Frequency Analysis Technique Applied to Transportation Freight, Goods and Services Data

BD-550-03

Systems Planning Office
605 Suwannee Street
Tallahassee, FL 32399-0450
(850) 414-4911

Project Manager: Huiwei Shen

February 2006

Final Report

Principal Investigator: Xin Li, Ph.D.

Research Team Members: Mohamed Abdel-Aty, Ph.D., P.E.
Martin Michalak, Research Assistant
Yuqiong Bai, Research Assistant
Chris Lee, Ph.D.
Anurag Pande, Ph.D.

University of Central Florida
Orlando, FL 32816

DISCLAIMER

The opinions, findings, and conclusions expressed in this publication are those of the authors and not necessarily those of the State of Florida Department of Transportation.

Technical Report Documentation Page

1. Report No.

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Frequency Analysis Technique Applied to Transportation Freight, Goods and Services Data

5. Report Date
February 2006

6. Performing Organization Code

7. Author(s)
Xin Li, Mohamed Abdel-Aty, Martin Michalak, Yuqiong Bai, Chris Lee, Anurag Pande

8. Performing Organization Report No.

9. Performing Organization Name and Address
University of Central Florida
Orlando, FL 32816

10. Work Unit No. (TRAIS)

11. Contract or Grant No.
BD-550-03

12. Sponsoring Agency Name and Address
Florida Department of Transportation
605 Suwannee Street, MS 30
Tallahassee, FL 32399

13. Type of Report and Period Covered
Final Report:
September 2004 – February 2006

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Patterns in transportation data sets can be used in planning. The detection of hidden periodic patterns requires sophisticated methods in data analysis. In this project, the frequency analysis method is applied in detecting hidden periodic patterns in transportation data sets. The recommended approach can be summarized as a two-stage strategy in which the frequency analysis method is used to provide a preliminary and quick detection of possible periodicities in the data set as a first stage; then in order to explain the causes for the detected hidden periodicities, a list of possible variables having similar periodicities is studied in a detailed regression analysis. The results from the regression analysis can be used to explain the possible causes for the detected periodicities.

17. Key Word
Data collection, Urban goods movement, Service agencies, Transportation planning, Frequency analyzers, Traffic counts, Peak hour traffic, Analysis

18. Distribution Statement
No restrictions.

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages

22. Price

Reproduction of completed page authorized

EXECUTIVE SUMMARY

We carried out the frequency analysis method in detecting hidden patterns in transportation data sets. Our approach can be summarized as a two-stage strategy in which we first use the frequency analysis method to do a preliminary and quick detection of possible periodicities in the data set. Then from the detected periodicities, in order to explain the cause for them, we try to come up with a list of possible variables having similar periodicities and do a detailed regression analysis. The results from the regression analysis can be used to explain the possible cause for the detected periodicities. Some details are presented in a draft of a research paper to be submitted for publication.

PROBLEM STATEMENT

There are many data sets collected in the study of transportation freight, goods, and services. In order to make these data sets useful for transportation planners and decision makers in improving or designing more flexible and efficient transportation systems, sophisticated data analysis must be carefully performed. One of such needs is in detecting the periodicity patterns in the data sets. A simple example of a periodicity pattern is the daily peak hours of the traffic counts. But there may be many hidden periodicity patterns that are not as easily observable. The recently developed tool of frequency analysis can be used in such study.

OBJECTIVES

The objectives of this research project are to develop:

- a quantitative analysis of the transportation data sets (such as freight, goods, and services), in particular the detection of hidden periodicities within the data set
- interpretation of the findings (periodicity patterns) and recommendations for planners and decision makers
- a guide line for applying the technique on other data sets
- a prototype of a user friendly computer environment for detecting hidden periodicities based on frequency analysis techniques

FINDINGS AND CONCLUSIONS

A large collection of transportation data sets are studied, including the 2004 release of the 2001 National Household Travel Survey from Bureau of Transportation Statistics (BTS), the Waterbourne Commerce: Foreign Traffic Vessel Entrance and Clearances data set from the Army Corps of Engineers, the Airline data set that contains daily records of actual departure times, arrival times, origins and destinations for non-stop domestic flights by major air carriers from 1988 to 2004 from BTS, the Visit Florida data that summarizes travel patterns of Florida residents and domestic/Canadian visitors to Florida in years 2000 and 2001, and the data from a private company, Reebie Associates, that describes commodity flow among 247 market areas in Florida and 16 counties bordering on Florida. To best demonstrate the use of our method of frequency analysis, we concentrated our study on the Airline data set, in particular the flight delay patterns.

The interpretation of the results from frequency analysis presented most challenges. Our approach can be summarized as a two-stage strategy in which we first use the frequency analysis method to do a preliminary and quick detection of possible periodicities (especially the hidden ones) in the data set. Then from the detected periodicities, in order to explain the causes for them, we try to come up with a list of possible variables having similar periodicities and do a detailed regression analysis and analysis of variance. The results from the regression analysis and analysis of variance are used to explain the possible causes for the detected periodicities.

A complete set of Matlab programs are developed for frequency analysis. A user friendly interface is provided for easy use of the method in dealing with other transportation data sets.

BENEFITS

Our proposed method has been successfully used in detecting hidden patterns in airline flight delay data set, together with some preliminary application in traffic volume data sets. With the proved applicability, our methodology will open the way to many applications in transportation planning and engineering. In the context of the transportation of freight and goods, if certain periodical patterns are detected, then this would help engineers, planners and decision takers prepare in advance for expected changes. For example, detecting periodical changes in freight movements would help in the anticipation of higher trucks demand at certain locations of the transportation network. This could help planners determine operation hours of trucks and the effect of trucks on the level of service of the roadways of interest, traffic engineers could adjust signal timing to meet the increased truck demand, infrastructure engineers could anticipate the effect of increased truck movements and patterns on pavement conditions, etc.

One of the main concerns in Florida is tourists and visitors. Our method can be beneficial to the state when it is used to detect the patterns and locations of increased tourist activities from meaningful data sets collected during an extensive period of time, helping decision makers and planners meet the periodical increase in demand. Traffic safety is another possible future application, where we can anticipate the patterns in traffic crashes, helping us to expect seasonal problems, and therefore prepare emergency vehicles and personnel.

Table of Contents

List of Figures

List of Tables

Introduction

The research team members from both departments (Mathematics and Civil and Environmental Engineering) have started to meet regularly for some preparatory work for the project since early September of 2004. This includes making the list of agencies/organizations from which our data sets are to be collected. Once we received the official notice to proceed with work on the project on October 11, 2004, we started to work on the project by contacting the selected agencies/organizations for transportation freight, goods and services data sets. For additional selection of data sets to be used in the project, CD-ROM's were ordered from BTS. From then, quarterly progress reports were prepared and sent to Florida Department of Transportation on a regular basis. All tasks were finished on time. This final report is structured as follows: Sections are divided according to the tasks. In each section, detailed findings and outcomes of the project will be provided. The report ends with two appendices, one for the draft of a research paper reporting some results through this project, another for a user guide for using the included Matlab codes/programs for detection of hidden patterns by the frequency analysis method.

Task 1: Obtain the data set and perform preliminary transformation or normalization

The main objective is to obtain as much data as possible and perform preliminary testing on the usability of these data sets. In the meantime, we also did an extensive literature survey on recent research on all related transportation issues like activity analysis and travel patterns.

Data Sets

We started with the data sources provided in the Appendix A of the 2002 report "Freight, Goods and Services Mobility Strategy Plan" by Metroplan Orlando. We found several interesting data sets. In particular, we searched the data sets from the Bureau of Transportation Statistics, Army Corps of Engineers, and Reebie Associates.

2001 National Household Travel Survey

In the attempt to include travel data sets in our study, we acquired CR-ROM's from BTS. This data set contains January 2004 release of 2001 NHTS national sample data. After a close look at the data set, we found there were 13 months of data. The data in each month are recorded in the days of the week (that is, Mondays, Tuesdays, Wednesdays, Thursdays, and Fridays), not in the date in the month (like April 12, May 26, etc.). In each day, the times of trips are recorded in hour and minute. We decided to try the starting time of the first trip in each household as our first sequence of data. We extracted the data for Florida households only. We looked at the data by using the frequencies in various time intervals (like 30 min., 1 hour, 2 hours, 4 hours, etc.). Since the samples sizes are different from a month to another month, we normalized the data by using proportions. The preliminary testing shows that the Frequency Analysis method does reveal the pattern as we can see from the plot of the sequence itself. But, the detection of the hidden pattern requires further investigation and on a larger data set.

To illustrate the results of our algorithm, we now show some plots of the detected frequencies. Each "+" sign represents a zero of some orthogonal polynomial generated by the algorithm. The angle between the horizontal axis and the line connecting a "+" sign and the origin is a detected *frequency*. The concentration of the "+" signs indicate a good detection.

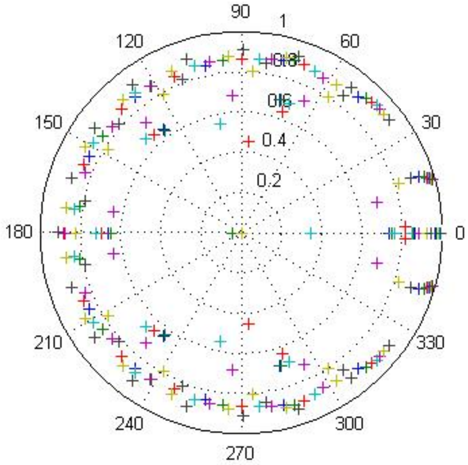


Figure 1. Florida Housholds Trips, frequencies in 1-hour interval, using degree 7 through 20

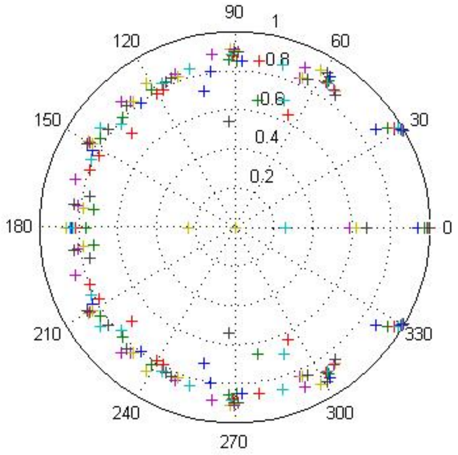


Figure 2. Florida Households Trips, frequencies in 2-hours interval, using degree 7 through 20

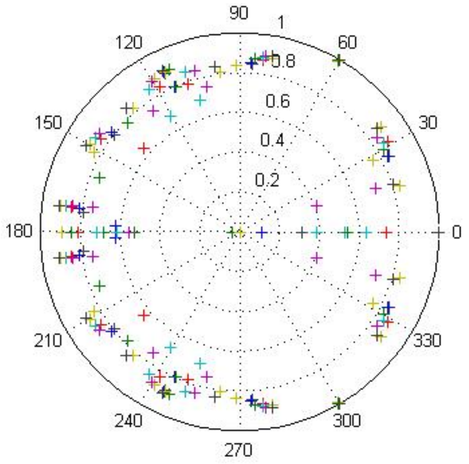


Figure 3. Florida Households Trips, frequencies in 4-hours interval, using degree 7 through 20

In all three cases, there are strong indications of patterns (where the "+" approaches the circle) with one or two detected frequencies. Fixing the degree (say, 6) of the orthogonal polynomial, we generated the following plot as the data sizes changed. The longer data set, the better the detection.

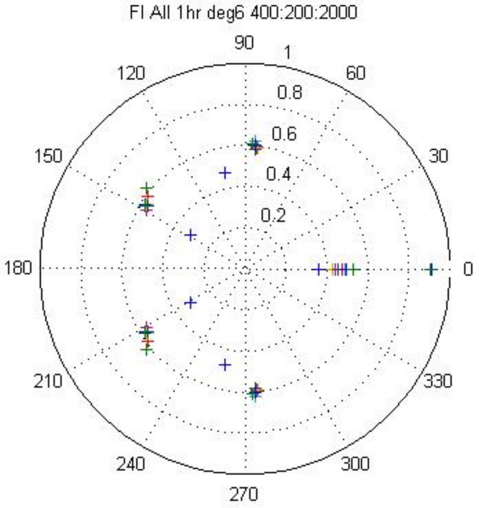


Figure 4. Florida Households Trips, frequencies in 1-hour interval, using degree 6, sample size varying between 400 and 2000

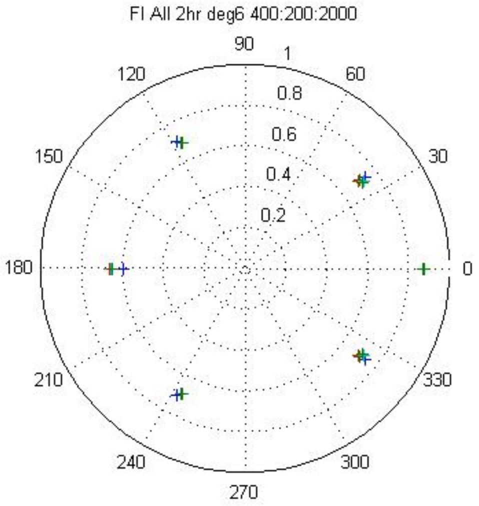


Figure 5. Florida Households Trips, frequencies in 2-hours interval, using degree 6, sample size varying between 400 and 2000

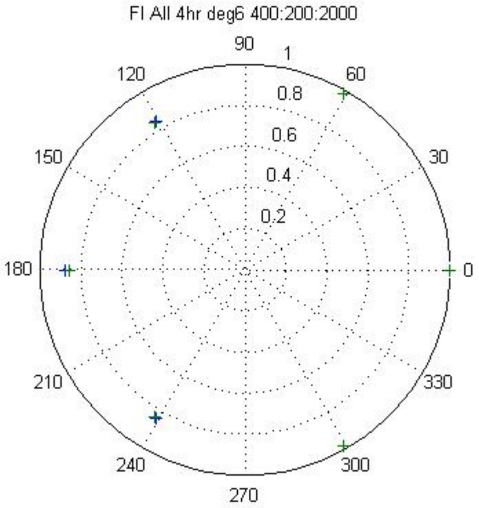


Figure 6. Florida Households Trips, frequencies in 4-hours interval, using degree 6, sample size varying between 400 and 2000

They all show the clear pattern (where the "+" signs converge) that we had expected: there is a peak in the morning. The above test indicates that the method of frequency analysis is capable in detecting the pattern in the data set (the known peak hours in this case). But for the hidden pattern, further study is required. One limitation of this data set is the relative small sample size. Of course, that the data was not recorded by dates is another major shortcoming as far as our methodology is concerned. We decided to continue working with this data set while we look for the other data sets that are more suitable for our method.

We also looked at Florida data of all starting time for "to work" trips. Here is a plot:

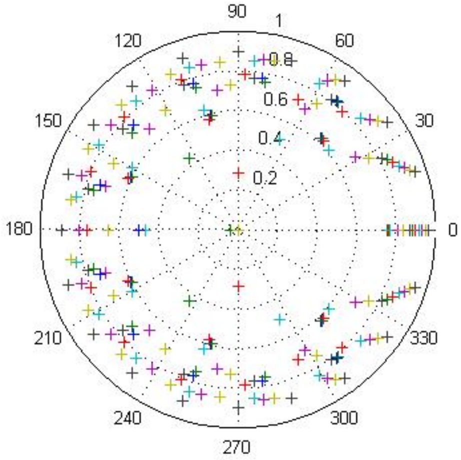


Figure 7. Florida Households "to-work" trips, 1-hour interval, using degree 7 to 20

Again, the tendency to the circle can be observed on the right semicircle. This indicates possible frequencies the data set processes have been detected.

Waterbourne Commerce: Foreign Traffic Vessel Entrance and Clearances

This data set is from the Army Corps of Engineers. It provides detailed raw data on foreign ships entering and leaving the US ports. The data is not aggregated: the temporal resolution of the data is, at least daily. The data spans from 1997 to 2002. It contains more than 86,000 records (about 10 per hour). We decided to further investigate this data set using our Frequency Analysis method. The data is freely available at: http://www.iwr.usace.army.mil/ndc/data/dataclen.htm

Airline Data:

This data set is the most promising in terms of the applicability of our method. The data is closely related to tourist travels. This data contains daily records of actual departure times, arrival times, origins and destinations for non-stop domestic flights by major air carriers from 1988 to 2004. The data is available at the Bureau of Transportation Statistics (BTS) web site (http://www.transtats.bts.gov). We decided to analyze the frequency of flight departures and arrivals using this data, and identify temporal repetitive patterns of tourist travels in Orlando (or Florida). We tried various ways of using the data sets, eventually moved to concentrate our effort on the flight delay patterns. Here are some experiments we performed on the data sets:

(1) Airline arrival total volume.

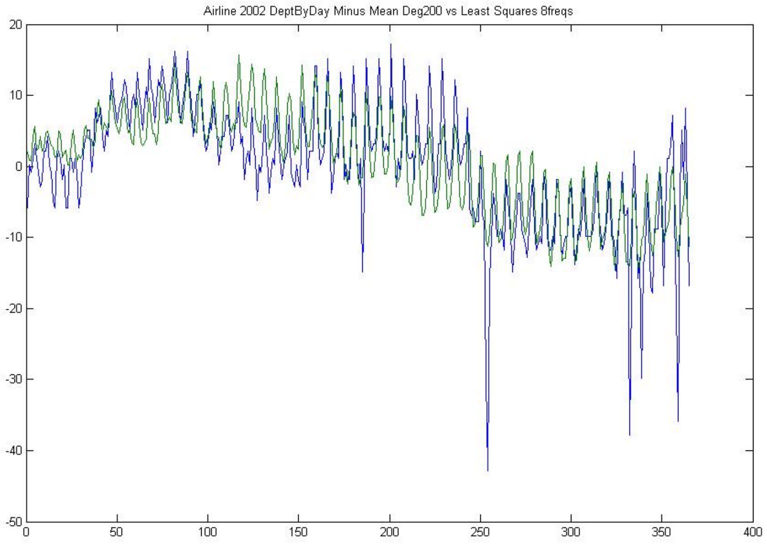


Figure 8. Airline arrival volume to Orlando International Airport plotted (in blue) along with a least squares best fit (in green) for the frequencies found from the frequency detection algorithm.

We stopped working with the raw arrival volume because the major trends in volume seemed to change substantially six to 24 months, completely dominating any long-term periodic effects.

(2) Airline arrival volume by departing airport. For these experiments, we examined the volume of arrivals at Orlando categorized by departure airport. This analysis revealed that airports in the same part of the country had similar patterns of flight volumes to Orlando.

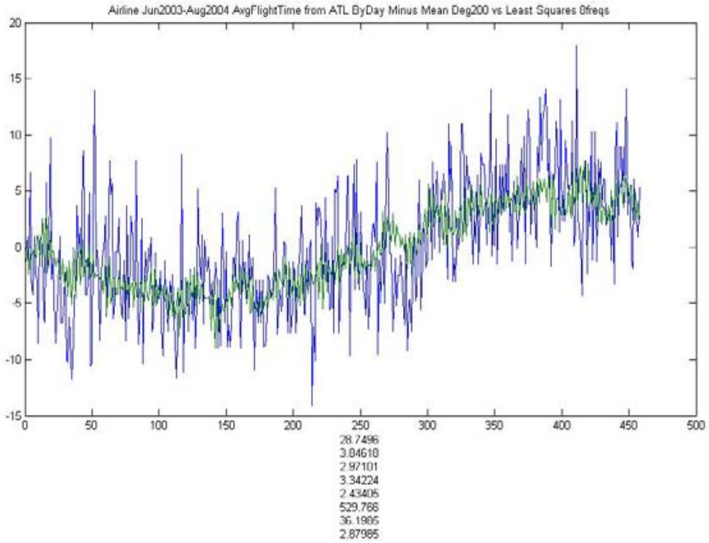


Figure 9. Airline arrival volume to Orlando International Airport plotted (in blue) from Atlanta Airport only along with a least squares best fit (in green) for the frequencies found from the frequency detection algorithm.

(3) Average delay per flight in time block.

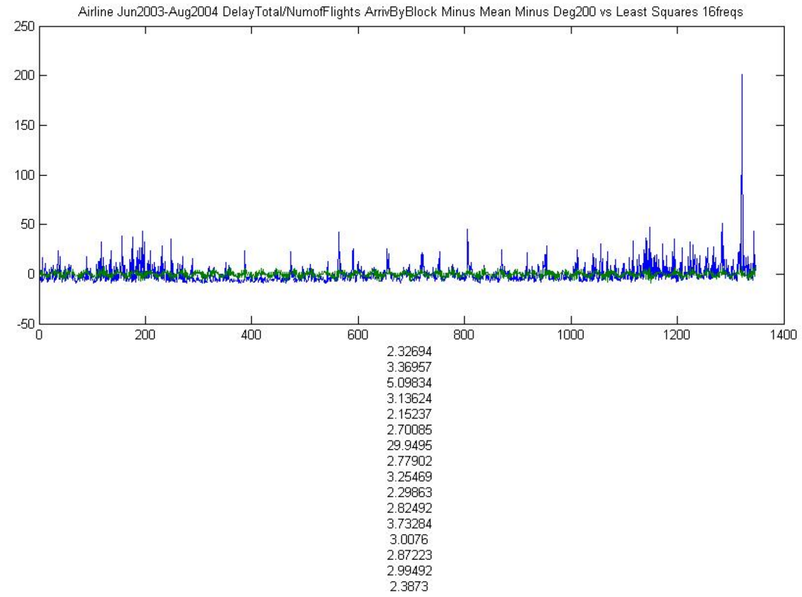


Figure 10. Average delay per flight of flights arriving at Orlando International Airport plotted (in blue) along with a least squares best fit (in green) for the frequencies found from the frequency detection algorithm.

When examining datasets for patterns of length less than a day, it was necessary to accumulate data into morning afternoon and evening activity periods to maintain a statically significant amount of data per data point and to eliminate zero activity periods during the late night.

Visit Florida Data

We carefully looked at the data sent to us by the project manager, Ms. Huiwei Shen. The data contains travel patterns of Florida residents and domestic/Canadian visitors to Florida. This data is a summary of annual travel records in the years 2000 and 2001. At the time of the research, we could not make much use of the data in its given hard copy form.

Reebie Associates Data Sets

These data sets were also sent to us by the project manager, Ms. Huiwei Shen. The data describes commodity flow among 247 market areas in Florida and 16 counties bordering on Florida. The database contains the information for 387,263 records by origin, designation, and commodity. Each record contains tonnage and value for each of the seven modes. Again, the data sets are aggregated and not in the raw data form, not immediately suitable for our method.

Literature Review (see the References for all cited publications)

From our preliminary study on the available data sets, we decided to concentrate on the flight delay patterns. Our decision is based on the following considerations: (i) The data set is fairly large and complete during a long time period from which we can easily choose a two-years or three-years period. (ii) Airline flight delay patterns will impact on the local traffic network. (iii) The cause of flight delay is itself an interesting problem. We will not aim at finding a complete answer to the causes. We believe that our understanding of the hidden periodic patterns inn the flight delays can help us to find the causes. (iv) Flight delay patterns reflect certain aspects of tourists travel patterns. In our study we considered delays of arriving flights to the Orlando International Airport. The same methodology may applied to other airports. Based on these concerns and facts, we chose flight delay patterns as the focus for our application of frequency analysis method. The following is a review of literature on the related research topics.

The increase in delays in the National Airspace System (NAS) has been the subject of studies in recent years. These reports contain delay statistics over the entire NAS, along with some data specific to individual airports. The various causal factors related to aircraft, airline operations, change of procedures and traffic volume are also discussed. The FAA describes the increase in delays and cancellations from 1995 through 1999. They found that the current system for collecting causal data does not provide the appropriate data for developing strong conclusions for delay causes and recommend changes to the current data collection system.

Systems and S.S. Allan (2001) examined delays at New York from September 1998 through August 2000 to determine major causes of the delay which occurred during the first year of ITWS use and delays that occurred with ITWS in operation that were "avoidable" if enhanced weather detection. The methodology used in the study has considered major causes of delays (convective weather inside and well outside the terminal area, and high winds) that have generally been ignored in previous studies of capacity constrained airports such as EWR. The research found that the usual paradigm of assessing delays only in terms of IMC and VMC conditions and the associated airport capacities is far too simplistic as a tool for determining which air traffic management investments best reduces the "avoidable" delays.

Lisa and David (2001) used the Detailed Policy Assessment Tool (DPAT) to model the propagation of delay throughout a system of airports and sectors. To estimate delays, throughputs, and air traffic congestion in a typical scenario of current operations in the US, DPAT models the flow of approximately 50,000 flights per day throughout the airports and airspace of the U. S. National Airspace System (NAS) and can simulate flights to analyze delays at airports around the world. They obtained results for local flight departure and arrival delays due to IMC, propagation for IMC, comparisons to VMC results, and a comparison of propagated delays to entire system.

Shomik, Mark and George (2002) presented an analysis of the possible impact of the application of slot controls as a demand management measure at San Francisco International Airport (SFO). A deterministic queuing model that uses an actual arrival schedule as input and simulates arrival delay based on available arrival capacity was used to estimate delay reduction potential of slot controls. The conclusions show the overall potential of slot controls to alleviate delay at SFO and their non-delay consequences.

Shangyao, Chi and Miawjane (2002) proposed a simulation framework, that is not only able to analyze the effects of stochastic flight delays on static gate assignments, but can also evaluate flexible buffer times and real-time gate assignment rules. The results of testing the framework on actual Chiang Kai-Shek airport operations were good, showing that the framework could be useful for airport authorities to perform gate assignments.
Mark Hansen (2002) analyzed runway delay externalities at Los Angeles International Airport (LAX) using a deterministic queuing model. This model allows estimating the delay impact of each specific arriving flight on each other specific arriving flight. The research finds that, despite being only moderately congested (average queuing delay only 4 min per arriving flight), individual flights can generate as much as 3 aircraft-hours of external delay impact on other flights, with an average impact of 26 aircraft-minutes and 3400 seat-minutes. About 90 percent of this impact is external to the airline as well as the flight, a consequence of the lack of a dominant airline at LAX.

Andrew Rosen (2002) measured the change in flight times resulting from infrastructure-constant changes in passenger demand. Results indicate that delays rise with the ratio of demand to fixed airport infrastructure, decreasing average flight times by close to seven minutes after the sharp decrease in demand in the fall of 2001. Flight time differences between the airlines in the sample are small, though the larger United had shorter average flight times in the winter quarter than America West, the smaller airline in the data sample.

Milan Janic(2003) presented a model for assessment of the economic consequences of large-scale disruptions of an airline single hub-and-spoke network expressed by the costs of delayed and cancelled complexes of flights. The model uses the scheduled and affected service time of particular complexes to determine their delays caused by disruption.

On-time performance of airlines schedule is key factor in maintaining current customer satisfaction and attracting new ones. However, flight schedules are often subjected to irregularity. Due to the tight connection among airlines resources, delays could dramatically propagate over time and space unless the proper recovery actions are taken. During the last decade, a considerable attention has seen been given to proactive schedule recovery models as a possible approach to limit flight delays associated with GDPs (Abdelghany et al., 2004a; Clarke, 1997). In these models, the impact of any reported flight delays, due to GDP or any other reason, is propagated in the network to determine any possible down-line disruptions (Monroe and Chu, 1995).

Yoshinori Suzuki (2000) proposed a new method of modeling the relationship between on-time performance and market share in the airline industry. The idea behind the method is that the passengers decision to remain (use same airline) or switch (use other airlines) at time t depends on whether they have experienced flight delays at time t-1 or not.

Khaled, Sharmila, Sidhartha, and Ahmed (2004) present a flight delay projection model, which projects flight delays and alerts for down-line operation breaks for large-scale airlines schedules. The results show that down-line schedule disruptions are proportional to the number of flights impacted by the GDP. Furthermore, in the recorded GDP instances, aircraft appears to be the reason for most flight delays predicted by the model.

Mark Hansen and Chieh Yu Hsiao (2005) analyzed the recent increase in flight delay in the US domestic system by estimating an econometric model of average daily delay that incorporates the effects of arrival queuing, convective weather, terminal weather conditions, seasonal effects, and secular effects. Results suggest that, controlling for these factors, delays decreased steadily from 2000 through early 2003, but that the trend reversed thereafter. The results identify key sources of recent increases in delay.

Mark Hansen, Yu Zhang (2005) investigated the interaction between LGA and the rest of the aviation system by estimating simultaneous equations of average LGA and National Airspace System (NAS) delay using two-stage least squares. The results demonstrate that the arrival delay impact of AIR-21 on LGA was in the form of increased Ground Delay Program (GDP) holding, and that while delay increased markedly under AIR-21 there were also observable improvements in the ability of LGA airport to handle traffic.

Mark Hansen and Dale Peterman (2004) used censored regression to analyze the delay impacts of the implementation of Traffic Management Advisor (TMA) metering at Los Angeles International Airport (LAX) in order to assess whether and how they have affected NAS performance.

The current method of valuing delay in benefit-cost analysis is insufficient for determining the distributional impacts of a technology change on users because it fails to account for the shifts in where benefits occur and to which users. Adib, Melissa and William (2004) proposed a theoretical framework for evaluating the distributive effects of technology changes that requires a new approach to the evaluation of delay and understanding efficiency in light of the state of the system. The framework defines different categories of delay per flight and a method for calculating the cost of each type of delay by stakeholder recognizing that the airlines have different business strategies and therefore have different preferences. A case study based on a recent study of the benefits of the Integrated Terminal Weather Service (ITWS) demonstrates that a detailed investigation of the breakdown of delay into components can lead to more accurate delay cost accounting. Cheng-Lung Wu (2005) explored the inherent delays of airline schedules resulting from limited buffer times and stochastic disruptions in airline operations. It is found that significant gaps exist between the real operating delays, the inherent delays (from simulation) and the zero-delay scenario. Results show that airline schedules must consider the stochasticity in daily operations. Schedules may become robust and reliable, only if buffer times are embedded and designed properly in airline schedules.

Task 2: Adapt the frequency analysis technique for transportation data sets.

We implemented the algorithm for the Frequency Analysis in Matlab programming environment. (Codes, a user friendly GUI, and a user’s guide (to be included in Appendix

B) will be delivered on a CD-ROM with the final report.) Then, we worked on the adaptation of the frequency analysis method to the BTS data sets and, in particular, the airline arrival/departure time data sets. One of the motivations to use the airline data set is that this is a huge data set of all the flight arrival and departure times for the past seventeen years from all the major airports in the country. We extracted the part with Orlando International Airport either as the arrival port or the departure port. This data set is more promising for our method than other BTS data sets. Another main motivation for us to use this data set is due to the strong interests of FDOT on the tourist related transportation issues. Since most tourists are coming to Orlando via air transportation, their arrival will impact on local and surrounding traffic networks. There are several well known high seasons for the airlines. Our analysis may help us find out if there are additional patterns and how local traffic may be affected. We now briefly describe the adaptation of the frequency analysis method to some of the data sets discussed above.

Waterborne Commerce Data Set

We did some additional work on the waterborne commerce data set at the major Florida ports: Miami, Jacksonville, and Tampa, one at a time and found out that the data was collected but only summaries on aggregated data sets were available. This makes it very difficult for our frequency analysis method since we need the raw data. We decided not to pursue this data set any further.

Airline Data Set

As we mentioned above, this data set is the most promising in terms of the applicability of our method. This data contains daily records of actual departure times, arrival times, origins and destinations for non-stop domestic flights by major air carriers from 1988 to 2004. We obtained the data from the Bureau of Transportation Statistics (BTS) web site (http://www.transtats.bts.gov). We performed the following analysis on temporal repetitive patterns of tourist travels in Florida, and Orlando in particular. In the view of the lack of tourist data sets, any detected pattern will be useful to the tourist travel study. We started by concentrating on the Orlando international airport. Once we have a better understanding of the Orlando related data set, it is easy to do a similar study on other major cities in Florida. Our analysis is related to traffic pattern analysis.

1. Estimation of number of passengers. Since the data set does not contain any information about the number of passengers, we need to find a way to estimate it. We proposed to use the aircraft type and the current airline seats configuration as the basis for our estimation. Three major airlines were considered. They are American Airline, US Airway, and Delta Airline. This estimation is useful when we interpret our analysis.
2. Arrival time analysis. Since the data set is huge and the times are recorded within minute, we designed a matrix for some interesting experiments by using different time intervals and different time periods. For example, we did analysis on the number of arrival flights in day-by-day for ten years; week-by-week for ten years and three years. Many more refined time intervals were used and a lot set of outputs of analysis was generated for our next task when interpretation process starts.

   The main adaptation of the frequency analysis method to the analysis of the airline data sets is by adding a least square fit using the spline functions. A spline function is a piecewise polynomial connected together smoothly. They are more flexible than polynomials in modeling or fitting curves. A cubic spline is a piecewise cubic polynomial that has continuous second derivative. One modification of our method was introduced: a cubic spline least square fit is applied before we do the frequency analysis in order to eliminate the trend within the data set. Some illustration of the resulting representation of some of the data sets are given below:

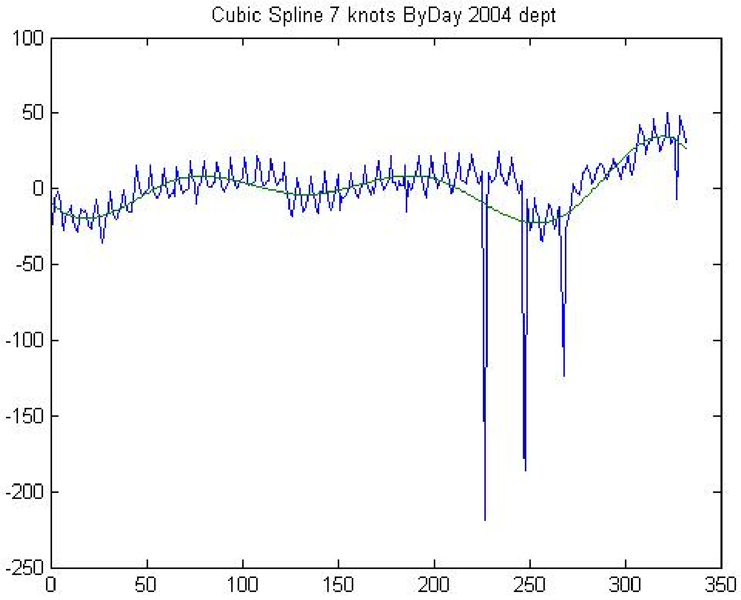


Figure 11. The cubic spline with 7 knots LS fit to the departure time (2004) by day

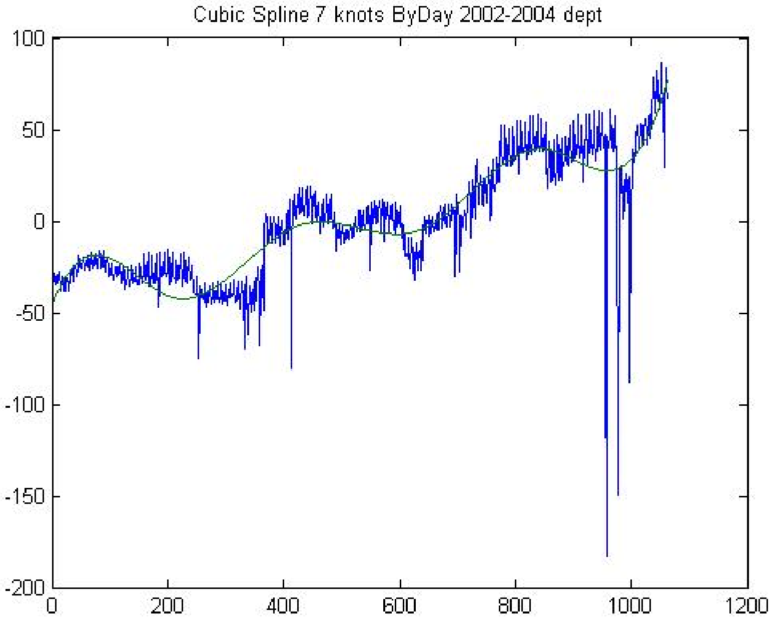


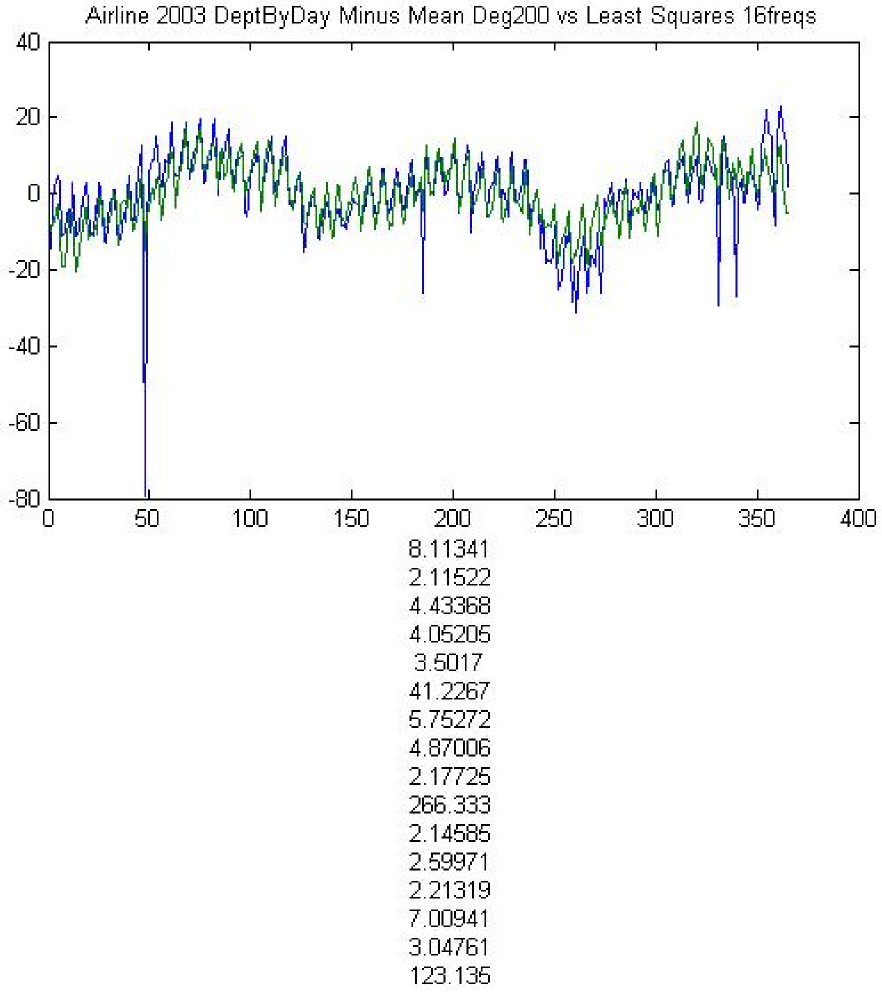
Figure 12.The cubic spline with 7 knots LS fit to the departure time (2002-2004) by day

Note that with only few knots for the cubic spline, we can model the trend very closely.

After the spline fitting, we take the difference of the data and the spline and do our frequency analysis on this difference. Here are some samples of our many outputs.

<u>Airline 2003 departure time subtracted by the mean (in blue) by 16 frequencies</u>

(The representing curve is in green. The most significant periods are listed below the plot with the most significant one at the bottom.)



```
8.11341
2.11522
4.43368
4.05205
 3.5017
41.2267
5.75272
4.87006
2.17725
266.333
2.14585
2.59971
2.21319
7.00941
3.04761
123.135
```

Figure 13. Airline 2003 departure time subtracted by the mean (in blue) by 16 frequencies

Airline 2002 departure time subtracted by the cubic spline with 11 knots (in blue) by 8 frequencies:

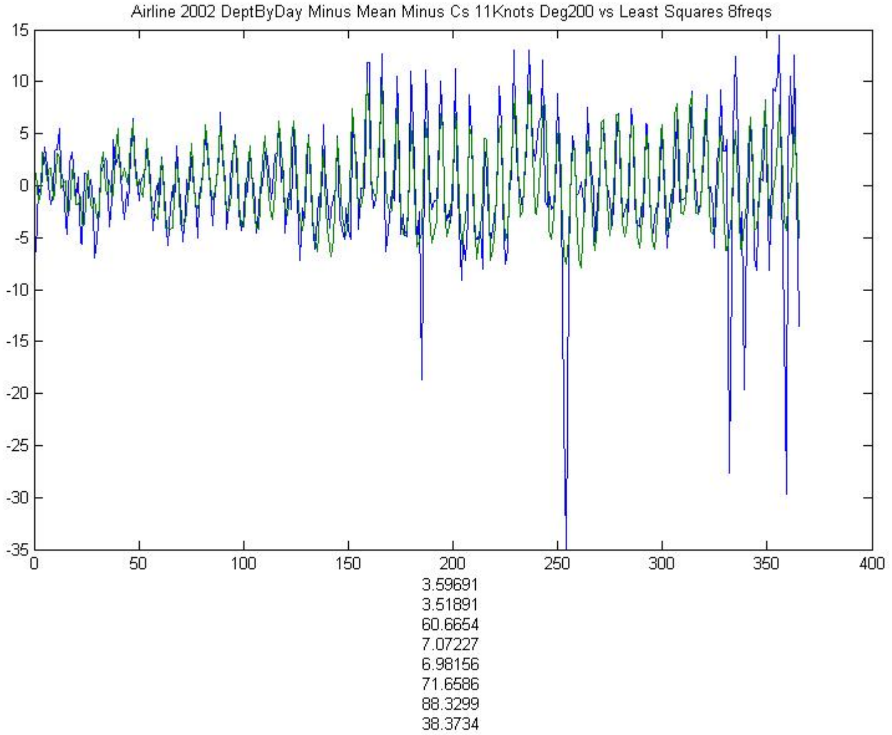


Figure 14. Airline 2002 departure time subtracted by the cubic spline with 11 knots (in blue) by 8 frequencies

Airline 2003 departure time subtracted by the mean (in blue) by 4 frequencies

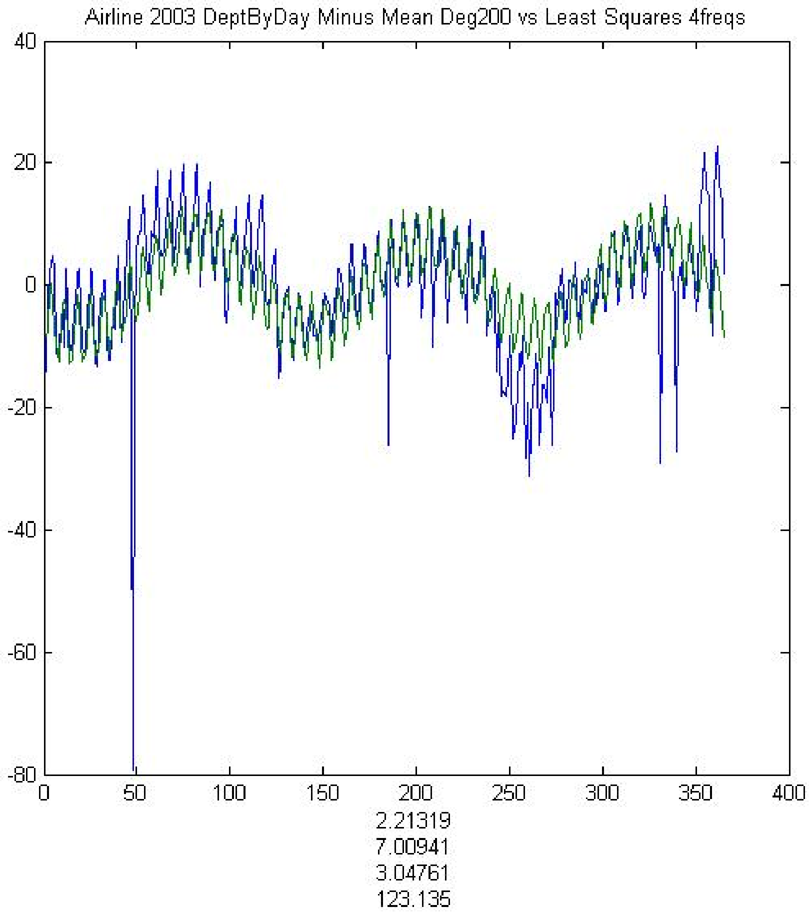


Figure 15. Airline 2003 departure time subtracted by the mean (in blue) by 4 frequencies

3. Delay time analysis. Comparing to the departure time, the delay time makes a better candidate for the study since the departure time was originally set by a human being while the delay time is due to many random causes. How to relate delays with their causes of weather, time, number of fights, original airports, airlines, etc. is a very interesting problem. We performed the similar analysis on the delay times as we did on the departure times. Many of the outputs of the frequency analysis for both departure times and arrival times are included in the attached CD (to be submitted with the final report).

Task 3: Develop numerical algorithms and related conceptual framework for data integration

In order to address the speed issue of the algorithm, we started to implement the more recently developed algorithms. For large data sets, the requirement on the performance of the algorithm will be high and real polynomial algorithms will be used in our Matlab implementation. To compare our frequency analysis method using orthogonal polynomials with the existing ones in the literature, we also started to implement other frequency estimation algorithms from statistics such as the algorithm named MUSIC (Multiple Signal Classification) in Matlab and package named PEST from time series analysis. These implementations will be applied on the same data sets to compare their performance.

More stable implementation of the algorithm have been studied to improve the numerical computational stability of our implementation. The stability of a numerical implementation of an algorithm measures how sensitive the numerical results are to the small change of the input data.

Standard statistical methods

For completeness and comparison, we used several standard statistical tools in our analysis. In particular, a tool named PEST from time series analysis (created by Brockwell and Davis from Colorado State University) was used to detect periodicities, and ARMA(p,q) parameters in our data sets. Another popular method was also applied from the Signal Processing Toolbox of Matlab. It is under the name MUSIC (with variations pmusic, rootmusic, ect.) in the Matlab toolbox. It does the frequency estimate via the MUSIC eigenvector method. Technical formulation and comparison results are reported below.

We made some experiments on comparisons of our frequency analysis method using orthogonal polynomials with other available methods (mainly from statistics) including: MUSIC and time series method as implemented in PEST program developed in the work of Blackwell and Davis (Brockwell, PJ and Davis, RA (1991). Time Series: Theory and Methods, 2nd ed., Springer-Verlag, New York.). We find that in most cases, our method is more straightforward and accurate and the run-times of these methods are comparable. Among the various numerical experiments, here's an example: Let us consider a digital signal (time series) given by

$$x(t)=2*\cos(t*\pi/6)+\cos(t*\sqrt{2})+0.5*\cos(t*2*\pi/3)$$

With our method of frequency analysis using polynomials orthogonal on the unit circle in the complex plane, we obtain the estimations of the frequencies $\pi/6$, $\sqrt{2}$, and $2\pi/3$ as 0.5239, 1.4281, and 2.1211 by using degree 8 (2 above the exact number of complex frequencies). A plot of the zeros of the orthogonal polynomial of degree 8 is given below:

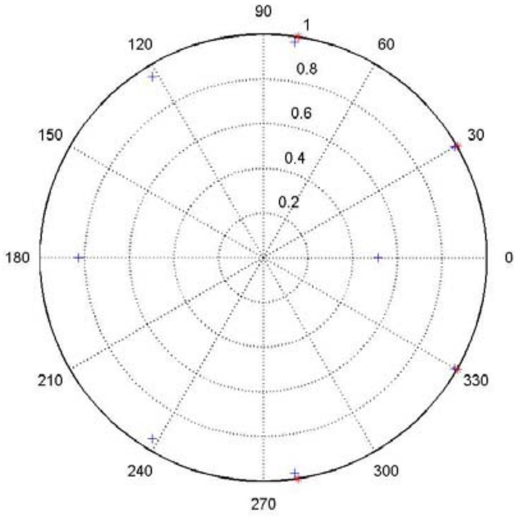


Figure 16. Zeros of orthogonal polynomial of degree 8

Using the MUSIC method, Matlab command pmusic gives us the following plot of the pseudospectrum:

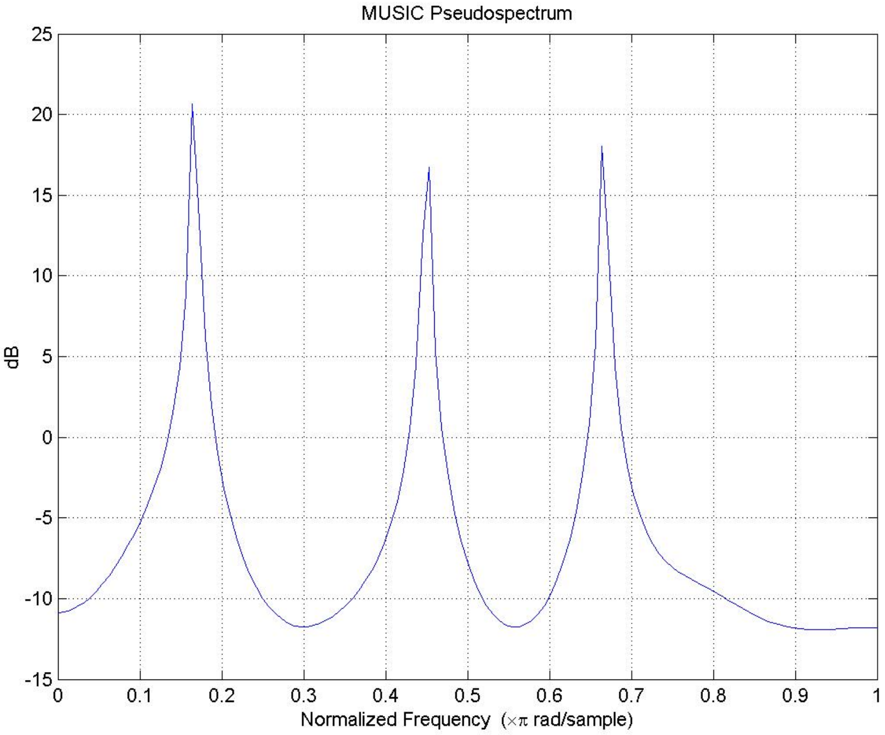


Figure 17. Pseudospectrum plot from MUSIC

From the plot, we may estimate the peaks which are corresponding to the estimations of the frequencies.

Finally, with the program PEST, we have to import the data into Excel and save it in the format for the program (CSV MS-DOS format, with extension .tsm). Then in the program, we specify the parameter of an AR model to 8 (again, intentionally 2 higher than the true number of frequencies). Here's a plot of the spectrums of the data and the estimation:

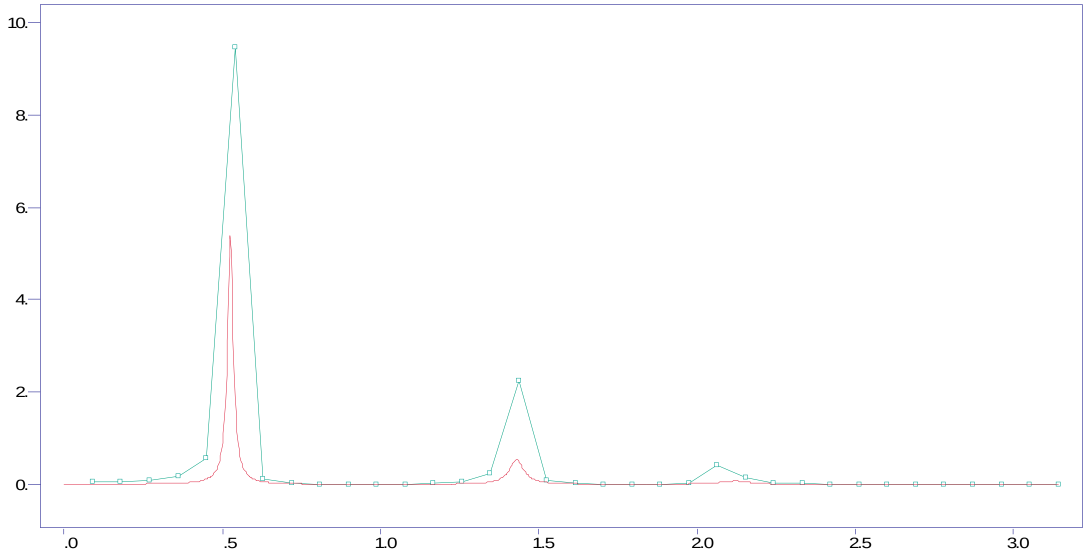


Figure 18. Spectrum plot from PEST

We found that in almost all cases with artificial data sets (when true frequencies are known), these methods are comparable in terms of their performance and accuracy. But when applied to real data sets, our method stands out in that we may choose the parameters to tailor our model to the data set so that we can pick the most significant frequencies.

Task 4: Integrate the numerical algorithms into GIS data sets

We explored the possibility of integration of Matlab and ArcView software.

i) **Linking via a third programming language**

ArcView has support within itself for writing programs in C or Visual Basic using Avenue. Avenue can use any COM compliant DLLs as library files. Matlab provides COM complaint DLLs that allow access to Matlab functionality. So, we can use either

Visual Basic or C to mediate between the two programs. We have not found any preexisting VB/C programs that perform the functions we need.

Judging by the way Avenue works, it seems that we could attach custom functions to any input type already supported by ArcView. Information regarding use of Matlab from VB:

http://www.dsprelated.com/groups/matlab/show/3903.php

Advantages: Very powerful. We should be able to accomplish everything we need using this technique.

Disadvantages: Requires dealing with an additional programming language. It will require a lot of extra overhead. It will require us to write much of the code from scratch.

ii) **Matlab Mapping Toolbox**

This toolbox allows ArcView objects to be accessed and edited within Matlab. It allows for modification of shapefiles. Mapping Toolbox homepage:

http://www.mathworks.com/products/mapping/

Advantages: Straightforward. It is probably easy to use. It is a well established product which is more likely to work properly and need not too much debugging.

Disadvantages: Somewhat limited. The toolbox does not allow for editing of all ArcView formats, and the user must work with only what the toolbox provides.

iii) **Using both applications simultaneously**

ESRI has developed an approach to groundwater modeling which uses both Matlab and GIS. Their approach involves switching between applications. Avenue in ArcView is used to format data for use in Matlab. Matlab .m files are then called to process the data and format it for use by ArcView. This approach requires both Avenue to access the appropriate .m files from ArcView, and the Mapping Toolbox to allow Matlab to access ArcView formats. ESRI Groundwater Project:

http://gis.esri.com/library/userconf/proc01/professional/papers/pap600/p600.htm

Advantages: Has been demonstrated to work. Will probably be flexible enough to accomplish whatever we need.

Disadvantages: Cumbersome in that it requires use of both applications. It requires use of Avenue and the Mapping Toolbox. This approach may be obsolete since it was designed for versions of ArcView that did not have the full COM support of later versions.

Task 5: Quantitative analysis

To help us visualize how close our detected patterns reflect the real data, we developed a procedure to show the least square fit using the detected frequencies to the real data. The quantitative analysis consists of the analysis of two delay models: average daily delay and single flight delay.

Average daily delay

Data description

The analysis is based on data from the Airline On-Time Performance Data from the Federal Aviation Administration (FAA) and the climatic data from the National Climatic Data Center (NCDC). Statistical models will be presented for the estimation of a vector of airport daily arrival delay or single flight arrival delay. These models are formulated using flight delay parameters and weather conditions at Orlando International Airport (MCO). The data used consists of the non-stop domestic flights on scheduled service by certificated carriers with the destination of MCO. This database contains departure delays and arrival delays for non-stop domestic flights by major air carriers, and provides such additional items as origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi out and taxi in times, air time, and non-stop distance. The flight data used is from 01/01/2002 through 12/31/2003 excluding the cancelled and diverted flights.

For frequency analysis, we formed a daily sequence where each data point was the average of the delays of all flights with a positive delay during the corresponding day. The mean was subtracted. Then, to test what patterns might be found at various timescales, we considered only the 1-320th days of the sequence. The Levinson algorithm was used to generate a polynomial of degree 50 to uncover the frequencies of periodic behavior in the sequence. The 8 strongest frequencies were then used to construct a model for the sequence using least squares, and the model and original sequence were plotted together.

This data set contained a seasonal pattern of 172 days (approximately half of a year), and a monthly pattern of 30.7 days. Here is an plot showing a seasonal pattern and a monthly pattern.

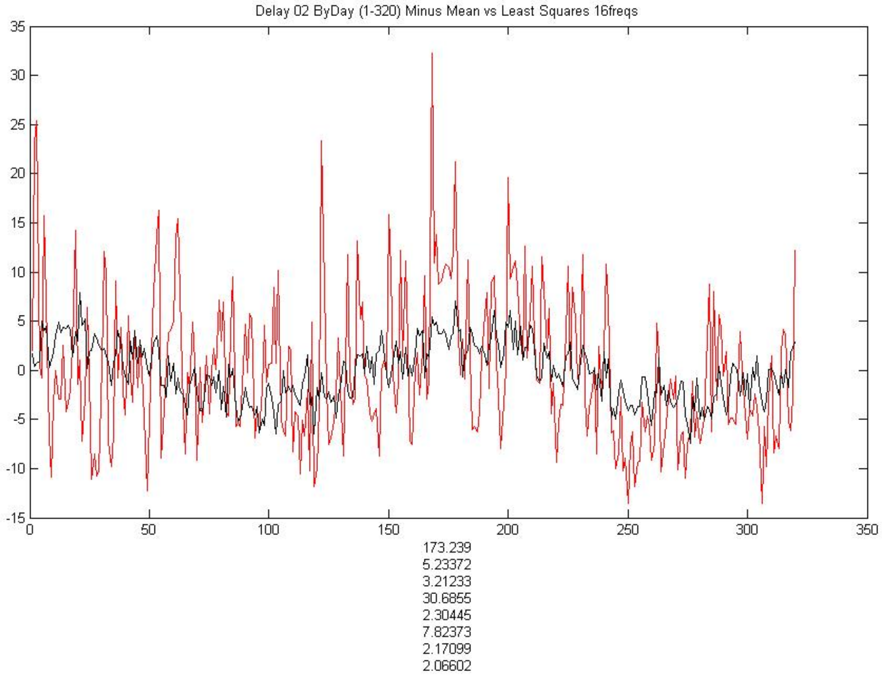


Figure 19. Seasonal (173 days) and monthly (30.7 days) patterns

We can also examine the daily and weekly pattern by considering morning, afternoon, and evening blocks. That is, we divide the time of a day into three blocks: the morning block covers 7:00am to 11:59 am, the afternoon block refers to 12:00pm to 5:59pm, and the evening block is from 6:00pm to 11:59pm. We intentionally omitted time period interval from 12:00am to 7:00am due to the fact that very few flights depart from or take off the airports we are considering. Therefore, every three blocks correspond to a day and every 21 blocks correspond to a week. In the following two plots, we show can observe clear daily patterns.

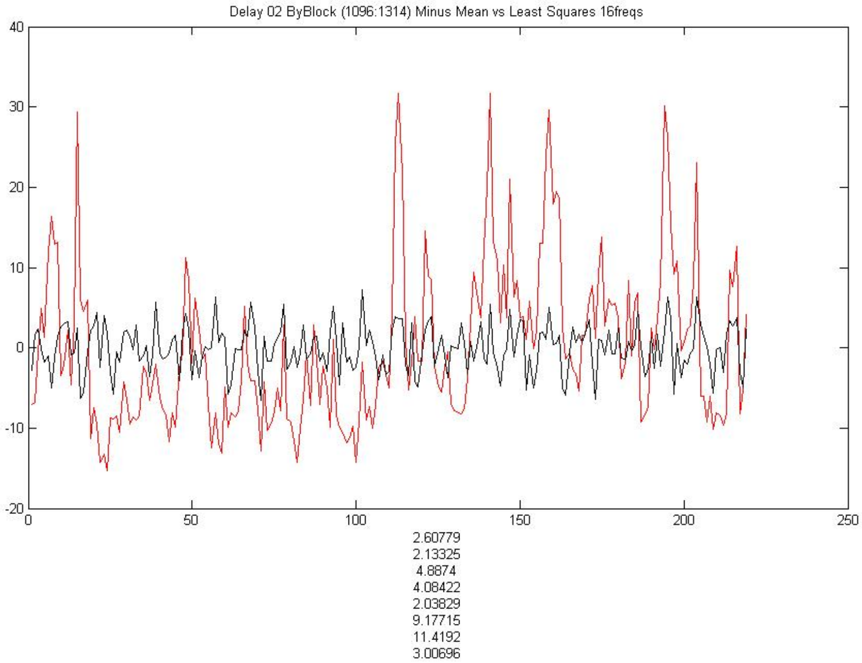


Figure 20. Daily pattern in morning, afternoon, and evening blocks (days 1-320)

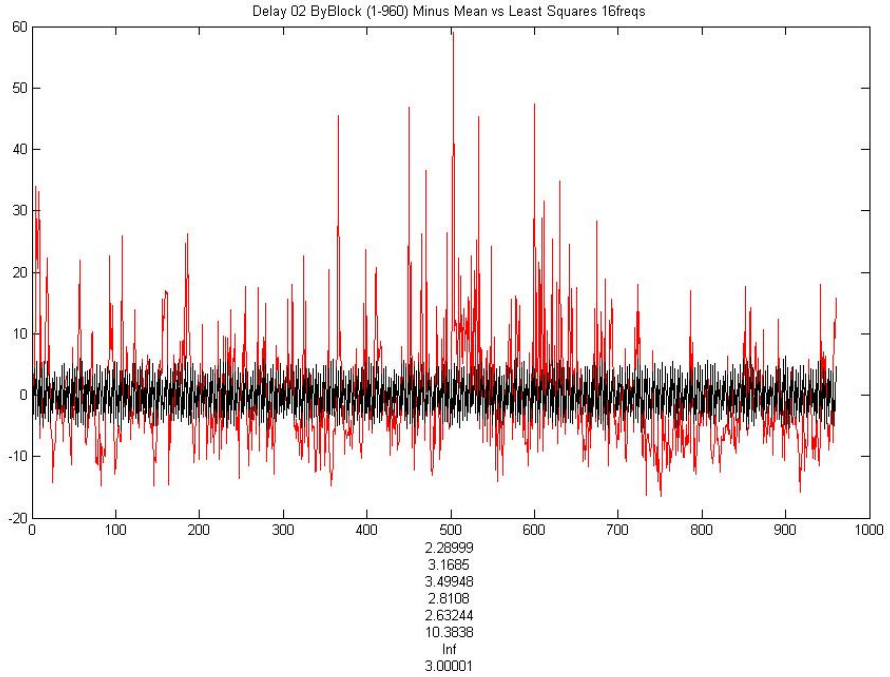


Figure 21. Daily pattern in morning, afternoon, and evening blocks (days 366-438)

Together with our analysis of the average daily delay using the frequency analysis method, we carried out a parallel analysis using more conventional method of analysis of variances and factor analysis via SAS in detecting patterns. The result is combined with the one obtained from the frequency analysis method. This approach provides a way to validate our methodology by using real data though our method has been validated with simulated data sets in our earlier completed tasks. We can get insight on our outputs with the usual tools and fills in the gaps between analysis and interpretation.

<u>Model variables</u>

(1) Average daily arrival delay

Arrival delay equals the difference of the actual arrival time minus the scheduled arrival time. The average daily delay metric reflects only positive delays; flights that arrive early are assigned zero delay in the calculation. For our delay metric, d(t), we used the average daily positive delay for all scheduled and completed flights from other airports to MCO airport in Orlando. It is the average of positive delay per flight per day. We collected daily data from the Inter-modal Transportation Database from 01/01/2002 through 12/31/2003.

(2) Maximum hourly flow rate

The airport capacity refers to the ability of the various facilities in the airport system in handling the aircrafts' activities in the airport. The critical factor of the capacity is the relationship between the demand and capacity and how the transportation system's service time is affected. As service time increases, system delay may increase and overall system reliability decreases.

The preferred measure of the capacity is the ultimate or saturation capacity, which gives the maximum number of aircrafts that, can be handled during a certain period under conditions of continuous demand. The hourly capacity is the maximum numbers of operation that can be handled in a one-hour period under specific operating conditions, in particular, weather conditions (ceiling and visibility), air traffic control, the aircraft mix and the nature of operations. In this work, the airport capacity is represented by the maximum hourly capacity in one day (including arrival and departure flights) according to the actual departure time.

(3) Arrival demand

The arrival demand is included as another variable that may capture the incidence of congestion in the airport system. The arrival demand vector is represented by the sum of completed arrival flights to MCO per day according to the actual arrival time.

(4) Flight duration

The flight time contributes to the flight delay to some extent. It is the airborne time for each flight.

(5) Space of Inter-arrival time

The space here means the intervals between two consecutive arriving flights. This inter-arrival time is calculated according to the scheduled time. The aim of space is to find the relationship of delay and the schedule operation of the airport.

(6) Airport precipitation

We control for adverse weather using information about the amount of rainfall for every day at MCO. The database contains daily observations about the inches of the rainfall indicating whether the station had heavy rainfalls during that day.

(7) Airport wind speed

Another typical factor can also have a significant impact on flights. Strong low-level winds or wind shear may require that planes are spaced farther apart. Strong crosswinds may make some runways unusable.

The arrival delay can thus be affected by windy conditions, either because of the direct effect of the wind or because of associated conditions such as wind shear. We thus include in our model the daily average wind speed at Orlando International Airport.

(8) Seasonal variables

Seasonal effects are captured using a set of dummy variables. The seasonal variables indicate the season of the observation day. The season has four classes, March-May, June-August, September to November, and December to February periods, which represent the four seasons pattern in Orlando. So the model includes 3 seasonal dummies. These variables capture the seasonal changes in the flight delay.

(9) Weekly variables

Weekly effects are captured using a set of dummy variables. The weekly variables indicate the weekday of the observation day. The weekly pattern is a hidden pattern and is decided by the model.

(10) Origin airport regional variables

There are 82 airports that have direct flights to Orlando. Here the areas of the origin airports are divided into four parts that are southeast, southwest, northeast and northwest (appendix A). And the daily flight number from each site to MCO is calculated as a variable, which indicated the location effect on the flight. So this group includes 4 variables, southeast, southwest, northeast and northwest.

(11) Interaction effects

We also investigated the interaction effects between up variables.

Single delay analysis

We also performed the analysis on *single flight delay* patterns based on the Airline On-Time Performance Data in the airline arrival/departure time data sets from BTS for Florida airports and Orlando International Airport (MCO). Major outputs are included below:

Data description

From the Airline On-Time Performance Data, the airline's on-time performance in MCO—the proportion of flights arriving within 15 minutes of schedule, for 2004 was 79.85 percent, down from 83.16 percent in 2003 and 82.55 percent in 2002. These delays are frustrating to air travelers and costly to airlines. They are also the concern in this research. What is the pattern of the delay? What is the contribution of various causal factors? Can we predict the delay with the flight schedule?

Our analysis is concentrated on the effect of the characteristic of flights on the on-time performance of arrival flights. A binary logistic regression is formulated including the seasonal effect, hourly arrival demand, regional effect, time effect, effect of flight distance, the effect of specific events, and the interaction.

Methodology

To introduce the factors into statistical model and test their main effects on delay of flights, the flights with arrival delay >=0 from other airports to MCO are identified, which are categorized into two groups: no delay (0<=arrival delay<15), delayed (15<=arrival delay).

Those factors include information of delayed flights, as well as the corresponding airport conditions. So that the dependent variable Y (delayed flights) here takes on two values: Y = 0 for no-delayed flights (with less than 15 minutes delayed), Y = 1 for delayed flights (with more 15 minutes delayed). The no-delayed flights and delayed flights respectively represent 61.42% and 38.48% of the whole data set.

Binary logistic regression is proper to use here when the dependent is a dichotomy (an event happened or not) and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects.

Data and variables

The model was estimated on a data set consisting of domestic flights with the destination of MCO. The data for the non-stop flights on scheduled service by certificated carriers to MCO were obtained from the Airline On-Time Performance Data on the Bureau of Transportation Systems website. We collected daily flight data with positive arrival delay from 01/01/2002 through 12/31/2003 excluding the cancelled and diverted flights. The factors are introduced as the same as delay model in section 3. Some important category variables are introduced as below:

The time effect is defined by a set of dummy variables. These variables indicate the scheduled arriving time of each delayed flights. Here we classify the scheduled arrival time into 3 classes: morning, afternoon and evening. They are 7am to 11:59 am, 12 am to 4:59 pm, and 5 pm to 11:59 pm. So the model includes 2 dummies to capture the scheduled arrival time of each delayed flight.

Seasonal effects are captured by a set of dummy variables. The seasonal variables indicate the seasons when the flights are scheduled. The season has four classes, spring (March-May), summer (June-August), fall (September to November), and winter (December to February), which represent the four season pattern in Orlando. So the model includes 3 seasonal dummies.

The regional effects are captured by a set of dummy variables, south, east, central, and west areas. So we have 3 regional dummies. The western states include Alaska, Arizona, California, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming. The central states include Arkansas, Colorado, Illinois, Iowa, Kansas, Louisiana, Minnesota, Missouri, Nebraska, North Dakota, Oklahoma, South Dakota, Texas, and Wisconsin. The eastern states include Connecticut, Delaware, District of Columbia, Indiana, Kentucky, Maine, Maryland, Massachusetts, Michigan, New Hampshire, New Jersey, New York, Ohio, Pennsylvania, Rhode Island, Vermont, Virginia, and West Virginia. The southern states include Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, and Tennessee.

The arrival demand is included as another variable that may capture the incidence of congestion of the airport in the hour when the cancelled flight occurs. The arrival demand vector is represented by the sum of completed arrival flights to MCO in the hour when the flight occurs according to the scheduled arrival time.

Airport wind speed and precipitation in Orlando international airport are another two weather variables from National Climatic Data Center (NCDC). The precipitation is hundredth inch of rainfall per hour, and the wind speed is speed of wind in mph per day.

The effect of flight distance is captured by the categories of the flight distance, which respectively represent the flight distance of 0 to 750 miles, 750 to 1000miles and greater than 1000 miles. The distance classes are categorized by the same frequency.

The space is used here another variable, which means the intervals between two consecutive arriving flights. This inter-arrival time is calculated the time between each flight and the before flight according to the schedule arriving time. The aim of space is to find the relationship of delay and the schedule operation of the airport. The intervals between arrival flights differ under different weather conditions, runway conditions and operating conditions. If the flight flows arrive smoothly, we can say that the space is constant and small which means the waiting and service time for each flight is controlled at a low level.

Task 6. Interpretation and recommendation

The interpretation of the results presented most challenges. Our approach can be summarized as a two-stage strategy in which we first use the frequency analysis method to do a preliminary and quick detection of possible periodicities in the data set. Then from the detected periodicities, in order to explain the cause for them, we try to come up with a list of possible variables having similar periodicities and do a detailed regression analysis. The results from the regression analysis can be used to explain the possible cause for the detected periodicities. Some details are presented in the draft of the paper included in Appendix A.

Average daily delay model analysis

i) The multiple regression methodology was initially attempted. We assume that the errors are normally, identically, and independently distributed. Initial experimentation revealed that these assumptions do not hold. In particular we found that the errors are not normal. About the random error ε, the assumption that ε is normally distributed is the least restrictive when we apply regression analysis in practice. When non-normality of the random error term is detected, it can often be rectified by applying the transformations.

After some experimentation a model is developed for the average daily delay. The R square values for this linear model is only 0.2755, which is not satisfactory. And different transformations of dependent variable are tried including, but the R square can’t be improved. So some other statistics methods will be used to analyze the daily delay, such as ANOVA and logistic regression.

ii) Analysis of variance (ANOVA) is used to study the effects of one or more independent (predictor) variables on the dependent (response) variable. Most commonly, ANOVA is used to test the equality of means by analyzing the total sum of squares (about the combined mean), which is partitioned into different components (due to model or due to random error).

The week differences are proved by F-test to be significant (p<0.0001). LSD test and TUKEY test both proved that the daily delay on Thursday and Friday are obviously higher than other weekdays. Tuesday and Saturday have the lowest daily delay in the week. This pattern should be related with the weekly schedule of the airport.

At the same time, the season differences are proved by F-test to be significant (p<0.0001). LSD test and TUKEY test both proved that in summer the daily delay is obviously higher than other three seasons, and in fall the daily delay is obviously lower than other three seasons. In spring the daily delay is lower than in winter. This pattern should be related with several reasons. In Orlando summer is a rainy season. The thunder may cause the interrupt of the operations of the airport, which increase the delay of the flights that schedule arrival times are during or directly after the bad weather.

iii) Logistic regression belongs to the group of regression methods for describing the relationship between explanatory variables and a discrete response variable. A logistic regression is proper to use when the dependent is categorized and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects.

The dependent variable (average daily arrival delay) can take on three values: Y = 0 for delay<5 min; Y=1 for delay >=5 and < 10 min; Y=2 for delay>=10min. They are created by the bins with approximately equal frequencies. The independent variables are the same as the linear model of daily arrival delay. Considering the week pattern of the daily delay, the week is classified into five levels: Monday and Friday have no significant difference and are combined into one level. Tuesday and Saturday are combined into one level. Wednesday, Thursday and Friday are individually one level.

Here we treat the arrival delay as a categorical outcome with three levels and keep the nature ordering presented in the data. There are usually three different ways of generalizing the logit model to handle ordered categories. We will use the Proportional Odds Model (Cumulative Logit Model). For this model, we have actually imposed the restriction that the regression parameters except the intercepts are the same for the two logit models. It implies that it doesn't make any difference how we categorize the dependent variable - the effects of the explanatory covariates are always the same. The result of modeling is showed below in table 1.

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr>ChiSq
Intercept 3	1	-1.4073	0.1974	50.8256	<.0001
Intercept 2	1	0.8087	0.1927	17.6189	<.0001
precipitation	1	0.0264	0.00364	52.4524	<.0001
Monday and Sunday	1	0.4423	0.1935	5.2256	0.0223
Wednesday	1	0.4659	0.2355	3.9136	0.0479
Thursday	1	0.9280	0.2402	14.9235	0.0001
Friday	1	1.3301	0.2457	29.3148	<.0001
spring	1	-0.2356	0.2026	0.2026	1.3520
summer	1	0.3872	0.2161	3.2104	0.0732
fall	1	-1.4521	0.2113	47.2132	<.0001

Effect	Point estimate	95% Wald Confidence Limits	
precipitation	1.027	1.019	1.034
Monday and Sunday vs Tuesday and Saturday	1.556	1.065	2.274
Wednesday vs Tuesday and Saturday	1.593	1.004	2.528
Thursday vs Tuesday and Saturday	2.530	1.580	4.051
Friday vs Tuesday and Saturday	3.781	2.336	6.120
spring vs winter	0.790	0.531	1.175
Summer vs winter	1.473	0.964	2.250
Fall vs winter	0.234	0.155	0.354

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.0	Somers' D	0.504
Percent Discordant	23.6	Gamma	0.516
Percent Tied	2.4	Tau-a	0.330
Pairs	174384	c	0.752

Table 1. Results of logistic regression of the daily arrival delay

From table 1, test statistic for the Proportional Odds Assumption is 8.4062 with the DF of 8, so the p value is 0.3948. The high p-value is desirable. For this problem, we find no reason to reject the proportional odds model.

Adjusting for other variables, the odds of having a higher arrival delay in spring will be 0.79 times of the odds in winter, the odds of having a higher arrival delay in summer will be 1.473 times of the odds in winter, and the odds of having a higher arrival delay in fall will be 0.234 times of the odds in winter.

On Monday and Sunday the odds of having a higher arrival delay will be 1.556 times of the odds in Tuesday and Saturday, on Wednesday the odds of having a higher arrival delay will be 1.593 times of the odds in Tuesday and Saturday, on Thursday the odds of having a higher arrival delay will be 2.530 times of the odds in Tuesday and Saturday, and on Friday the odds of having a higher arrival delay will be 1.556 times of the odds in Tuesday and Saturday. From the odds ratio, on Thursday and Friday the airport is showed to have the higher probability to have delay more than 10 minutes.

For each 10*0.01=0.1 inch increase with the precipitation, the odds of having more arrival delay increases by exp(0.0264*10)-1= 30.2%. There are no significant interactions.

Compared with linear regression model, the multiple logistic regression model shows the same seasonal pattern and weekly pattern.

The daily delay on Thursday and Friday are obviously higher than other weekdays. Tuesday and Saturday have the lowest daily delay in the week. In summer the daily delay is obviously higher than other three seasons, and in fall the daily delay is obviously lower than other three seasons. In spring the daily delay is lower than in winter. At the same time the variable of precipitation contributes to the delay

Single flight delay results of the model of low and high delay

Parameter	Definition
Thursday and Friday	The flight takes place on Thursday and Friday
summer	The flight takes place in summer
winter	The flight takes place in winter
Fall	The flight takes place in fall
evening	The flight takes place between 5pm to 11:59pm
afternoon	The flight takes place between 12pm to 4:59pm
distance between 750 and 1000	The flight distance is in between 750 and 1000 miles
distance > 1000	The flight distance is larger than 1000 miles
precipitation	Hundredth of inches of the precipitation per day
log_space	The log transform of the space

Table 2. Definition of delay model independent variables

The final logit model is presented in Table 3 and 4 and 5.

Parameter	Estimate	StandardError	Wald Chi-Square	Pr > ChiSq
Intercept	-1.0827	0.0287	1423.1896	<.0001
winter	0.1432	0.0215	44.5419	<.0001
Fall	-0.2595	0.0233	123.7075	<.0001
summer	0.2191	0.0211	107.7185	<.0001
Evening	0.7179	0.0210	1166.2651	<.0001
Afternoon	0.2889	0.0226	163.9135	<.0001
Thursday and Friday	0.1172	0.0164	50.8101	<.0001
distance > 1000 miles	-0.0338	0.0214	2.4898	0.1146
distance between 750 and 1000 miles	0.1921	0.0182	111.4865	<.0001
precipitation	0.0226	0.00111	414.5916	<.0001
Log_space	-0.0198	0.00905	4.7965	0.0285

Table 3. Model estimation for significant independent variables

Effect	Point Estimate	95%Wald Confidence Limits	
Winter vs spring	1.154	1.106	1.203
Fall vs spring	0.771	0.737	0.808
Summer vs spring	1.245	1.194	1.297
Evening vs moring	2.050	1.967	2.136
Afternoon vs moring	1.335	1.277	1.395
Thursday and Friday vs other weekdays	1.124	1.089	1.161
distance>1000vs distance<750	0.967	0.927	1.008
distance in[750,1000]vs distance<750	1.212	1.169	1.256
precipitation	1.023	1.021	1.025
log_space	0.980	0.963	0.998

Table 4. Odds Ratio Estimates

Criterion	Intercept Only	intercept & Covariates
AIC	99612.098	96759.608
SC	99621.319	96861.041
-2 Log L	99610.098	96737.608

Testing Global Null Hypothesis: BETA=0			
Likelihood Ratio	2872.4895	10	<.0001
Score	2785.6663	10	<.0001
ald	2553.7276	10	<.0001
Likelihood Ratio	2872.4895	10	<.0001

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	60.9	Somers' D	0.224
Percent Discordant	38.4	Gamma	0.226
Percent Tied	0.7	Tau-a	0.106

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
17.4662	8	0.1226

Table 5. Model Fit Statistics

From Table 5, Hosmer and Lemeshow Goodness-of-Fit Test statistic is 17.4662 with the DF of 8. The resulting p value of 0.1226, shown in Table 5, suggests that the model fits well. We find no reason to reject the odds model at 5% confidence level.

Four factors including season influences, time influences, distance influences, precipitation and space show significant association with the likelihood of flights of delay.

In the ‘odds ratio’ column in table 4, which are obtained from the parameter estimates. For the dummy variable ‘summer’, which indicates the flight takes place in summer, the odds ratio is 1.245. The odds ratio of 1.245 tells us that the predicted odds of ‘delay flights’ for summer are 1.245 times the odds for other seasons. In other words, the predicted odds of delay are about 24.5% higher when the season is summer. The dummy variable ‘winter’ indicates the flight takes place in winter, the odds ratio is 1.154. This implies that the predicted odds of delay are about 15.4% higher when the season is winter. The dummy variable ‘fall’ indicates the flight takes place in fall, the odds ratio is

0.771. This implies that the predicted odds of delay are about 23% lower in winter than other seasons.

The odds ratios for the arrival time show the relative ratios of high delay between different times (morning, afternoon, an evening) for each flight. Compared to morning, the odds of delay in the afternoon could be 1.335 times higher and the odds in the evening could be 2.05 times higher. At evening the flights are much more likely to be delayed than in morning. The results testified that the time will definitely contribute to delay.

The odds ratio of the distance variables is interesting. For each flight with the flight distance between 750 to 1000 miles, the odds of having arrival delay will increase by 1.212-1=21.2%. While for the flight with the flight distance larger than 1000 miles, the odds of having arrival delay will decrease 1-0.967=3.3% than the flight with the flight distance less than 750 miles.

The odds ratio for the precipitation shows that for each 10*0.01=0.1 inch increase with the precipitation, the odds of having more arrival delay increases by exp(0.0226*10)-1= 25.3%.

The odds ratio for the log space shows that as the space increases, the probability of the flights being delayed will decrease. Since the space is the inter-arrival time for two successive flights, when the space increases, the service time for each flight will increase, so that the efficiency of the operation will decline.

References:

1 Abdelghanya, K. F. et al., (2004) A model for projecting flight delays during irregular operation conditions, Journal of Air Transport Management, Volume 10, Issue 6, November 2004, Pages 385-394

2 Aisling, J.R., and J.B.,Kenneth, (1999) An assessment of the capacity and congestion levels at European airports, Journal of Air Transport Management

3 Allan, S.S., S.G. Gaddy, and J.E. Evans, (2001) Delay Causality and Reduction at the New York City Airports Using Terminal Weather Information, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, Lexington, Massachusetts

4 Bureau of Transportation Statistics, Airline On-Time Statistic. U.S. Department of Transportation. Washington, D.C. http://www.bts.gov/programs/airline_information.

5 C. F. Bracciali; Xin Li; A. Sri Ranga, Real orthogonal polynomials in frequency analysis, Math. Comp. **74** (2005), 341-362.

6 Hansen, M., and C. Y. Hsiao (2005), Going South? An Econometric Analysis of US Airline Flight Delays from 2000 to 2004, TRB 2005-2762

7 Hansen, M., and D. Peterman, (2004) Throughput Impacts of Time-based Metering at Los Angeles International Airport, TRB 2004-003212

8 Hansen, M., et al. (1998) Empirical Analysis of Airport Capacity Enhancement Impacts: A Case Study of DFW Airport,

9 Hansen, M., (2002) Micro-level analysis of airport delay externalities using deterministic queuing models: a case study, Journal of Air Transport Management Volume 8, Issue 2 , March 2002, Pages 73-87

10 Hansen, M., Y. Zhang, (2005) Operational Consequences of Alternative Airport Demand Management Policies: The Case of LaGuardia Airport, Transportation Research Part E

11 Janic, M., (2003) Large-schale Disruption of an Airline Network: a Model for Assessment of the Economic Consequences, TRB 2003-000471

12 Kanafani, A., M. R. Ohsfeldt, and W. J. Dunlay, (2004) Comprehensive Evaluation of Investments – A new method of evaluating impacts of technologies in air traffic management, TRB 2004-002684

13 L. Daruis, O. Njåstad and W. Van Assche, Para-orthogonal polynomials in frequency analysis, *Rocky Mountain J. Math.*, 33 (2003), 629-645.

14 Mehndiratta, S.R., M. Kiefer, and G. C. Eads, (2002) Analyzing the Impact of Slot Controls: The Case of San Francisco International Airport, TRB 2002-00055

15 Mueller, E.R. and G. B. Chatterji, (2002) Analysis of Aircraft Arrival and Departure Delay Characteristics, AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical

16 Rosen, A. (2002), Flight Delays on US Airlines: The Impact of Congestion Externalities in Hub and Spoke Networks, Department of Economics, Stanford University

17 Schaefer, L., and D. Miller, (2001) Flight Delay Propagation Analysis with the Detailed Policy Assessment Tool, Proceedings of the 2001 IEEE Systems, Man, and Cybernetics Conference

18 Suzuki, Y., (2000), The relationship between on-time performance and airline market share: a new approach, Transportation Research Part E 36 (2000) 139-154

19 Wang, P. T., C. R. Wanke, F. P. Wieland, (2004) Modeling Time and Space Metering of Flights in the National Airspace System, Proceedings of the 2004 Winter Simulation Conference

20 W.B. Jones, O. Njåstad and E.B. Saff, Szego polynomials associated with Wiener-Levinson filters, *J. Comput. Appl. Math.*, 32 (1990), 387-407.

21 Wu, C. (2005), Inherent delays and operational reliability of airline schedules, Journal of Air Transport Management Volume 11, Issue 4 , July 2005, Pages 273-282

22 Yan, S., C.Y. Shieh and M. Chen, (2002) A simulation framework for evaluating airport gate assignments, Transportation Research Part A: Policy and Practice Volume 36, Issue 10 , December 2002, Pages 885-898

Appendix A: A draft of a research paper to be submitted for publication

A Two-stage Approach to Identify Flight Delay Patterns

1 Introduction

Many major airports around the world have significant delay problems as a result of an imbalance between capacity and demand. Flight delay is a complex phenomenon, because a flight can be out of schedule due to problems at the original airport, at the destination airport, or during the airborne. A combination of these factors often occurs. Delays can sometimes also be attributable to airlines. Some flights are affected by reactionary delays, due to late arrival of previous flight. These reactionary delays can be aggravated by the schedule operation.

On-time performance of airlines schedule is key factor in maintaining current customer satisfaction and attracting new ones. Flight schedules are often subjected to irregularity. Due to the tight connection among airlines resources, delays could dramatically propagate over time and space unless the proper recovery actions are taken. Even if complex, flight delays are nowadays measurable. And there exist some pattern of flight delay due to the schedule performance and airline itself. Some results extracted from the case study on Orlando International Airport (MCO) can help to understand better the phenomenon

This paper addresses a two-stage approach including mathematic method and statistical models, which are developed and used to help one airport evaluate its on-time arrival performance and finding the pattern of flight delay. Using the method, the airport is able to examine the trend of flight delay and make effective strategies for maintaining high on-time performance.

The remainder of this paper is organized as follows. Section 2 is the literature review. Section 3 presents our methodology, including model specification, data sources, and estimation procedures. Estimation results are presented and discussed in Section 4. Conclusions and suggestions for further research are presented in Section 5.

2 Literature Review

The increase in delays in the National Airspace System (NAS) has been the subject of studies in recent years. The Federal Aviation Administration (FAA) describes the increase in delays and cancellations from 1995 through 1999. They found that the current system for collecting causal data does not provide the appropriate data for developing strong conclusions for delay causes and recommend changes to the current data collection system [17].

Allan et al. (2001) examined delays at New York City Airports from September 1998 through August 2000 to determine major causes of delay that occurred during the first year of an Integrated Terminal Weather System (ITWS) use and delays that occurred with ITWS in operation that were “avoidable” if enhanced weather detection. The methodology used in the study has considered major causes of delays (convective weather inside and well outside the terminal area, and high winds) that have generally been ignored in previous studies of capacity constrained airports such as Newark International Airport (EWR). The research found that the usual paradigm of assessing delays only in terms of Instrument Meteorological

Conditions (IMC) and Visual Meteorological Conditions (VMC) and the associated airport capacities is far too simplistic as a tool for determining which air traffic management investments best reduces the "avoidable" delays.

Schaefer and Miller (2001) use the Detailed Policy Assessment Tool (DPAT) to model the propagation of delay throughout a system of airports and sectors. To estimate delays, throughputs, and air traffic congestion in a typical scenario of current operations in the U. S., DPAT models the flow of approximately 50,000 flights per day throughout the airports and airspace of the U. S. National Airspace System (NAS) and can simulate flights to analyze delays at airports around the world. They obtained results for local flight departure and arrival delays due to IMC, propagation for IMC, comparisons to VMC results, and a comparison of propagated delays to entire system.

Shomik et al. (2002) presents an analysis of the possible impact of the application of slot controls as a demand management measure at San Francisco International Airport (SFO). A deterministic queuing model that uses an actual arrival schedule as input and simulates arrival delay based on available arrival capacity is used to estimate delay reduction potential of slot controls. The conclusions show the overall potential of slot controls to alleviate delay at SFO and their non-delay consequences.

Mehndiratta et al. (2002) propose a simulation framework, that is not only able to analyze the effects of stochastic flight delays on static gate assignments, but can also evaluate flexible buffer times and real-time gate assignment rules. The results of testing the framework on actual Chiang Kai-Shek airport (Taiwan) operations were good, showing that the framework could be useful for airport authorities to perform gate assignments.

Hansen (2002) analyzes runway delay externalities at Los Angeles International Airport (LAX) using a deterministic queuing model. The model allows estimating the delay impact of each specific arriving flight on each other specific arriving flight. The research finds that, despite being only moderately congested (average queuing delay only 4 min per arriving flight), individual flights can generate as much as 3 aircraft-hours of external delay impact on other flights, with an average impact of 26 aircraft-minutes and 3400 seat-minutes. About 90 percent of this impact is external to the airline as well as the flight, a consequence of the lack of a dominant airline at LAX.

Rosen (2002) measures the change in flight times resulting from infrastructure-constant changes in passenger demand. Results indicate that delays rise with the ratio of demand to fixed airport infrastructure, decreasing average flight times by close to seven minutes after the sharp decrease in demand in the fall of 2001. Flight time differences between the airlines in the sample are small, though the larger United had shorter average flight times in the winter quarter than America West, the smaller airline in the data sample.

Janic(2003) presents a model for assessment of the economic consequences of large-scale disruptions of an airline single hub-and-spoke network expressed by the costs of delayed and cancelled complexes of flights. The model uses the scheduled and affected service time of particular complexes to determine their delays caused by disruption.

During the last decade, a considerable attention has been given to proactive schedule recovery models as a possible approach to limit flight delays associated with Ground Delay Programs (GDP) (Abdelghany et al., 2004; Clarke, 1997). In these models, the impact of any reported flight delays, due to GDP or any other reason, is propagated in the network to determine any possible down-line disruptions. (Monroe and Chu, 1995).

Suzuki (2000) proposes a new method of modeling the relationship between on-time performance and market share in the airline industry. The idea behind the method is that the passengers decision to remain (use same airline) or switch (use other airlines) at time t depends on whether they have experienced flight delays at time t-1 or not.

Abdelghanya et al. (2004) present a flight delay projection model, which projects flight delays and alerts for down-line operation breaks for large-scale airlines schedules. The results show that down-line schedule disruptions are proportional to the number of flights impacted by the GDP. Furthermore, in the recorded GDP instances, aircraft appears to be the reason for most flight delays predicted by the model.

Hansen and Hsiao(2005) analyze the recent increase in flight delay in the US domestic system by estimating an econometric model of average daily delay that incorporates the effects of arrival queuing, convective weather, terminal weather conditions, seasonal effects, and secular effects. Results suggest that, controlling for these factors, delays decreased steadily from 2000 through early 2003, but that the trend reversed thereafter. The results identify key sources of recent increases in delay.

Hansen and Zhang (2005) investigated the interaction between LaGuardia Airport (LGA) and the rest of the aviation system by estimating simultaneous equations of average LGA and National Airspace System delay using two-stage least squares. The results demonstrate that the arrival delay impact of the Aviation Investment and Reform Act for the 21st Century (AIR-21) on LGA was in the form of increased Ground Delay Program (GDP) holding, and that while delay increased markedly under AIR-21 there were also observable improvements in the ability of LGA airport to handle traffic.

Mark Hansen and Dale Peterman (2004) uses censored regression to analyzes the delay impacts of the implementation of Traffic Management Advisor (TMA) metering at Los Angeles International Airport (LAX) in order to assess whether and how they have affected NAS performance.

The current method of valuing delay in benefit-cost analysis is insufficient for determining the distributional impacts of a technology change on users because it fails to account for the shifts in where benefits occur and to which users. Adib, Melissa and William(2004) proposes a theoretical framework for evaluating the distributive effects of technology changes that requires a new approach to the evaluation of delay and understanding efficiency in light of the state of the system. The framework defines different categories of delay per flight and a method for calculating the cost of each type of delay by stakeholder recognizing that the airlines have different business strategies and therefore have different preferences. A case study based on a recent study of the benefits of the Integrated Terminal Weather Service (ITWS) demonstrates that a detailed investigation of the breakdown of delay into components can lead to more accurate delay cost accounting.

Wu (2005) explores the inherent delays of airline schedules resulting from limited buffer times and stochastic disruptions in airline operations. It is found that significant gaps exist between the real operating delays, the inherent delays (from simulation) and the zero-delay scenario. Results show that airline schedules must consider the stochasticity in daily operations. Schedules may become robust and reliable, only if buffer times are embedded and designed properly in airline schedules.

The analysis of delays in the NAS has been the subject of several studies in recent years. These reports contain delay statistics over the entire NAS along with some data specific to individual airports. Statistical models and simulation method are used, including deterministic queuing model, censored regression, and econometric model etc. The various causal factors related to aircraft, airline operations, change of procedures and traffic volume are also discussed. Besides the above factors, this paper will also consider the characters of single flight and their effect on flight delay. This paper will detect the pattern of delay from the airport level and flight level in which delays occur, give basic statistics on their magnitudes and frequencies.

3 Methodology and Data

3.1 Data Description

The analysis is based on data from the Airline On-Time Performance Data from the Federal Aviation Administration (FAA) and the climatic data from the National Climatic Data Center (NCDC). Statistical models are presented for the estimation of a vector of airport average daily arrival delay or single flight arrival delay. These models are formulated using flight delay parameters and weather conditions at Orlando International Airport (MCO). The data used consists of the non-stop domestic flights on scheduled service by certificated carriers with the destination of MCO. This database contains departure delays and arrival delays for non-stop domestic flights by major air carriers, and provides such additional items as origin and destination airports, flight numbers, scheduled and actual departure and arrival times, cancelled or diverted flights, taxi out and taxi in times, air time, and non-stop distance. The flight data used is from 01/01/2002 through 12/31/2003 excluding the cancelled and diverted flights.

3.2 Methodology

The aim of the study is to identify the pattern of airline delay in Orlando International Airport. We propose a two-stage approach when studying the data set to complete the process of detection and interpretation. In the first stage, we use frequency analysis technique to quickly detect any periodicities with in the data set. In the second stage, we try to explain the cause of the detected frequencies (as many as we can) by a detailed regression analysis.

Stage I. In this stage we use frequency analysis to detect periodicities. The periodicity could be interpreted as a daily, weekly, monthly, or seasonal pattern. Since the two years (2002-2003) average daily arrival delay data is over a long period of time, to detect daily pattern better, we zoomed-in our analysis by looking at blocks of 72 days each. For example, the first 72 days' average daily arrival delay (minus global mean) is shown in Figure 1.

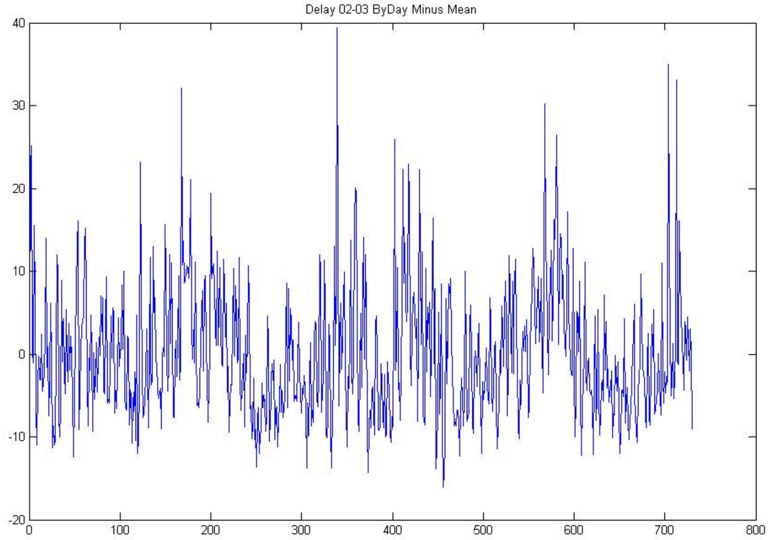


Figure 1. The average daily arrival delay vs. time (in days)

Let $f(t)$ denote the variable of the average daily arrival delay at day t. In order to eliminate the trend effect in $f(t)$, we employed a Least Square fit by cubic spline functions.

$$S_3(t) \approx f(t).$$

For example, we may subtract the values of LS fit function $S_3(t)$ from the corresponding values of $f(t)$ in Figure 1 and obtain the difference $d(t) = f(t) - S_3(t)$.

Then the problem can be formulated as the hidden pattern analysis by considering the model

$$d(t) = \sum_{k=1}^{p} a_k e^{i\lambda_k t} + e(t)$$

where (p, a_k, λ_k) are parameters to be estimated and $e(t)$ is the error term. We estimate the parameters by using the frequency analysis method that makes use of polynomials that are orthogonal on the unit circle. The numerical implementation of this method relies on the Levinson algorithm and its extension. The algorithm we use can be described as follows. For convenience, we will use $x(t)$ for $d(t)$ in the following discussion.

1) Generate the moments from $x(t), t = 1, 2, ..., N,$ and call these moments $r(h)$:

$$r(h)=\sum_{k=1}^{N-h}x(k)x(k+h)\text{ , }h=0,1,2,...,N\text{ .}$$

2) Apply the Levinson algorithm to the moment sequence $r(h)$ to generate the coefficients of the polynomials $\Phi_p(t)$ orthogonal on the unit circle with respect to weight on $[0,2\pi]$: $\frac{1}{2\pi}\left|\sum_{k=1}^{N}x(k)e^{-i(k-1)\theta}\right|^2 d\theta$, $t=e^{i\theta}$.

3) Find all the zeros of the orthogonal polynomial $\Phi_p(t)$ of degree p: $z_j=r_je^{i\theta_j}, j=1,2,...,p$

4) Use the arguments θ_j of the zeros in Step 3 to estimate the frequencies.

For example, if $x(t)=2\cos(\frac{\pi}{6}t)+\cos(\sqrt{2}t)$, $t=1,2,3,...N$, the zeros of $\Phi_4(t)$ are plotted (for N=40 to 200) in the following figure indicated by "+". The frequencies $\pm\frac{\pi}{6}$ and $\pm\sqrt{2}$ are indicated on the plot by "*".

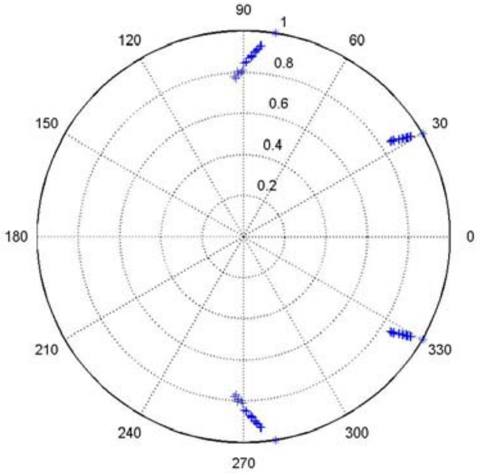


Figure 2. Illustration of zeros approaching the true frequencies

We see that the zeros are attracted to the correct "frequency points". For details on the method described here together with its improvement, please refer to the paper by C. F. Bracciali, Xin Li, and A. Sri Ranga (2005). When applying this algorithm to $d(t)$ above, we find the estimated frequencies (as listed in Figure 3 below). Figure 3 shows us the data with the least square fit by using the detected frequencies:

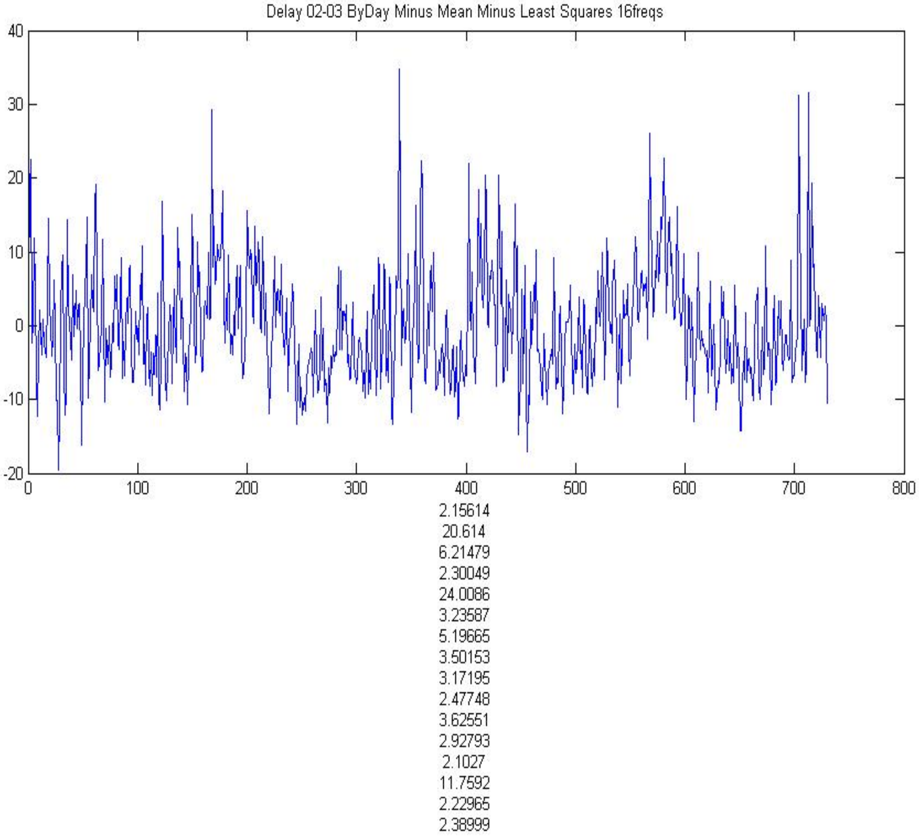


Figure 3. Plots of data and its representation using detected frequencies

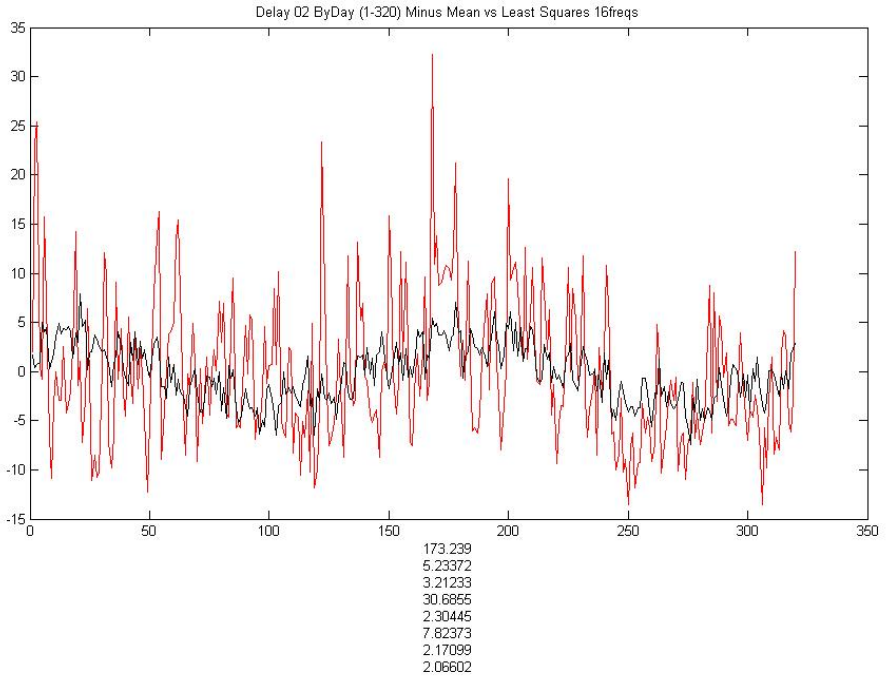
Stage II. Using the detected pattern from Stage I, in this stage, we use regression models to do a regression analysis in order to further detect the cause factors for the delay. The variables used in the regression are shown in appendix A.

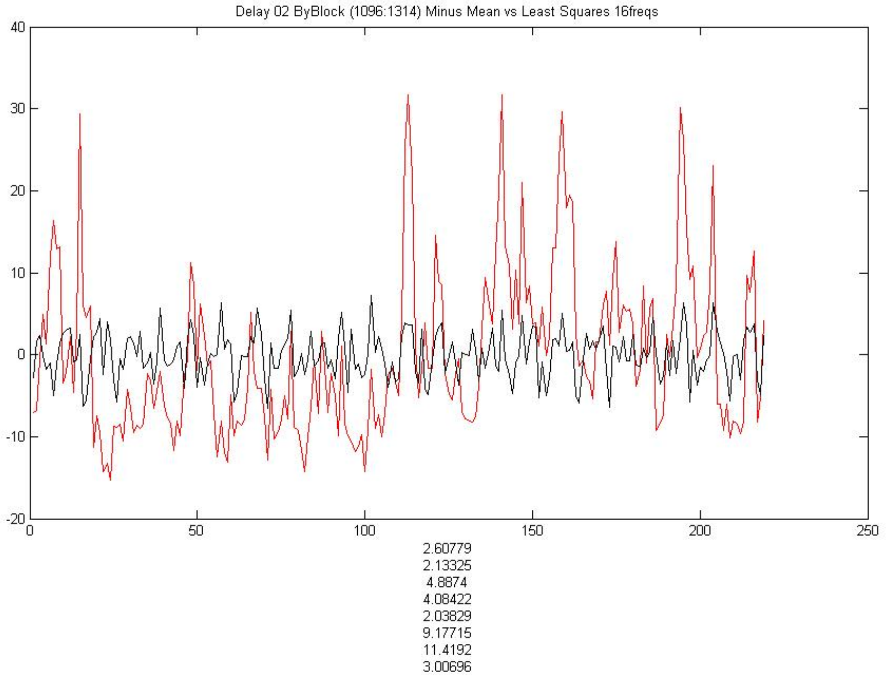
4 Results

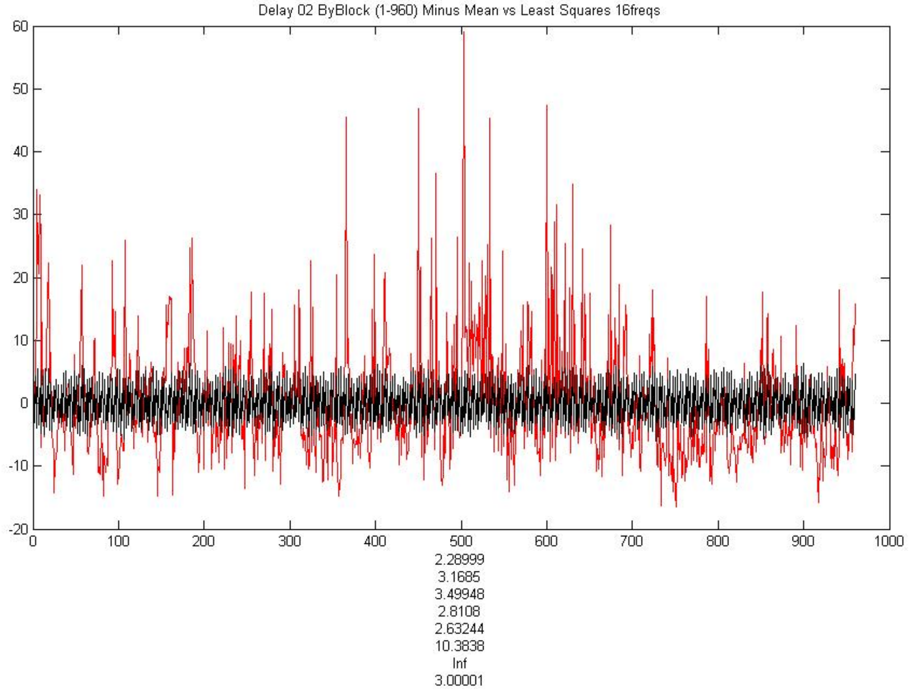
4.1 Pattern detection and interpretation using frequency analysis technique

For this experiment we formed a daily sequence where each data point was the average of the delays of all flights with a positive delay during the corresponding day. The mean was subtracted. Then, to test what patterns might be found at various timescales, we considered only the 1-320th days of the sequence. The Levinson algorithm was used to generate a polynomial of degree 50 to uncover the frequencies of periodic behavior in the sequence. The 8 strongest frequencies were then used to construct a model for the sequence using least squares, and the model and original sequence were plotted together.

This data set contained a seasonal pattern of 172 days (approximately half of a year), and a monthly pattern of 30.7 days. Here is an image showing a seasonal pattern. The graph is shown in figure 4.


Delay 02 ByDay (1-320) Minus Mean vs Least Squares 16freqs
35
30
25
20
15
10
5
0
-5
-10
-15
0
50
100
150
200
250
300
350
173.239
5.23372
3.21233
30.6855
2.30445
7.82373
2.17099
2.06602


Delay 02 ByBlock (1096:1314) Minus Mean vs Least Squares 16freqs
40
30
20
10
0
-10
-20
0
50
100
150
200
250
2.60779
2.13325
4.8874
4.08422
2.03829
9.17715
11.4192
3.00696



4.2 Pattern explanation by regression analysis of average daily delay

The multiple regression methodology was initially attempted. We assume that the errors are normally, identically, and independently distributed. Initial experimentation revealed that these assumptions do not hold. In particular we found that the errors are not normal. About the random error ε, the assumption that ε is normally distributed is the least restrictive when we apply regression analysis in practice. When non-normality of the random error term is detected, it can often be rectified by applying the transformations. After some experimentation a model is developed for the airport average daily delay. The R square values for this linear model is only 0.2755, which is not satisfactory. And different transformations of dependent variable are tried including, but the R square can't be improved. So some other statistics methods will be used to analyze the daily delay, such as ANOVA and logistic regression.

Analysis of variance (ANOVA) is used to study the effects of one or more independent (predictor) variables on the dependent (response) variable. Most commonly, ANOVA is used to test the equality of means by analyzing the total sum of squares (about the combined mean), which is partitioned into different components (due to model or due to random error).

The day of week differences are proved by F-test to be significant ($p<0.0001$). LSD test and TUKEY test both proved that the daily delay on Thursday and Friday are obviously higher than other weekdays. Tuesday and Saturday have the lowest daily delay in the week. This proved the weekly pattern of the delay. This pattern should be related with the weekly schedule of the airport.

At the same time, the seasonal differences are proved by F-test to be significant ($p<0.0001$). LSD test and TUKEY test both proved that in summer the daily delay is obviously higher than the other three seasons, and in the fall the daily delay is obviously lower than the other three seasons. In spring the daily delay is lower than in winter. This pattern should be related to the fact that summer is the rainy season in Orlando. The thunder may cause the interrupt of the operation of the airport, which increases the delay of the flights. That is to say the average daily delay shows seasonal pattern, which is the results of the detection of frequency analysis technique.

Logistic regression belongs to the group of regression methods for describing the relationship between explanatory variables and a discrete response variable. A logistic regression is proper to use when the dependent is categorized and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects. The dependent variable (average daily arrival delay) can take on three values: Y = 0 for delay<5 min; Y=1 for delay >=5 and < 10 min; Y=2 for delay>=10min. The average daily delay metric reflects only positive delays; flights that arrive early are assigned zero delay. For our delay metric, d(t), we used the average daily positive delay for all scheduled and completed flights from other airports to MCO airport in Orlando. It is the average of positive delay per flight per day. They are created by the bins with approximately equal frequencies.

Considering the week pattern of the daily delay, the week is classified into four levels: Monday, Wednesday, and Friday have no significant difference and are combined into one level. Tuesday and Saturday are combined into one level. Thursday and Friday are individually one level.

Here we treat the arrival delay as a categorical outcome with three levels and keep the natural ordering presented in the data. There are usually three different ways of generalizing the logit model to handle ordered categories. We use the Proportional Odds Model (Cumulative Logit Model). For this model, we have actually imposed the restriction that the regression parameters except the intercepts are the same for the two logit models. It implies that it doesn't make any difference how we categorize the dependent variable - the effects of the explanatory covariates are always the same. The significant variables are shown in Table 1 and the result of the modeling is shown in Table 2, 3 and 4.

Table 1 Definition of average daily delay model independent variables

Parameter	Definition
Precipitation	Hundredth of inches of the precipitation per day
Mon. Wed. and Sun.	The flight takes place on Monday, Wednesday and Sunday
Thursday	The flight takes place on Thursday
Friday	The flight takes place on Friday
Spring	The flight takes place in spring
summer	The flight takes place in summer
fall	The flight takes place in fall

Table 2 Results of logistic regression of the airport average daily arrival delay

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr>ChiSq
Intercept 3	1	-1.4073	0.1974	50.8322	<.0001
Intercept 2	1	0.8085	0.1927	17.6116	<.0001
precipitation	1	0.0264	0.00364	52.5429	<.0001
Mon. Wed. and Sun.	1	0.4505	0.1767	5.2256	0.0223
Thursday	1	0.9280	0.2402	14.9235	0.0001
Friday	1	1.3301	0.2457	29.3148	<.0001
spring	1	-0.2356	0.2026	1.3520	0.2450
summer	1	0.3870	0.2161	3.2104	0.0733
fall	1	-1.4519	0.2113	47.2132	<.0001

Table 3 Odds Ratio Estimates of the airport average daily arrival delay

Effect	Point estimate	95% Wald Confidence Limits	
precipitation	1.027	1.019	1.034
Mon. Wed. and Sun. vs Tuesday and Saturday	1.569	1.110	2.218
Thursday vs Tuesday and Saturday	2.530	1.580	4.051
Friday vs Tuesday and Saturday	3.781	2.336	6.120
spring vs winter	0.790	0.531	1.175
Summer vs winter	1.473	0.964	2.250
Fall vs winter	0.234	0.155	0.354

Table 4 Model Fit Statistics of the airport average daily arrival delay

Association of Predicted Probabilities and Observed Responses

Percent Concordant	73.6	Somers' D	0.503
Percent Discordant	23.3	Gamma	0.519
Percent Tied	3.1	Tau-a	0.330
Pairs	174384	c	0.752

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
5.5864	7	0.5888

From Table 4, test statistic for the Proportional Odds Assumption is 5.5864with the DF of 7, so the p value is 0.5888. The high p-value is desirable. For this problem, we find no reason to reject the proportional odds model.

Adjusting for other variables, the odds of having a higher arrival delay in summer will be 1.473 times of the odds in winter, and the odds of having a higher arrival delay in fall will be 0.234 times of the odds in winter. The odds of having a higher arrival delay in spring and in winter show no significant difference. This shows the significant difference between the delays in seasons and proves the seasonal pattern of delay.

On Monday, Wednesday and Sunday the odds of having a higher arrival delay will be 1.569 times of the odds in Tuesday and Saturday, on Thursday the odds of having a higher arrival delay will be 2.530 times of the odds in Tuesday and Saturday, and on Friday the odds of having a higher arrival delay will be 3.781 times of the odds in Tuesday and Saturday. From the odds ratio, on Thursday and Friday the airport is shown to have the higher probability to have delay more than 10 minutes.

For each 10*0.01=0.1 inch increase with the precipitation, the odds of having more arrival delay increases by exp(0.0264*10)-1= 30.2%. There are no significant interactions.

Compared with linear regression model, the multiple logistic regression model shows the same seasonal pattern and weekly pattern.

The daily delay on Thursday and Friday are obviously higher than other weekdays. Tuesday and Saturday have the lowest daily delay in the week. In summer the daily delay is obviously higher than other three seasons, and in fall the daily delay is obviously lower than other three seasons. In spring the daily delay is lower than in winter. At the same time precipitation is found to contribute to the daily delay

4.3 Pattern explanation by delay model on the single flights

To introduce the factors into statistical model and test their main effects on delay of flights, the before-time flights (with arrival delay <0) from other airports to MCO are not identified. The Bureau of Transportation Statistics (BTS) compiles delay data for the benefit of passengers. They define a delayed flight as one in which the aircraft fails to release its parking brake less than 15 minutes after the scheduled departure time. The FAA is more interested in delays indicating surface movement inefficiencies and will record a delay when an aircraft requires 15 minutes or longer. So the single flights here are categorized into two groups: no delay (0<=arrival delay<15), delayed (15<=arrival delay).

So that the dependent variable Y (delayed flights) here takes on two values: Y = 0 for no-delayed flights (with less than 15 minutes delayed), Y = 1 for delayed flights (with more 15 minutes delayed). The no-delayed flights and high-delayed flights respectively represent 61.42% and 38.48% of the whole data set. Binary logistic regression is proper to use here when the dependent is a dichotomy (an event happened or not) and can be applied to test association between a dependent variable and the related potential factors, to rank the relative importance of independents, and to assess interaction effects.

We collected all the single flight data from 01/01/2002 through 12/31/2003 excluding the before-time arrival flights, the cancelled and diverted flights. The variable factors include information of delayed flights, as well as the corresponding airport conditions (Appendix B).

Airport precipitation in Orlando international airport is a weather variable from National Climatic Data Center (NCDC). The precipitation is hundredth inch of rainfall per hour, different from the precipitation in the model of average daily delay, which is hundredth inch of rainfall per day.

The space is used here another variable, which means the intervals between two consecutive arriving flights. This inter-arrival time is calculated the time between each flight and the before flight according to the schedule arriving time. The aim of space is to find the relationship of delay and the schedule operation of the airport. The intervals between arrival flights differ under different weather conditions, runway conditions and operating conditions. If the flight flows arrive smoothly, we can say that the space is constant and small which means the waiting and service time for each flight is controlled at a low level. The results of the model of single flight are shown below.

Table 5 Definition of delay model independent variables

Parameter	Definition
Thursday and Friday	The flight takes place on Thursday and Friday
summer	The flight takes place in summer
winter	The flight takes place in winter
Fall	The flight takes place in fall
evening	The flight takes place between 5pm to 11:59pm
afternoon	The flight takes place between 12pm to 4:59pm
distance between 750 and 1000	The flight distance is in between 750 and 1000 miles
distance > 1000	The flight distance is larger than 1000 miles
precipitation	Hundredth of inches of the precipitation per day
log_space	The log transform of the space

The final logit model is presented in Table 6 and 7 and 8.

Table 6 Model estimation of single flight arrival delay

Parameter	Estimate	StandardError	Wald Chi-Square	Pr > ChiSq
Intercept	-1.0827	0.0287	1423.1896	<.0001
winter	0.1432	0.0215	44.5419	<.0001
fall	-0.2595	0.0233	123.7075	<.0001
summer	0.2191	0.0211	107.7185	<.0001
Evening	0.7179	0.0210	1166.2651	<.0001
Afternoon	0.2889	0.0226	163.9135	<.0001
Thursday and Friday	0.1172	0.0164	50.8101	<.0001
distance > 1000 miles	-0.0338	0.0214	2.4898	0.1146
distance between 750 and 1000 miles	0.1921	0.0182	111.4865	<.0001
precipitation	0.0226	0.00111	414.5916	<.0001
log_space	-0.0198	0.00905	4.7965	0.0285

Table 7 Odds Ratio Estimates of single flight arrival delay

Effect	Point Estimate	95%Wald Confidence Limits	
Winter vs spring	1.154	1.106	1.203
Fall vs spring	0.771	0.737	0.808
Summer vs spring	1.245	1.194	1.297
Evening vs moring	2.050	1.967	2.136
Afternoon vs moring	1.335	1.277	1.395
Thursday and Friday vs other weekdays	1.124	1.089	1.161
distance>1000vs distance<750	0.967	0.927	1.008
distance in[750,1000]vs distance<750	1.212	1.169	1.256
precipitation	1.023	1.021	1.025
log_space	0.980	0.963	0.998

Table 8 Model Fit Statistics of single flight arrival delay

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	60.9	Somers' D	0.224
Percent Discordant	38.4	Gamma	0.226
Percent Tied	0.7	Tau-a	0.106

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
17.4662	8	0.1226

From Table 8, Hosmer and Lemeshow Goodness-of-Fit Test statistic is 17.4662 with the DF of 8. The resulting p value of 0.1226, shown in Table 8, suggests that the model fits well. We find no reason to reject the odds model at 5% confidence level.

Four factors including season influences, time influences, distance influences, precipitation and space show significant association with the likelihood of flights of delays.

The 'odds ratio' column in Table 7 is obtained from the parameter estimates. For the dummy variable 'summer', which indicates the flight takes place in summer, the odds ratio is 1.245. The odds ratio of 1.245 indicates that the predicted odds of 'delay flights' for summer are 1.245 times the odds for spring. In other words, the predicted odds of delay are about 24.5% higher in summer than in spring. The dummy variable 'winter' indicates that when the flight takes place in winter, the odds ratio is 1.154. This implies that the predicted odds of delay are about 15.4% higher in winter than in spring. The dummy variable 'fall' indicates that when flight takes place in fall, the odds ratio is 0.771. This implies that the predicted odds of delay are about 23% lower in fall than in spring. This shows the same results as the frequency analysis and the average daily delay model, that the flight arrival delay shows seasonal pattern.

The odds ratios for the arrival time show the relative ratios of high delay between different times (morning, afternoon, an evening) for each flight. Compared to morning, the odds of delay in the afternoon could be 1.335 times higher and the odds in the evening could be 2.05 times higher. In the evening the flights are much more likely to be delayed than in the morning. The results testify that the time contribute to delay. And this finding corresponds to the result of frequency analysis that arrival delay shows daily pattern.

The odds ratio of the distance variables is interesting. For each flight with the flight distance between 750 to 1000 miles, the odds of having arrival delay will increase by 1.212-1=21.2%. While for the flight with the flight distance larger than 1000 miles, the odds of having arrival delay will decrease 1-0.967=3.3% than the flight with the flight distance less than 750 miles.

The odds ratio for the precipitation shows that for each 10*0.01=0.1 inch increase with the precipitation, the odds of having more arrival delay increases by exp(0.0226*10)-1= 25.3%.

The odds ratio for the log space shows that as the space increases, the probability of the flights being delayed will decrease. Since the space is the inter-arrival time for two successive flights, when the space increases, the service time for each flight will increase, so that the efficiency of the operation will decline.

5 Conclusions

This paper is concerned with the detection of the pattern of flight delay. This methodology enables us to investigate several issues that have not heretofore received much consideration. First, pattern of arrival delay at the flight level and the airport level are individually measured and compared. The pattern is detected by frequency analysis method, and then interpreted by regression method. The seasonal, weekly and daily patterns are proved. However some hidden pattern cannot be proved by the regression method such as the half-week pattern yet.

Second, the characters of single flight and their effect on flight delay are considered. So the patterns of delay from the flight level in which delays occur are analyzed, and the significant reasons of delay are given out. The precipitation, flight distance, season, weekday, arrival time and the space between two successive arriving flights are found to contribute to arrival delay of flights.

Third, the effect of a flight on the immediate flight is considered. We measure the time interval of two consecutive flights and analysis its effect on the flight delay. The results show that the space between two successive arriving flights increases, the probability of the flights being delayed will decrease.

References: (omitted to save space)

Appendix A Variables used in the regression model

Variables	Definition
Flight arrival delay	Difference of the actual arrival time minus the scheduled arrival time
Maximum hourly flow rate	Maximum numbers of operation that can be handled in a one-hour period under specific operating conditions
Arrival demand	Number of completed arrival flights to MCO per day according to the scheduled arrival time.
Flight duration	Airborne time for each flight
Space of Inter-arrival time	Intervals between two consecutive arriving flights
Airport precipitation	Daily observations about the inches of the rainfall
Airport wind speed	Daily average wind speed at MCO, speed of wind in mph per day.
Seasonal variables	Indicate the seasons when the flights are scheduled, spring (March-May), summer (June-August), fall (September to November), and winter (December to February).
Weekly variables	Indicate the weekday when the flights are scheduled
Time variables	Indicate the scheduled arriving time of each delayed flights, morning (7am to 11:59 am), afternoon (12 am to 4:59 pm), and evening (5 pm to 11:59 pm).
Origin airport regional variables	The regional effects are captured by a set of dummy variables, south, east, central, and west areas (definition is in Appendix B).
Flight distance	Categories of flight distance, which respectively represent the distance of 0 to 750 miles, 750 to 1000miles and greater than 1000 miles.

Appendix B Definition of Regional variables

Areas	Definition
South	States of Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, and Tennessee
East	States of Connecticut, Delaware, District of Columbia, Indiana, Kentucky, Maine, Maryland, Massachusetts, Michigan, New Hampshire, New Jersey, New York, Ohio, Pennsylvania, Rhode Island, Vermont, Virginia, and West Virginia
Central	States of Arkansas, Colorado, Illinois, Iowa, Kansas, Louisiana, Minnesota, Missouri, Nebraska, North Dakota, Oklahoma, South Dakota, Texas, and Wisconsin
West	States of Alaska, Arizona, California, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming

Appendix B: User Guide for Using the Programs

In this appendix, information on how to install and use the program contained in the CD ROM is given.

I. Installation of the program

To install the program, change to your CD-ROM drive and double click the program named "install.com" or its shortcut (PEGASUS) icon.

The installation process will create a directory named "fdotucf" on your C: drive automatically.

One the installation process is finished, you can start the program from DOS command line or in Windows environment by going to the directory C:\fdotucf and double click the program file named "run.bat".

You may also run from the CD ROM by using the command "runCD.bat [CDROM Drive letter]" from the CD ROM itself. For example, if the CD ROM drive letter is D, then once you change to D: drive, you can issue "runCD D" to run the program.

To uninstall the program, just delete the directory fdotucf on your C: drive.

Please read more on how to use the program in next section.

II. Using the Pattern Detection Graphic User Interface

The purpose of this program is provide a user friendly interface when using the techniques developed in the FDOT project to determine the periodicities (or frequencies) in a given data set. These frequencies can then be used to identify hidden patterns within the data. In the process, an approximation of the data will be constructed from the determined frequencies and plotted. This plot can be used to estimate how well the details of the data are being captured by the determined frequencies. A pre-fit using cubic splines or a line can be applied to improve the approximation or suppress certain trends. The steps required to perform the frequency detection are detailed below.

Step I: Preparing a Data File

Before beginning to use the pattern detection software, it is necessary to prepare a data file. The data file should be a sequence of numbers, one number per line, stored in plain text (.txt) format. To find patterns in time, the sequence should be organized by time in equally spaced intervals, and each interval during the period of observation should have a datum in the file. For instance, if observations are made by day, then the data in the file should be ordered by the order of the days, each datum should represent exactly one day, and each day should have a datum in the file. A sample data file SampleData.txt has been included.

Step II: Opening a Data File

Run the pattern detection software by clicking the program "run.bat" in the directory C:\fdotucf. Click the Open Data File button. A directory listing will appear. Browse to the location of the of the data file of your choice, select it, and click the Open button. The file will load and its name will appear above the Graph Region.

Step III: Select Basic Detection Options

Enter a number into the Deg. of Orthogonal Polynomial text box, or use the default. This number must be an integer and must be greater than zero. This number will determine the degree of orthogonal polynomial generated by the Levinson algorithm. The number also determines how many frequencies can be used in the reconstruction of the data, and can also affect the precision with which frequencies are determined.

Enter a number into the Number of Significant Frequencies text box, or use the default. This number must be an integer greater than or equal to zero and must be at most half of the degree of the polynomial (the value in the text box above). This number determines how many of the detected frequencies are used in the reconstruction of the given data. Increasing this number will improve the fit. Entering "0" (zero) into this text box will allow you to view either the cubic spline or linear fit without additional frequency detection if either of these options is selected (see below).

Step IV: Run the Pattern Detection

Click the Run button. The fitting algorithm will execute and the graph of the reconstruction of the data will appear in the Graph Region. The frequencies used in the reconstruction will be converted to periods (more useful for interpretation) and the periods will be listed in the Most Significant Periods box with the most important periods appearing at the bottom.

To plot the original data along with the reconstruction, check the Plot Original Data checkbox, and click the Run button to update the graph.

Additional Features:

Removing a Linear Trend:

To remove a remove a linear trend from the data, activate the Take out linear trend radio button and click Run to update the plot and significant periods. A linear fit will be performed on the data and subtracted out before the frequency detection is performed. Removing a linear trend can help strengthen higher frequency patterns. To see just the linear fit without frequency detection, enter zero into the Number of Significant Frequencies text box. If you no longer wish the program to search for linear trends, select one of the other radio buttons.

Removing a Cubic Spline Fit:

To remove a remove a cubic spline fit from the data, activate the Take out cubic spline trend radio button. The Divisions text box will become active. Enter a nonnegative integer into this box or use the default. The number in this box will determine how many knots are used in the cubic spline fit. The data will be divided into the entered number of divisions and a knot placed at the ends of each division. Thus the number of knots will be one greater than the number of divisions, with all knots equally spaced and one at each endpoint.

Click Run to update the plot and significant periods. A cubic spline fit will be performed on the data and subtracted out before the frequency detection is performed. As with a linear fit, removing a cubic spline fit can help strengthen higher frequency patterns. To see just the cubic spline fit without frequency detection, enter zero into the Number of Significant Frequencies text box. If you no longer wish the program to search for cubic spline trends, select one of the other radio buttons.

Deactivating Cubic Spline Fit/Linear Fit:

Select the No preprocessing radio button to stop searching for linear or cubic spline trends.

III. List of programs

Here is a complete list of programs developed and used in this project with brief descriptions on what they do and how to use them.

computeAllSubmodelsofModelCorrs(vectData, intDeg, intNumTerms, vectPeriods)

It computes correlations with the original data. The correlations are computed with the model which uses all of the periods in "vectPeriods".

cosmat(freq,num)

This routine is no longer used since it duplicates a Matlab function

cubicSplinesLeastSquares(vectData,intDivisions)

This program beahves just like "cubicSplineLeastSquaresNoPlot", but also plots the cubic spline best fit against the data vector.

cubicSplinesLeastSquaresNoPlot(vectData,intDivisions)

This program takes as input a data vector, vectData, and returns a cubic spline fit with intDivisions + 1 number of equally spaced knots. A primary use for this program is to subtract out a cubic spline fit for a dataset to eliminate long period patterns.

cubicSplinesTest(intDomainSize,intGenDivisions, intTestDivisions, intLevDeg, intLevNumTerms)

The purpose of this program is to test how well a function of random cubic splines can be modeled with the sum of a cubic spline model with fewer knots and a trigonometric model generated with the Levinson algorithm. Specifically, it is to see how well the trigonometric model captures the details lost by reducing the cubic spline to fewer knots. "intDomainSize" is the size of the domain, "intGenDivisions" determines how many knots to use in the randomly generated cubic spline model, "intTestDivisions" determines the number of knots to use in the approximating cubic spline model, *"intLevDeg" is the degree used in the approximating Levinson algorithm, and "intLevNumTerms" is how many of the resulting frequencies to use in the model. The program plots the random cubic spline vs the cubic spline fit, vs the final fit and returns the error between the random cubic spline and the final fit*

generateRandomCSFunctions(intDomainSize,intGenDivisions, intNumFunctions)

This program generates random functions by constructing random linear combinations of cubic splines. "intGenDivisions" determines the number of cubic splines used to construct each function and "intDomainSize" determines at how many places the resulting functions will be evaluated. "intNumFunctions" is the number of functions to be constructed in this way. The evaluations of each resulting function is turned into a column vector and the column vectors assembled into a matrix. The matrix is returned as the output of the program.

getLeastSquaresModel(vectData,intDeg,intNumTerms)

Similar to "plotDataVsLeastSquares", this program takes a data vector, "vectData", and constructs a least squares best fit using trigonometric sines and cosines. The frequencies for these trigonometric functions are determined by a Levinson algorithm of degree "intDeg', and then the best "intNumTerms" frequencies are used in the reconstruction.

graphroots3(x,n,minlength,lengthstep,maxlength,cap)

graphroots3 displays results of the algorithm for a data that has been truncated at various indices on the same graph. The arguments are (datavector, degree, min, step, max, caption). That is to say, the data vector is truncated to the length "min", and that number is iteratively increased by "step" until it reaches "max". "Max" must be less than the length of the data vector and "max" must equal "min" + n"step" for some whole number n, or an error will result.*

graphrootsDeg(x,degmin,degmax,cap)

graphrootsDeg displays results from the same dataset for different degrees on the same graph. Its arguments are (datavector, mindegree, maxdegree, caption). Degree is increased by one from mindegree until it reaches maxdegree. We have been using mindeg=7 and maxdeg=20.

graphrootsMulti(x,deg,cap)

This function graphs several data sets on the same graph. datamatrix is a matrix made up of data vectors. Each data vector must be of the same length.

labelAxes(strArgument, intDegree, intNumTerms, strCaption)

This program adds the most significant periods to the bottom of a graph generated by "plotDataVsLeastSquares". The most significant will appear last. "strArgument", "intDegree" and "intNumTerms" should be given the same values as they were given in "plotDataVsLeastSquares" when the graph was first generated. "strCaption" is a text arguement. Its contents will be added as a caption to the top of the graph.

labelAxesAvg(strArgument, intDegree, intNumTerms, strCaption)

This function is similar to "labelAxes". However, instead of listing all of the most significant periods, this program rounds to the nearest half-unit and discards any duplicates.

levinson2(x,n)

Levinson Algorithm. This is the core of the program.

minsearchDataVsLeastSquares(vectPeriods, vectData)

This is a slightly modified and reorganized variation of "plotDataVsLeastSquaresNoLevin"

It returns the negative of the correlation of a vector of input data, "vectData", with the least squares best fit using trigonometric functions with periods from "vectPeriods". By using a minimization technique on the negative correlation with respect to "vectPeriods", we can find the periods which maximize the correaltion to the data.

plotDataVsLeastSquares(vectData,intDeg,intNumTerms)

This function takes a dataset, analyzes it, reconstructs a least squares fit of the data, and then plots the data along with the fit on the same plot (original data in red, fit in black). It also produces a matrix of frequencies in terms of significance (magnitude of the associated zero), and a matrix of least squares amplitudes in terms of frequencies. "vectData" is the input vector, "degree" is the number of degrees to be used in the Levinson algorithm, and "intNumTerms" is the number of frequencies to be used in the least squares reconstruction.

plotDataVsLeastSquaresNoLevin(vectData, vectPeriods)

This program takes a data vector and a vector of periods and performs a least squares best fit to the data using trigonometric functions with the given periods. The resulting best fir is plotted (in black) along with the original data (in red). The correlation of the best fir with the data is also returned. This program is used primarily to test how a particular fit, generated by other means, behaves when periods are added or removed.

plotDataVsLeastSquaresRnd(vectData,intDeg,intNumTerms)

This function is similar to "plotDataVsLeastSquares", but whenever a pair of frequencies are within a half-unit of eachother, one is discarded before computing the least square fit.

polyFit2(vectInput, intDeg)

This program takes a data vector, performs a polynomial fit of the specified degree, and then constructs a polynomial vector made up of the created polynomial evaluated at all of the indicies of the *data vector. The vector this constructed is the same size as the datavector. (use datavector-polyFit2(datavector, n) to get a dataset with the fit removed).*

setupDataVector2(inputArray)

This function turns an array into a row vector by simply listing each row into a single row.

trigMatrix(freqvect,ampvect,num)

This program generates a matrix generates a matrix whose columns are various sine/cosine functions evaluated over a given domain. "freqvect" is a vector containing the frequencies on the trigonometric functions to be evaluated. "ampvect" is a vector containing the amplitudes of the cosine and sine for each frequency (in that order). Clearly, "ampvect" must be exactly twice as long as "freqvect". These trigonometric functions will be evaluated at every integer btween 1 and "num", and the resulting matrix will be returned. This routine needs to be improved to use builtin Matlab matrix construction techniques.

computeAllSubsetsofFrequeciesCorrs(vectData, intDeg, intNumTerms, vectPeriods)

Given a vector of periods "vectPeriods", this program takes every possible subset of the listed periods and uses "plotDataVsLeastSquaresNoLevin" to compute the correlations between the given data and the trigonomteric models generated using each such subset. The correlations are returned as a vector.