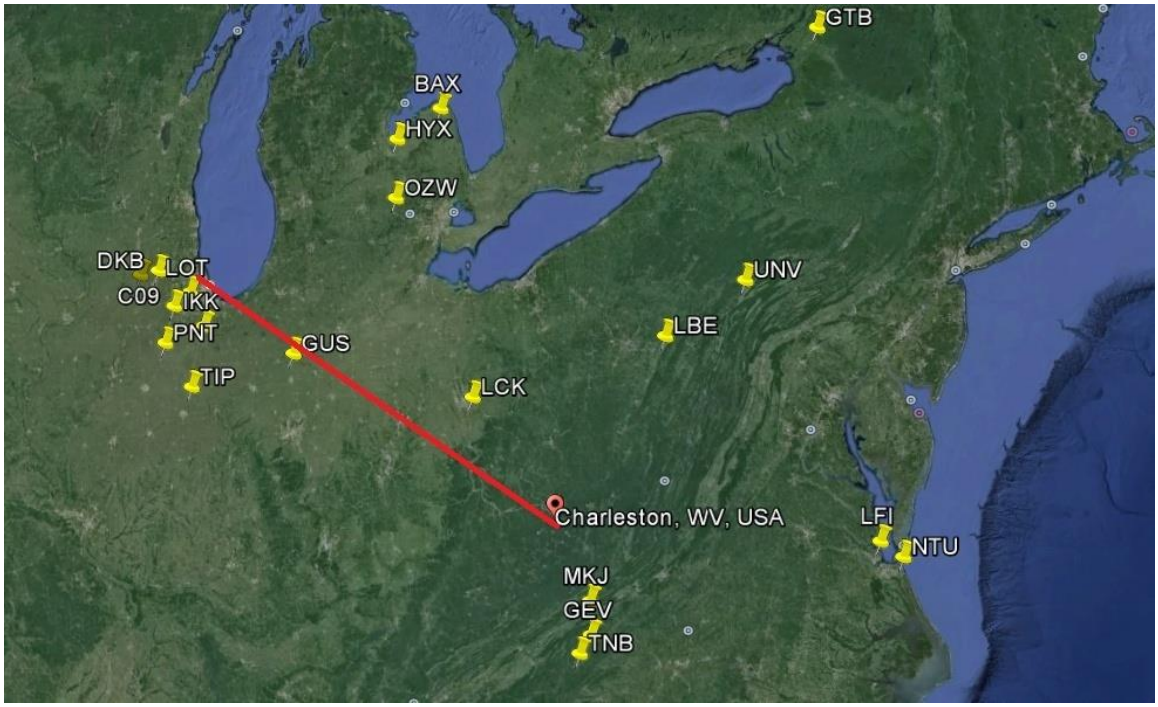

Figure 5: Stations Selected for Chicago-Charleston Route

We assume that the importance of the stations in western Michigan arises from the impacts the Great Lakes have on the regional climate (Scott & Huff, 1996). This effect may well be influencing the Chicago area and the Detroit area in a similar fashion and thereby causing the correlation observed in this example. It is possible that the station in New York was selected due to a similar correlation.

Finally, the reasons for the Chesapeake Bay's importance cannot be properly explained at this time. However, even leaving the potential actual causal connections unexplained, the results allow us to be confident about the validity of the model's station selection. We will discuss these results in the following section.

Training and Testing Periods

This study requires two separate datasets - one to train the model and one to test it against. Testing a model against its own training data would drastically skew the results in its favor. There are two rather different approaches to obtaining these two separate datasets:

- Using sets from different time periods
- Using the same dataset but only training the model on some of the included days while testing against the remaining days. Leave-one-out is a common example of this approach

Using data from different time periods will introduce a significant amount of uncertainty as weather patterns change from year to year, so we would likely skew the results in a negative fashion, effectively undervaluing the model's validity. However, training and testing against different subsets of the same dataset is a much less realistic indicator of the model's validity as it does not properly represent the ultimate purpose of the forecasting system. As weather phenomena can often be rather stable and consistent within a given season, they can vary greatly between different years. As a consequence of this long term variability, validation would be positively biased (Pers, Albrechtsen, Holst, Sørensen, & Gerds, 2009).

It is the study's declared goal to forecast flight delays in a different time period from the one the model was trained on. Therefore we chose to train against 2013 data and test against data from 2012. While this - as previously stated - has the potential of undervaluing our model, it also eliminates the risk of overvaluing it. We will attempt to

predict a prior year using data from a more recent time period to eliminate any learning effect that may have taken place in the carrier's network optimization algorithms.

Probabilistic Predictions Based On Historic Data

The model's merit needs to be evaluated against an established base line. We will derive such by creating a forecast based on historic data alone, calculating the true positive rate, true negative rate and combined accuracy for said probabilistic forecast and compare our model's performance against it.

For each forecast period, we will calculate the historic probability of irregularity occurring for any given flight and then create a forecast based on that probability. For example, for the month of July 2012, 17 out of 94 flights encountered irregularities. Hence the probability of such an irregularity for any flight was 18.1%. Based on this figure, the probabilistic forecast would have selected a random 17 flights as irregular. The results would have been a forecast predicting three true positives (TP), 14 false negatives (FN), 14 false positives (FP) and 63 true negatives (TN):

Actual Irregularities: 17

Thereof forecast as irregular (TP): 17*0.181=3

as normal (FN): 17+(1-0.181)=14

Actual Normal Operations: 77

Thereof forecast as irregular (FP): 77*0.181=14

as normal (TN): 77+(1-0.181)=63

Knowing TP, TN, FP and FN, one can calculate true positive rate, true negative rate and overall accuracy using the definitions widely accepted in the machine learning community (Zhenhua, 2009):

True Positive Rate:$TPR = \frac{\text{TP}}{\text{TP+FN}}$

True Negative Rate: TNR= $\frac{\text{TN}}{\text{FP+TN}}$

Accuracy: $ACC = \frac{\text{TP+TN}}{\text{TP+FP+TN+TF}}$

Using these definitions, we arrive at a rather low true positive rate of only 0.181, a true negative rate of 0.954 and an overall accuracy of only 0.181:

True Positive Rate:$TPR = \frac{\text{TP}}{\text{TP+FN}} = \frac{3}{3+14} = 0.181$

True Negative Rate: TNR= $\frac{\text{TN}}{\text{FP+TN}} = \frac{63}{66} = 0.954$

Accuracy: $ACC = \frac{\text{TP+TN}}{\text{TP+FP+TN+TF}} = \frac{17}{94} = 0.181$

True positive rate and accuracy are expected to be identical due to the fact that this forecast was created using historic probabilities.

Initial Results

Training on Annual Data

Table 6 shows the results from our initial iteration of training the model on a flight from Chicago's O'Hare airport to Fargo, ND. We used weather and flight data for the entire year of 2013 and then tested the resulting model against data from 2012. This achieved impressive overall improvements over a probabilistic forecast based on historic data.

However, especially during the summer months, TPR was negatively affected. January's results are also disappointing, yet the very small number of positive samples in that month is prone to causing low true positive rate.

Table 6: Results using Annual Training Data (A). Probabilistic Forecast (R) listed for comparison.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
TP	0	3	5	0	0	0	0	0	0	7	3	7	25
TN	37	37	57	45	77	100	77	90	72	35	64	43	734
FP	9	17	10	1	1	0	0	0	0	11	7	27	83
FN	2	4	7	1	10	9	17	11	6	6	5	7	85
TPR_R	.042	.115	.152	.021	.114	.083	.181	.109	.077	.220	.101	.167	.119
TPR_A	0	.429	.417	0	0	0	0	0	0	.538	.375	.5	.227
TNR_R	.998	.983	.969	.998	.984	.992	.953	.986	.993	.926	.987	.962	.982
TNR_A	.804	.685	.851	.978	.987	1	1	1	1	.761	.902	.614	.898
ACC_R	.042	.115	.152	.021	.114	.083	.181	.109	.077	.220	.101	.167	.119
ACC_A	.771	.656	.785	.957	.875	.917	.819	.891	.923	.712	.848	.595	.819
ACC_{Gain}	.729	.541	.633	.936	.761	.834	.638	.782	.846	.492	.747	.428	.7

Training on Monthly Data

When training the model on data for a specific month, we would expect the accuracy of the forecast to improve. Seasonality is inherent in weather patterns and therefore correlations learned for the summer months may be utterly unhelpful in forecasting delays in November or December. Compared to the previous approach using the entire year's data for training, Table 7 shows inconsistent changes using training data from just one month of the training year to forecast the same month in the testing year. In total, Table 8 reports a virtually unchanged average accuracy of 0.808, or a marginal 0.004 lower than in the previous approach. It is notable that the monthly training intervals appear to have improved the overall accuracy during the winter months. This is driven by an increased true negative rate while TPR is considerably lower. This is most likely a

result of the considerably smaller training sets during the winter months that make it more difficult for our machine to define an accurate threshold.

Table 7: Results using Monthly Training Data (M). Probabilistic Forecast (R) listed for comparison.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
TP	0	2	3	0	2	0	3	3	1	3	7	5	29
TN	40	47	59	38	70	100	70	79	69	37	1	54	664
FP	6	7	8	8	8	0	7	11	3	9	64	16	147
FN	2	5	9	1	8	9	14	8	5	10	7	9	87
TPR_R	.042	.115	.152	.021	.114	.083	.181	.109	.077	.220	.101	.167	.125
TPR_M	0	.286	.25	0	.2	0	.176	.273	.167	.23	.125	.357	.25
TNR_R	.998	.984	.969	.998	.984	.992	.953	.986	.993	.926	.987	.962	.980
TNR_M	.869	.870	.881	.826	.897	1	.909	.878	.958	.804	.944	.771	.819
ACC_R	.042	.115	.152	.021	.114	.083	.181	.109	.077	.220	.101	.167	.125
ACC_M	.834	.804	.785	.809	.818	.917	.777	.812	.897	.678	.861	.702	.748
ACC_{Gain}	.792	.689	.633	.788	.704	.834	.596	.703	.82	.458	.76	.535	.623

Table 8: Accuracy Improvements Using Monthly Data

ACC_{Year}	.771	.656	.785	.957	.875	.917	.819	.891	.923	.712	.848	.595	.812
ACC_{Month}	.834	.804	.785	.809	.818	.917	.777	.812	.897	.678	.861	.702	.808
Diff:	.063	.148	0	-.148	-.057	0	-.042	-.079	-.026	-.034	.013	.107	-.004

By using single months for training we have drastically reduced the size of the training set. This may well counteract any improvements achieved by the more analog weather patterns in matching seasons of different years. It is possible that using identical months from a larger number of years for training might generate substantial improvement. Such an approach, however, lies outside the computational limitations of this study. Within our scope we have to assume that using monthly data does not provide sufficient improvements in forecast validity to justify the substantially higher computational

requirements. Using annual data, only one expensive training cycle and twelve cheaper testing cycles are needed. Monthly data would require twelve cycles each.

Optimized Results

As previously stated, certain thresholds were employed in the preprocessing of the NOAA data. For example, any temperature in excess of 31C was classified as potential cause for operational irregularities. These relatively liberal thresholds result in a large share of positive features which adds substantial noise. We had intentionally accepted this noise in order to maximize the true positive rate. However, considering the rather disappointing results demonstrated above, those thresholds need to be reconsidered. Allowing the machine to optimize its own thresholds exponentially increase the computing resources required for preprocessing. It does, however, have the potential to yield a significantly better forecast.

After several computing intensive iterations, the machine arrived at the following improved thresholds:

- High Temperature Threshold: 20% of observations over 35C (95F)
- Low Temperature Threshold: 30% of observations under 0C (32F)
- Wind Speed Threshold: 30 kts
- Cloud Cover Threshold: average cover of greater than 80%
- Rain Threshold: 70% of observations
- Reduced Visibility Threshold: 10% of observations
- Low Pressure Threshold: 1002mbar
- Thunderstorm Threshold: Any reported thunderstorm any time

Table 9 shows that this adjustment somewhat improved the forecast's TPR based on annualized training data. True negative rate and accuracy are both lower throughout the year – especially during the winter months. However, our primary goal was to increase the true positive rate. In this context, these results present a notable improvement.

Table 9: Improvements Achieved by Optimizing Features

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
TP	2	5	5	0	5	1	4	3	2	10	6	13	56
TN	11	11	29	36	50	97	73	83	68	21	25	7	511
FP	35	43	38	10	28	3	4	7	4	25	46	63	306
FN	0	2	7	5	5	8	13	8	4	3	2	1	58
TPR_M	0	.429	.417	0	0	0	0	0	0	.538	.375	.5	.227
TPR_{OptA}	1	.714	.583	0	.5	.111	.235	.273	.333	.769	.750	.929	.491
TNR_M	.804	.685	.851	.978	.987	1	1	1	1	.761	.902	.614	.898
TNR_{OptA}	.239	.204	.567	.783	.641	.970	.948	.923	.944	.457	.350	.100	.626
ACC_M	.771	.656	.785	.957	.875	.917	.819	.891	.923	.712	.848	.595	.819
ACC_{OptA}	.271	.263	.570	.701	.625	.899	.819	.851	.897	.525	.392	.239	.610

Building on our previous approach, we also ran the entire procedure using monthly training data. Table 12 reveals considerable improvements over both the previous results (TPR_M, TNR_M, ACC_M) - using the original feature definitions and monthly data – and the values achieved by using the improved feature definitions and annual training data (TPR_{OptA}, TNR_{OptA}, ACC_{OptA}).

Table 10: Detailed Improvements

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Total
TP	19	5	12	0	9	2	8	11	2	13	7	8	
TN	9	51	41	41	52	85	28	38	48	18	33	68	
FP	16	3	26	5	26	15	49	52	24	28	38	2	
FN	4	2	0	1	1	7	9	0	4	0	1	6	
True Positive Rate													
TPR_M	0	.429	.417	0	0	0	0	0	0	.538	.375	.5	.227
TPR_{OptA}	1	.714	.583	0	.5	.111	.235	.273	.333	.769	.750	.929	.491
TPR_{OptM}	.826	.714	1.00	0.00	.900	.223	.471	1.00	.334	1.00	.875	.571	.660
True Negative Rate													
TNR_M	.869	.870	.881	.826	.897	1	.909	.878	.958	.804	.944	.771	.819
TNR_{OptA}	.239	.204	.567	.783	.641	.970	.948	.923	.944	.457	.350	.100	.626
TNR_{OptM}	.360	.944	.612	.891	.667	.850	.364	.423	.667	.391	.465	.971	.634
Accuracy													
ACC_M	.834	.804	.785	.809	.818	.917	.777	.812	.897	.678	.861	.702	.748
ACC_{OptA}	.271	.263	.570	.701	.625	.899	.819	.851	.897	.525	.392	.239	.610
ACC_{OptM}	.583	.918	.671	.872	.693	.798	.383	.485	.641	.525	.506	.905	.665

Finally, we introduced two additional wind direction features to improve classification. Previously we only considered westerly and northerly winds. These two features are sufficient to describe all four wind quadrants as previously described, but of course the Naïve Bayes approach only considers true features which could have undervalued the importance of easterly and southerly winds. Interestingly, the overall accuracy of the forecast dropped slightly to .61, almost exclusively as a result of a lowered True Negative Rate. This reduced performance prompted us to reattempt the calculation without any wind direction features (using only nine instead of the original eleven features). The results listed in table 11 seem to indicate that wind direction does not improve forecast

quality at all, but might instead be introducing unnecessary noise. This can be explained by the fact that most stations have predominant wind directions and that the observational period is quite long.

Table 11: Adding and Removing Wind Features

	TPR	TNR	ACC
11 Features	.66	.634	.665
13 Features	.66	.604	.61
9 Features	.63	.705	.69

CHAPTER IV

DISCUSSION

Findings

When forecasting flight delays, one can rely on historic probabilities and derive a forecast based on the assumption that these probabilities will describe the behavior of future flights. We have created such a forecast for an example route and tested it against a more sophisticated approach using national meteorological data to predict irregular flight operations. Compared to the probabilistic forecast, the model achieved significantly better results underscoring the validity of our approach.

As described in the previous section, we achieved accuracy improvements of 0.54 or by considering weather data instead of basing a forecast on historic probabilities alone. However, the study's main goal is to create a forecast of irregular operations (true positives). The true positive rate improved by 0.537. Table 12 summarizes these improvements.

Table 12: Improved TPR

TPR_R	.042	.115	.152	.021	.114	.083	.181	.109	.077	.220	.101	.167	.123
TPR_{OptM}	.826	.714	1.00	0.00	.900	.223	.471	1.00	.334	1.00	.875	.571	.660
Diff:													+.537

Paired with an accuracy of 66.5%, an overall true positive rate of 66.0% may not seem impressive to the casual reader, yet the results are clearly much better than the probabilistic forecast alternative.

We have shown that a Naïve Bayes approach can be used to successfully select a small subset of varying size from the comprehensive set of weather stations in the contiguous United States in order to create a valid forecast. This study further demonstrated that said forecast can be drastically optimized by adjusting the thresholds used for feature analysis in the preprocessing stage.

Future Research and Application

While our work demonstrates a promising new approach to forecasting flight delays using just one known dataset and ignoring the multitude of unknown variables, it also creates vast potential for improvements on method, thresholds, weights and features. A significantly larger set of operational performance data will likely not only validate our approach but could vastly improve the result's validity.

In addition to the mentioned improvements on data selection and processing, one must also consider the fact that the scope of this study is restricted to a very narrow methodological approach to the problem. Specifically, we are using binary feature expressions throughout our entire work while other approaches including decision trees would be able to handle non-binary feature expressions and could potentially improve findings significantly. Future research may hence be based on a replication of this study using advanced feature selection and analysis algorithms such as the aforementioned decision trees. Within the binary system, the proposed forecast method is - with the exception of data preprocessing of course - universally applicable to any n-dimensional feature system and forecast model.

Finally, we have demonstrated that the forecast validity can be drastically improved by dynamically adjusting not only the feature weights within the forecasting sequence, but also the thresholds used in preprocessing and feature generation. This was accomplished by manipulating individual features and using a hill climbing approach. This optimization is relatively computing expensive and also does not guarantee a global optimum. We therefore propose more sophisticated approaches based genetic algorithms (Fonseca & Fleming, 1993) or differential evolution (Storm & Price, 1995) for future research on the subject matter. These algorithms will be able to more efficiently optimize weights, particularly as the n-dimensional feature space is not likely to be convex but will instead include numerous local maxima (Bianchi & Jakubowicz, 2013). Further improvements may be achieved by redesigning the experiment based on support vector machines (Suykens & Vanderwalle, 1999) or random forest approaches (Liaw & Wiener, 2002).

Taking these limitations into account, our work merely lays the foundation for the described further research and should not be understood as a comprehensive model to be used in a real world implementation. Through additional research we will be able to significantly improve on the validity of the created forecast and hopefully reduce computing cost drastically. This will enable us to proactively complete the learning procedure for a large number of routes and operators and create a database of forecasting models. These models can then be quickly deployed to process current meteorological data to create on-demand real time forecasts for these routes. This short response time will allow us to not only serve the academic community or interested members of industry, but also the general travelling public at large.

The relatively low computing resources required to create real time forecasts using forecast models from the preprocessed database mean that a relatively lightweight server or server bank can be used to provide the results using a web service interface that can be accessed by a wide array of clients. The traveling public, for example, might use a readily available mobile application to retrieve a delay or cancellation forecast for their next flight. They will then be able to make an informed decision and potentially consider alternative flight options.

REFERENCES

Agustsson, H., & Olafsson, H. (2009). Forecasting wind gusts in complex terrain. *Meteorology and Atmospheric Physics* , 173-185.

Ball, M., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A., et al. (2010). *Total Delay Impact Study.* NEXTOR.

Ball, M., Barnhart, C., Nemhauser, G., & Odoni, A. (2006). Managing Air Traffic and Airline Operations for Schedule Reliability. *Operations Research Handbook on Transportation, Gilbert Laporte and Cynthia Barnhart* .

Bianchi, P., & Jakubowicz, J. (2013). Convergence of a Multi-Agent Projected Stochastic Gradient Algorithm for Non-Convex Optimization. *IEEE Transactions on Automatic Control.*

Billet, J. A., & Titlow, J. (2010). An examination of stability changes allowing for increased gustiness over the Mid Atlantic from tropical systems. *29th Conference on Hurricanes and Tropical Meteorology.* Tucson, AZ: American Meteorological Society.

Black, J. (2010). A Look at Airport Delays Due to Ceiling and Visibility Conditions. *American Meteorological Society* .

Bruce, P. J. (2011). Decision-making in airline operations: the importance of identifying decision considerations. *International Journal of Aviation Management* , 89-104.

Cai, J., Wang, H., & Zhou, D.-X. (2002). Gradient learning in a classification setting by gradient descent. *Journal of Approximation Theory* , 674-692.

Currie, G. N., Dickey, R. W., Duclos, R., & Price, W. L. (1974). Forecasting traffic in an air transport network. *Operations Research Society of America and Institute of Management Sciences.* San Juan.

Dacre, H. F., Grant, A., & Johnson, B. T. (2013). Aircraft observations and model simulations of concentration and particle size distribution in the Eyjafjallajokull volcanic ash cloud. *Atmospheric Chemistry and Physics* , 1277-1291.

DeGaetano, A. T., & French, J. R. (1991). Black Hills precipitation climatology based on surface wind direction. *South Dakota School of Mines and Technology* .

Dell'Olmo, P., & Lulli, G. (2003). A new hierarchical architecture for Air Traffic Management: Optimisation of airway capacity in a Free Flight scenario. *European Journal of Operational Research* , 179-193.

Ding, J., & Li, H. (2011). The Flight Delays Propagation Analysis Model for Hub Airports based on the Danger Model Theory. *International Journal of Digital Content Technology* , 229.

Dray, L. M., Evans, A., Vera-Morales, M., Reynolds, T. G., & Schafer, A. (2008). Network and Environmental Impacts of Passenger and Airline Response to Cost and Delay. *AIAA Aviation Technology, Integration and Operations (ATIO) Conference.* American Institute of Aeronautics and Astronautics Inc.

Federal Aviation Administration. (2014, 07 24). Aeronautical Information Manual. *AIM* .

Federal Aviation Administration. (2008). *Density Altitude.* Washington, DC.

Gronlund, S. D., Dougherty, M. R., Durso, F. T., Canning, J. M., & Mills, S. H. (2005). Planning in Air Traffic Control: Impact of Problem Type. *The International Journal of Aviation Psycholog* , 269-293.

Hauf, T., & Sasse, M. (2002). The impact of thunderstorms on landing traffic at Frankfurt airport (Germany) - A case study. *10th Conference on Aviation, Range, and Aerospace Meteorology* (pp. 160,161). Portland, OR: American Meteorological Society.

Hebert, J. E., & Dietz, D. C. (1997). Modeling and analysis of an airport departure process. *Journal of Aircraft* , 43-47.

Idris, H. R. (2002). Observation and analysis of departure operations at Boston Logan International Airport. *Dissertation Abstracts International* .

Jiang, L. (2012). Learning Instance Weighted Naive Bayes from labeled and unlabeled data. *Journal of Intelligent Information Systems* , 257-268.

Konrad, C. E. (1994). Moisture trajectories associated with heavy rainfall in the Appalachian region of the United States. *Physical Geography* , 227-248.

Kossin, J. P., & Schubert, W. H. (2001). Mesovortices, Polygonal Flow Patterns, and Rapid Pressure Falls in Hurricane-Like Vortices. *Journal of the Atmospheric Sciences* , 2196.

Lacher, A. R., & Ball, C. G. (2002). Gridded Congestion Forecast - A mechanism for shared situational awareness to manage en route congestion? *AIAA Aircraft Technology, Integration, and Operations (ATIO) Forum.* Los Angeles, CA: American Institute of Aeronautics and Astronautics, Inc.

Li, J., & Ding, J. (2008). Analysis of Flight Delay Propagation Using Bayesian Networks. *Acta Aeronautica et Astronautica Sinica* , 1598-1604.

Li, J., Heap, A. D., Potter, A., & Daniell, J. J. (2011). Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling and Software* , 1647-1959.

Lindholm, T. A., Hage, F., Wade, C., & Rasmussen, R. M. (1997). Weather support to Ground Deicing Decision Making System (WSDDM) real-time snowfall rate display, operational utility, and benefits. *Conference on Aviation, Range and Aerospace Meteorology* (pp. 152-157). Long Beach, CA: American Meteorological Society.

Lorentson, M. (2011). 15th Conference on Aviation, Range, and Aerospace Meteorology. *The use of aviation system performance metrics and measures in qualitative forecast evaluation.* Los Angeles, LA: American Meteorological Society.

Matus, A., Hudnall, L., Murray, J. J., & Krueger, A. (2010). The Impacts on Air Traffic of Volcanic Ash from the 2009 Mt. Redoubt Eruption. *14th Symposium on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface* . Atlanta, GA: American Meteorological Society.

Mayer, R. H., & Sprong, K. R. (2008). Improving Terminal Operations - Benefits of RNAV Departure Proceedures. *AIAA Aviation Technology, Integration and Operations (ATIO) Conference.* American Institute of Aeronautics and Astronautics Inc.

Meyer, D., Saghi, B., & Tarnai, G. (2008). Safety management of traffic growth in air transportation. *Periodica Polytechnica, Transportation Engineering* , 69-72.

Missoni, E., Nikolic, N., & Missoni, I. (2009). Civil Aviation Rules on Crew Flight Time, Flight Duty, and Rest: Comparison of 10 ICAO Member States. *Aviation, Space, and Environmental Medicine* , 135-138.

Mueller, E. R. (2002). Analysis of aircraft arrival and departure delay characteristics. *AIAA Aircraft Technology, Integration, and Operations (ATIO) Forum.* Los Angeles, CA: American Institute of Aeronautics and Astronautics, Inc.

National Oceanic and Atmospheric Administration. (2014). METAR Training Manual.

National Oceanic and Atmospheric Administration. (2014). *National Climatic Data Center*. Retrieved 07 10, 2014, from National Oceanic and Atmospheric Administration: http://cdo.ncdc.noaa.gov/qclcd_ascii/

Nettleton, D., Orriols-Puig, A., & Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review* , 275-306.

Nogami, J. (1995). Concept of the future air traffic management and sensitivity analysis on the propagation of congestion delay through airlines' flight network. *33rd Aircraft Symposium* (pp. 677-680). Hiroshima: Japan Society for Aeronautical and Space Sciences.

Partain Jr, J. (2008). NWS Alaska Region: Science and Data Needs in an Era of Changing Climate. *Symposium on Linkages Among Societal Benefits, Prediction Systems and Process Studies for 1-14 day Weather Forecasts.* Boston, MA: American Meteorological Society.

Pers, T. H., Albrechtsen, A., Holst, C., Sørensen, T. I., & Gerds, T. A. (2009). The Validation and Assessment of Machine Learning: A Game of Prediction from High-Dimensional Data. *PLoS ONE* , 6287.

Pisano, D. (2008). Blind Landings: Low-Visibility Operations in American Aviation, 1918–1958. *ISIS* , 194-196.

Robinson, P. (1989). The influence of weather on flight operations at the Atlanta Hartsfield International Airport. *Weather and Forecasting* , 461-468.

Rubin, J. (1977). A survey of techniques for airline crew scheduling. *Operations Research Society of America and Institute of Management Sciences, Joint National Meeting.* Las Vegas.

Sanderson, M. (1993). Prevailing trade winds: weather and climate in Hawaii. *University of Hawaii Press* , 126.

Scott, R. W., & Huff, F. A. (1996). Impacts of the Great Lakes on regional climate conditions. *Journal of Great Lakes Research* , 845-863.

Seelhorst, M., & Hansen, M. (2014). *Flight Cancellation Behavior and Aviation System Performance.* National Center of Excellence For Aviation Operations Research.

Shavell, Z. A. (2001). Effects of schedule disruptions on the economics of airline operations. *Effects of schedule disruptions on the economics of airline operations* .

Slattery, R. A., & Cheng, V. H. (1997). Sensitivity of en-route scheduling to variable separation in the terminal area. *AIAA Guidance, Navigation, and Control Conference.* New Orleans, LA: American Institute of Aeronautics and Astronautics Inc.

Sood, N., Mulgund, S., Wanke, C., & Greenbaum, D. (2007). A Multi-Objective Genetic Algorithm for Solving Airspace Congestion Problems. *AIAA Guidance, Navigation, and Control Conference and Exhibit.* Hilton Head, SC: American Institute of Aeronautics and Astronautics Inc.

Sun, D., Clinet, A., & Bayen, A. (2011). A dual decomposition method for sector capacity constrained traffic flow optimization. *Transportation Research Part B* , 880-902.

van Es, G. W., van der Geest, P. J., & Nieuwpoort, T. M. (1999). Safety aspects of aircraft operations in crosswind. *11th Annual European Aviation Safety Seminar* (pp. 275-322). Alexandria, VA: Flight Safety Foundation.

Wang, T.-C., & Tsao, C.-H. (2012). Time-Based Separation for Aircraft Landing Using Danger Value Distribution Flow Model. *Mathematical Problems in Engineering* .

Wang, X., & Regan, A. (2006). Dynamic yield management when aircraft assignments are subject to swap. *Transportation Research Part B: Methodological* , 563-576.

Zhenhua, L. (2009). *Computational Intelligence and Intelligent Systems: 4th International Symposium on Intelligence Computation and Applications.* Springer.

Zobell, S. M., Ball, C. G., & Sherry, J. E. (2001). Rerouting as a strategy for collaborative weather problem resolution. *AIAA, Aircraft, Technology Integration, and Operations Forum.* Los Angeles, CA.